UNIVERSIDADE DE LISBOA

FACULDADE DE LETRAS

Português Controlado para a Tradução Automática: Português → Italiano

Marianna Buchicchio

Relatório de projeto orientado pela Prof.ª Doutora Palmira Marrafa, especialmente elaborado para a obtenção do grau de Mestre em Tradução

2017

ÍNDICE

Agradecimentos

À Professora Doutora Palmira Marrafa, orientadora deste trabalho de projeto. Antes de mais, agradeço por me ter ensinado a "olhar para os dados", frase enigmática, mas que foi essencial durante a elaboração deste trabalho. Obrigada pela paciência, pelos conselhos extremamente preciosos e por ter acreditado em mim desde o início do Mestrado. Obrigada por me ter oferecido a oportunidade de participar na Conferência de Varsóvia.

À minha família. À Mamma, ao Babbo e ao Giulio. Grazie. Obrigada, porque sem o vosso apoio não teria conseguido. Obrigada por terem ouvido os meus caprichos, os meus desabafos, as minhas queixas e por me terem apoiado sempre e incodicionadamente. Sempre lontani ma sempre vicini, sempre nel mio cuore. Quero agradecer também a uma parte de mim que, infelizmente, já não está comigo. Obrigada Nonna! Obrigada por ter sido a minha segunda mãe, mas, sobretudo, a minha primeira e grande fã.

Ao Miguel. Obrigada por seres a pessoa fantástica que és. Obrigada pela paciência e pelo amor. Estar ao meu lado nos momentos de felicidade é fácil, mas é nos momentos difíceis que se vê o valor de uma pessoa. Ajudaste-me a ultrapassar os meus medos e os meus limites e acompanhaste-me neste caminho desde o início. Não tenho palavras para exprimir a gratidão, só sei dizer que isto tudo foi possível porque estiveste ao meu lado. O teu apoio e o teu amor foram essenciais. Un grazie "più grande del mondo".

Aos meus amigos "italianos, mas portugueses", que foram a minha família aqui em Portugal. Obrigada por terem sido os melhores que podia desejar. Obrigada por terem partilhado esta viagem comigo.

Às amigas de uma vida inteira. À Camilla, à Vanessa, à Samantha e à Giulia. Não há distância quando a amizade é verdadeira. Obrigada por serem as melhores de sempre.

A Lisboa, à minha Madeira e à língua portuguesa, a minha grande paixão.

O último agradecimento vai para mim própria porque, não obstante os medos e as condições adversas, fui tenaz e cheguei ao fim. Dei-me uma oportunidade e realizei o meu grande sonho: viver e estudar em Portugal. Posso finalmente dizer que consegui.

RESUMO

Neste trabalho propõe-se um fragmento de português controlado para a Tradução Automática. A linguagem controlada proposta é do tipo *Machine-oriented Controlled Language* (MOCL), ou seja, orientada para a máquina, em concreto, para a redação de textos a serem traduzidos por um sistema de tradução automática. Em termos de cobertura lexical, este fragmento não se destina a um domínio específico, pelo que, dado o seu carácter geral, pode ser utilizado para escrever textos diversos.

O português controlado para a tradução automática para italiano proposto neste trabalho tem como objetivo a simplificação das estruturas dos textos de partida que causam problemas de tradução automática e, consequentemente, a eliminação, das ambiguidades, entre outros aspetos indutores de maus resultados, para se obter um *output* gramatical. Discute-se também uma "alternativa" à aplicação da linguagem controlada utilizando uma ferramenta disponibilizada pelo sistema de tradução automática SYSTRANet, demostrando-se que é possível resolver alguns problemas de tradução através da aplicação do português controlado e do auxílio de tal ferramenta. Por outras palavras, este fragmento de português controlado permite a redação controlada de textos cuja tradução produzirá resultados satisfatórios no que respeita aos fenómenos cobertos pelas regras apresentadas.

PALAVRAS CHAVE: tradução automática, linguagem controlada, português controlado para a tradução automática PT > IT

ABSTRACT

In this study we propose a fragment of controlled Portuguese for machine translation into Italian. The Controlled Language here proposed is a Machine-oriented Controlled Language (MOCL) and can be used to write texts that are destined to Machine Translation. This fragment, because of its peculiarities, is not created to cover a specific technical area but, instead, to write and than translate texts that belong to different areas.

The scope of controlled Portuguese for machine translation into Italian here proposed is to simplify the structures of the source text which cause translation problems and, consequently, to eliminate ambiguities in order to obtain an acceptable output. In addition, we also discuss an "alternative" to the "traditional" controlled language, by using a tool offered by the machine translation system SYSTRANet. We demonstrated that is possible to resolve some of the translation problems analysed by using the controlled Portuguese and this translation tool. In other words, the fragment of controlled Portuguese here proposed allows the writing and the translation of texts written in this controlled language, so that the resulting translation is acceptable and "correct" because it follows the rules here presented.

KEY WORDS: machine translation, controlled language, controlled Portuguese for machine translation PT > IT

1. INTRODUÇÃO

O objetivo deste trabalho é criar um conjunto de regras, gerais e específicas, para o controlo da língua portuguesa para a tradução automática para italiano. Para o efeito, utilizou-se o sistema de tradução automática SYSTRANet, versão gratuita do sistema de tradução automática SYSTRAN e disponível online. A linguagem controlada, em termos gerais, é um conjunto de restrições a aplicar no controlo de textos e pode ser utilizada para facilitar a leitura de um texto ou para melhorar a comunicação numa dada língua (HOCL, *Human-oriented Controlled Language*) ou para a redação de textos a serem processados por uma máquina (MOCL, *Machine-oriented Controlled Language*). Independentemente do tipo de linguagem controlada, as aplicações são múltiplas e podem ser utilizadas para a representação do conhecimento, para a redação de textos técnicos, para a simplificação e o melhoramento de textos em língua natural e para o controlo do desempenho de sistemas de tradução automática (Marrafa *et al*., 2012:153).

A criação de regras de linguagem controlada, em termos gerais, visa a eliminação das estruturas problemáticas na elaboração de textos, para determinados efeitos, numa dada língua. Neste sentido, nas linguagens controladas orientadas para os humanos (HOCL), controla-se a língua para que a leitura do texto resulte mais fácil para os falantes não nativos, com o objetivo de facilitar a compreensão e a comunicação. Por outro lado, nas linguagens controladas orientadas para uma máquina (MOCL) pode acontecer que, na fase do controlo da língua, o *input* seja degradado e que resulte agramatical. Neste caso, a degradação do *input* é feita em função da obtenção de um *output* gramatical na língua de chegada. Este é também o objeto de estudo deste trabalho, em que se propõe o controlo da língua portuguesa para a tradução automática para italiano, com o controlo de estruturas que causam problemas de tradução.

No que diz respeito às linguagens controladas orientadas para a máquina (MOCL), é desejável que as linguagens controladas permitam a obtenção de melhores resultados pelo menos para os sistemas de um dado paradigma. Como a criação de uma linguagem controlada de ampla cobertura não cabia nos objetivos deste trabalho, optou-se por criar um conjunto de regras para o controlo do português para a tradução automática para italiano, tendo-se usado o sistema de

tradução SYSTRANet, versão gratuita e disponível online do sistema de tradução automática SYSTRAN.

Procede-se aqui a uma apresentação das várias tipologias de sistemas de tradução automática, assentando nessa base a análise e o enquadramento histórico e teórico da área. Consequentemente, começa-se pela descrição dos paradigmas de tradução automática (orientados para o conhecimento linguístico, orientados para os dados e híbridos), tendo em conta as diferentes abordagens à tradução automática de cada paradigma.

Neste contexto, e tomando como base de inspiração o proposto em Marrafa *et al*., (2011), propõe-se, subsequentemente, um fragmento de português controlado para a tradução automática para italiano em que são analisadas especificidades da língua portuguesa e da língua italiana.

1.2 METODOLOGIA

Para a criação do fragmento de português proposto neste trabalho, partiu-se da perspetiva histórica e teórica da tradução automática a partir dos anos 40, anos em que a tradução automática era utilizada sobretudo nos Estados Unidos para a tradução russo-inglês, passando posteriormente por um período de estagnação (sobretudo na segunda metade da década de 60), mas recebendo novos estímulos à investigação nos anos a seguir, sobretudo na Europa e no Japão. Este enquadramento histórico serviu como base para a descrição dos paradigmas de tradução automática orientados para o conhecimento linguístico, para os dados e os híbridos. Na descrição dos paradigmas orientados para os dados, apresentam-se os primeiros sistemas de tradução automática baseados em dicionários e, a seguir, os mais recentes RBMT (*Rule-based Machine Translation*), com referência às suas diferentes abordagens, nomeadamente o *transfer* e a interlíngua. No que diz respeito aos paradigmas orientados para os dados, descrevem-se os sistemas SMT (*Statistical-based Machine Translation*) e EBMT (*Example-based Machine Translation)*, ambos baseados em *corpora*, mas que diferem quanto aos mecanismos de tradução. Por isso, para os sistemas EBMT, foi preciso aprofundar também a questão da "Tradução por Analogia". Deu-se relevo aos sistemas RBMT e SMT porque o sistema de tradução automática SYSTRAN é fundamentalmente *rule-based*, ainda que hoje em dia seja disponibilizado numa versão híbrida, que combina os módulos baseados em regras com uma componente estatística. Por

último, no que respeita aos paradigmas híbridos, face ao que é a matéria central deste trabalho, são mencionadas as principais estratégias de hibridização, com particular importância dada aos sistemas que integram as regras típicas dos sistemas *Rule-based* com o alinhamento estatístico.

No que diz respeito às linguagens controladas, é apresentada uma panorâmica das origens até às aplicações mais modernas, citando os exemplos mais importantes de linguagem controlada que servem como base na criação do fragmento do português controlado.

Definem-se dois critérios fundamentais para a criação de uma linguagem controlada - a legibilidade e a traduzibilidade -, que orientam dois tipos diferentes de linguagem controlada: HOCL (*Human-oriented Controlled Language*) para a legibilidade e MOCL (*Machine-oriented Controlled Language*), para a traduzibilidade. A estas duas tipologias de linguagem controlada correspondem duas abordagens distintas: a abordagem naturalista, típica das HOCL, na qual a linguagem controlada visa a simplificação de textos produzidos numa dada língua e em que permanecem ainda estruturas ambíguas; e a abordagem formalista utilizada nas MOCL, na qual as regras de linguagem controlada são bem definidas e o processamento por parte de uma máquina resulta mais fácil. Por último, é descrito o processo de criação de regras, que podem ser proscritivas, que definem as estruturas que não são permitidas, ou prescritivas, as quais indicam as estruturas permitidas.

As regras deste fragmento são proscritivas, porque são baseadas numa língua específica, o português, e descrevem as estruturas não permitidas. Contudo, é necessário acrescentar que além da determinação das estruturas não permitidas, as regras especificam também quais são as estruturas e as construções que é oportuno utilizar, para fornecer uma alternativa à proscrição e para deixar indicações claras para o uso correto de tal linguagem, sendo, nesse sentido, também prescritivas. Por último, o fragmento de português controlado proposto neste trabalho é orientado para a tradução automática, razão por que satisfaz o critério de traduzibilidade. Quanto ao fragmento de linguagem controlada proposto, as regras são divididas entre regras gerais, ou seja, algumas restrições gerais para a redação de texto que indicam o que é preciso evitar e regras específicas, que identificam as estruturas que é necessário evitar e as que é preciso utilizar. Para a identificação das estruturas ambíguas, foi feita uma análise contrastiva das especificidades da língua

portuguesa e da italiana com o auxílio de artigos escritos por especialistas e de gramáticas descritivas. Como no âmbito deste trabalho não cabia a elaboração de uma linguagem controlada de ampla cobertura, optou-se por estudar as especificidades relativas a modo, modalidade, tempo e aspeto. No que diz respeito ao modo, é estudada a variação no uso dos modos verbais em frases finitas e não finitas, dando-se particular atenção à utilização do infinitivo flexionado e do futuro do conjuntivo. Quanto à modalidade, no âmbito da modalidade epistémica, é analisada a utilização do verbo modal *dever* e, por outro lado, no âmbito da modalidade deôntica, o uso da expressão verbal *ter* + *de*. Por último, analisa-se o uso do pretérito perfeito simples, que corresponde ao *passato prossimo* e ao *passato remoto* em italiano, e do aspeto progressivo, mostrando os pontos de divergência nas duas línguas. Tendo como base esta análise foi criado um *corpus* composto de exemplos tirados das gramáticas utilizadas e de outros criados para o efeito. Cada exemplo foi testado com o sistema de tradução automática SYSTRANet e de seguida controlado e testado novamente. Um exemplo é composto por quatro frases:

(1a) Ao rever o amigo, deu-lhe um longo beijo.
(1b) *Alla revisione l'amico, gli ha dato un lungo bacio.
LC: (1c) Quando reviu o amigo, deu-lhe um longo beijo.
(1d) Quando ha rivisto l'amico, gli ha dato un lungo bacio.

em que o número identifica o exemplo e as letras têm as seguintes correspondências:

- (1): número do exemplo;
- (a): frase na língua de partida;
- (b): resultado da tradução automática de (a);
- LC:(c): frase escrita em linguagem controlada;
- (d): resultado da tradução automática de (c).

Os exemplos mostram os fenómenos a evitar, os resultados agramaticais de tradução automática identificados por "*", como é usual, a aplicação das regras e o

resultado final do controlo. A cada exemplo segue-se a descrição dos fenómenos a evitar, indicando-se quando há casos de agramaticalidade e/ou de ambiguidade.

Na fase de teste dos exemplos ocorreram fenómenos que não cabem nos objetivos deste trabalho, mas que foram igualmente abordados e explicados nas notas de rodapé. Neste caso, foram criadas regras para o controlo e foram explicitadas as estruturas a utilizar no controlo.

Por último, discute-se uma "alternativa" ao controlo da linguagem feita através de uma ferramenta que o próprio sistema de tradução automática disponibiliza, o *My Dictionary*, onde se propõe uma pequena amostra de exemplos, retomados do *corpus*, que apresentam fenómenos que, no entanto, não cabem no objetivo deste trabalho, como é o caso da ambiguidade lexical.

1.3 ESTRUTURA DA DISSERTAÇÃO

No *capítulo 2* fala-se do âmbito da tradução automática, apresentando uma definição e destacando as suas caraterísticas principais e finalidades. A tradução automática é inserida dentro de uma perspetiva histórica geral como enquadramento desta área, desde os seus inícios até aos nossos dias, passando por períodos de estagnação e de importância cruciais. A seguir, apresentam-se os paradigmas de tradução automática, nomeadamente os orientados para o conhecimento linguístico, os orientados para os dados e os híbridos. Na descrição dos paradigmas de tradução automática orientados para o conhecimento linguístico são abordados os sistemas baseados em dicionários e os sistemas baseados em regras (RBMT, *Rule-based Machine Translation*) e as abordagens *transfer* e interlíngua. A seguir, no âmbito dos sistemas orientados para os dados, é descrita a abordagem baseada em *corpora* e os dois sistemas que a seguem: os sistemas baseados em estatística (SMT*, Statistical-Based Machine Translation*) e os sistemas baseados em exemplos (EBMT, *Example-based Machine Translation*). Por último, são descritos os sistemas híbridos e dá-se uma breve perspetiva sobre as metodologias de hibridização principais. Neste capítulo são abordados também os sistemas de tradução automática online e, em particular, o sistema SYSTRAN, da versão *rule-based* inicial à versão híbrida.

No *capítulo 3* fala-se das linguagens controladas, dentro de uma breve perspetiva histórica, referindo as aplicações. A seguir, são tratados os fenómenos de legibilidade e traduzibilidade, aplicáveis respetivamente às linguagens

controladas orientadas para os humanos (HOCL*, Human-oriented Controlled Languags*) e para as máquinas (MOCL, *Machine-oriented Controlled Language*) à luz da abordagem naturalista e formalista. São abordados também os critérios a seguir para a concepção e o desenvolvimento de uma linguagem controlada, através de duas metodologias diferentes para a criação de regras, proscritivas e prescritivas. Apresentam-se também algumas diretrizes principais para a criação de uma linguagem controlada orientada para uma máquina, que servem de base para a criação do fragmento de português controlado proposto neste trabalho.

No *capítulo 4* apresenta-se o português controlado e a metodologia utilizada na fase de criação das regras, gerais e específicas. Tais regras especificas são criadas para o controlo das variações dos modos verbais em frases finitas, não finitas e imperativas, para o controlo de *dever* e da expressão verbal *ter* + *de* no que diz respeito à modalidade epistémica, deôntica e de capacidade interna. Na secção relativa a tempo e aspeto, propõem-se regras para o controlo do pretérito perfeito simples, para o controlo do progressivo e das expressões verbais *ir* + *gerúndio* e *andar a* + *infinitivo*.

No *capítulo 5* discute-se uma "alternativa" à linguagem controlada para abordar questões como a ambiguidade lexical, a tradução de siglas, sequências e nomes próprios.

No *Anexo* são apresentadas as regras gerais e específicas do fragmento de português controlado para a tradução automática para italiano.

Por último, apresentam-se as *Conclusões*, em que são descritos os desafios encontrados na criação do fragmento de linguagem controlada proposto neste trabalho, remetendo também para futuras aplicações.

2. A TRADUÇÃO AUTOMÁTICA

Neste capítulo são apresentados elementos que servem de enquadramento histórico e teórico à Tradução Automática.

Nesta primeira secção é dada uma definição do conceito de tradução automática referindo-se as finalidades da mesma na divulgação de informação.

Na *secção 2.1*, relativa à perspetiva histórica geral da tradução automática, é feita uma análise dos marcos cruciais para o desenvolvimento da área ao longo do século XX.

Na *secção 2.2*, são descritos os paradigmas de tradução automática orientados para o conhecimento linguístico, nomeadamente os "sistemas" baseados em dicionários e os sistemas baseados em regras (RBMT, *Rule-based Machine Translation*) e as abordagens *transfer* e interlíngua. São abordados também os paradigmas orientados para os dados, ou seja, os sistemas baseados em estatística (SMT, *Statistical-based Machine Translation*) e os sistemas baseados em exemplos (EBMT, *Example-based Machine Translation*), que seguem a abordagem baseada em *corpora*. São descritos também os híbridos, sendo o SYSTRANet o sistema de tradução automática utilizado para a criação do fragmento de linguagem controlada proposto neste trabalho.

Apresentam-se também os sistemas de tradução automática online, que se enquadram em diferentes paradigmas e abordagens.

A seguir, na *secção 2.3,* descreve-se o sistema de tradução automática SYSTRAN, sistema de tradução automática utilizado neste trabalho para a tradução dos exemplos que servem de sustentação empírica à linguagem controlada aqui proposta.

A tradução automática situa-se no cruzamento das ciências da computação, da Linguística e de outras ciências relacionadas com a inteligência artificial. Tem um papel central na sociedade contemporânea, sobretudo por razões sociais e políticas, devido à importância estratégica de áreas do mundo em que há comunidades multilingues e onde a tradução é fundamental para a interação humana. Além disso, a tradução automática tem importância crucial em transações comerciais e em

diversas áreas científicas[1]. Um outro ponto a favor da tradução automática é a rápida expansão da Internet, o meio de comunicação mais utilizado a nível mundial, que faz com que a tradução se torne numa ponte de ligação entre falantes de línguas diferentes (Quah, 2006:89). Neste contexto, a tradução revela-se crucial para a comunicação, sendo que um dos maiores problemas que temos de enfrentar hoje em dia é a impossibilidade de dispor de um tradutor humano em todas as situações em que a tradução se revela necessária. Neste sentido, a tradução automática é uma "alternativa" rápida e económica, sobretudo na tradução em tempo real. Há, contudo, tradutores que têm uma conceção errada sobre a tradução automática, quando a veem como ameaça ao trabalho humano. Neste sentido, é preciso dizer que a tradução automática não vai eliminar o trabalho do homem porque o volume de traduções é bastante elevado, cresce rapidamente e não há tradutores suficientes para satisfazer a procura de forma eficiente. Além disso, é improvável que a tradução automática elimine completamente e a curto prazo o trabalho dos tradutores humanos, dadas as limitações dos sistemas disponíveis hoje em dia no mercado (Arnold *et al.,* 1994:8), e que não se preveem totalmente ultrapassáveis. Os obstáculos nesta área são também de carácter linguístico, como a ambiguidade (lexical e estrutural), a complexidade sintática e as estruturas agramaticais. Tal coloca problemas no que diz respeito à extração do significado, o que consequentemente cria problemas nos textos de *output*. Assim sendo, é preciso intervenção humana no controlo do *input* para se obter um *output* aceitável, com recurso ao uso das chamadas sublínguas e de linguagens controladas.

As finalidades da tradução automática são variadas, graças a uma procura cada vez maior de traduções de tipo técnico e científico ou de traduções de manuais de instruções. Esta lista não é exaustiva, e como há um volume elevado de traduções em muitas línguas e os tradutores não conseguem satisfazer a procura, a tradução automática adquire um papel central na divulgação de informação, nomeadamente na sua disseminação, assimilação e troca. Na disseminação de informação, as traduções devem ser de "alta" qualidade, no sentido em que a maior parte da informação do texto de partida é transferida para o texto de chegada e as traduções são publicáveis. O conceito de qualidade é subjetivo e varia em termos de

[1] Terá, ainda, importância filosófica, pois é uma tentativa de automatizar uma atividade que requer reflexão sobre as áreas em que é preciso o conhecimento humano do mundo (Arnold *et al.,* 1994:5).

fidelidade, adequação, inteligibilidade, estilo e registo. O *output* produzido é "imperfeito" e pode ser preciso um controlo do texto de partida com a restrição do *input* através de sublínguas e linguagens controladas. No caso das sublínguas é utilizado o vocabulário específico do domínio ao qual pertence a tradução em questão. Por outro lado, no caso das linguagens controladas, em termos gerais, opera-se um controlo sobre o léxico de modo a reduzir as ambiguidades e sobre aspetos estruturais problemáticos, entre os quais a ambiguidade estrutural. Na assimilação, a tradução é utilizada para permitir uma compreensão rápida do texto, sem objetivos de publicação. Com a "explosão" da informação na segunda metade do século XX, jornalistas, analistas e investigadores precisam cada vez mais de informações disponíveis apenas em outras línguas que não dominam e para este efeito basta que os textos traduzidos sejam "compreensíveis" e não "perfeitos" do ponto de vista linguístico. Com a difusão dos computadores, este é o tipo de tradução mais procurado. A última categoria da divulgação de informação é a troca de informações, associada ao conceito de "acesso à informação". São traduções económicas e rápidas, que podem ser feitas para sites, blogues e na tradução em tempo real de conversas online (*chat room*), através da extração de informação diretamente de bancos de dados de escrita e de fala, disponíveis na Internet. Respondem à necessidade de ter uma tradução rápida, ainda que não seja "perfeita", mas capaz de comunicar o conteúdo do texto na língua de partida. É nesta área que se coloca o foco da investigação na tradução automática hoje em dia.

2.1 Perspetiva histórica geral da tradução automática

Nesta secção é apresentada uma perspetiva histórica geral da tradução automática, desde os seus inícios até aos anos 2000, que serve de enquadramento para o estudo dos paradigmas de tradução automática e das diferentes abordagens existentes hoje em dia.

Os primeiros passos na área da tradução automática foram dados no início dos anos 30 do século XX, com a criação do *Cerveau Mécanique* do franco-arménio Georges Artsruni e do primeiro tradutor automático criado pelo russo Petr Troyanskii. De particular importância é o tradutor automático criado pelo investigador russo, que propõe um dicionário bilingue e um esquema, baseado no esperanto, para a codificação gramatical, a análise e a síntese linguística, que hoje tem na base os conceitos de *transfer* e de interlíngua. Este tradutor pode ser

considerado o precursor dos atuais tradutores automáticos porque a tradução era efetuada através de três etapas: a pré-edição do texto a traduzir, em que um falante nativo da língua de partida analisava e separava as unidades linguísticas, a etapa mecânica, em que a máquina encontrava as correspondências entre as unidades linguísticas do texto de partida e do texto de chegada, e a última era a da pós-edição dos textos, corrigidos por parte de um falante nativo da língua de chegada. Pode dizer-se que as primeiras abordagens à tradução automática eram de tipo direto, ou seja, de tradução palavra a palavra. Os progressos da investigação nesta área pararam durante uns anos devido à Segunda Guerra Mundial, tendo as ideias de Artsruni e Troyanskii sido abandonadas até à segunda metade dos anos 40, época em que surgiram os primeiros computadores.

O período entre a segunda metade dos anos 40 e os primeiros anos da década de 50 é considerado o início da tradução automática que conhecemos hoje. O impulsionador é o matemático americano Warren Weaver, que entrou em contacto com o Professor Norbert Wiener do Massachusetts Institute of Technology (MIT) e com Andrew Donald Booth para pedir a opinião de um informático e de um linguista sobre a criação de um sistema de tradução automática. Os resultados obtidos ficaram longe do esperado e, por isso, Warren Weaver em 1949 divulgou um memorando, *Translation*[2], no qual explicava o interesse e a necessidade da tradução automática na tradução de textos técnicos e científicos, com o objetivo de difundir a ideia de que o futuro da tradução encontrava-se na tradução automática.

Nos anos seguintes, no período da Guerra Fria até à segunda metade dos anos 60, o interesse dos Estados Unidos focou-se na tradução de documentos que vinham da União Soviética e a investigação na área da tradução automática orientou-se para a criação de sistemas de tradução automática bilingue (nomeadamente entre russo e inglês). Eram "sistemas" diretos, ou seja, de tradução palavra a palavra, efetuada através de dicionários bilingues, sem análise sintática ou lexical e, por isso, a qualidade do *output* na maioria dos casos não era aceitável. Nesta altura, dado o objetivo das traduções feitas através de sistemas automáticos, o que interessava não era a qualidade, mas o conteúdo do texto. Nesta primeira fase é preciso falar de tradução mecânica e não de tradução automática propriamente dita, dada a abordagem direta utilizada nesta altura. Fora dos Estados Unidos, a

[2] Publicado em Locke e Booth (1955).

pesquisa continuou em particular na União Soviética e na Europa, com o objetivo de criar um tradutor completamente automático.

Em 1951, Bar-Hillel do MIT declara a impossibilidade de uma tradução completamente automática e foi um dos primeiros a favor da tradução mista, feita através de um sistema automático coadjuvada por um tradutor humano, dada a pouca qualidade dos *outputs* produzidos. Em 1952, precisamente no MIT, teve lugar a Primeira Conferência sobre a Tradução Automática e Bar-Hillel falou pela primeira vez do controlo da gramática do texto de partida na tradução automática, ideia que nas décadas seguintes vai dar início às linguagens controladas para a tradução automática que conhecemos hoje em dia. A ideia que Bar-Hillel defendia era a da intervenção humana nas fases de pré-edição e pós-edição dos textos. É nesta altura que surgem também as primeiras críticas à tradução automática.

O ano que trava a pesquisa na área da tradução automática nos Estados Unidos é o 1966, ano do Relatório ALPAC (*Automatic Language Advisory Committee*). A comissão analisa os resultados da tradução automática e dá relevância a três pontos principais: a qualidade, a rapidez e os custos das traduções. O que emerge do Relatório é:

> "The Committee believes strongly that the quality of translation must be adequate to the needs of the requester. The production of a flawless and polished translation for a user-limited readership is wasteful of both time and money. On the other hand, production of an inferior translation when one of archival quality is called for is even more wasteful of resources. It seems clear to the Committee that, in many cases, translations of adequate quality are not being provided" (ALPAC, 1966:16).

No que concerne aos custos das traduções afirmou-se:

> "Cost is important because in many cases it is the only measure the government can sensibly use in deciding how its translation is to be done. As we have seen, it varies considerably—from $9 to $66 per 1.000 words. Machines are probably inappropriate for some forms of translations, such as very high quality diplomatic translation and literary translation. But translations of scientific material can be done with or without machine aids. As to quality and speed, at extra cost, better quality and higher speed can be attained if long texts are split into segments. Thus, cost for a particular result is the criterion that the government should apply in deciding on means of translation" (ALPAC 1966:17-18).

O Relatório acrescentava também que os atrasos na entrega das traduções eram o resultado da demora nas fases de pré e pós edição dos textos e termina num tom pessimista quanto às perspetivas de futuro para a tradução automática:

> " "Machine Translation" presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing. In this context, there has been no machine translation of general scientific text, and none is in immediate prospect" (ALPAC, 1966:19).

O Relatório ALPAC leva ao fim de muitos projetos de tradução automática, sobretudo nos Estados Unidos.

Os anos 70 são considerados um período de estagnação, mas a investigação na área da tradução automática continua no Canadá, na Comunidade Económica Europeia (CEE) e no Japão e há ainda alguns grupos de investigação americanos que trabalham na combinação linguística russo-inglês. A CEE e o governo do Canadá tinham a necessidade de investir em várias combinações linguísticas, dada a natureza multicultural e multilingue destes territórios. É neste contexto que o governo canadiano financia o projeto METEO, para a tradução do inglês para francês de informações meteorológicas, que se tornou operativo em 1977. O METEO é definido como o primeiro sistema de tradução completamente automático. Uns anos depois (em 1986, mais precisamente), a Comunidade Económica Europeia comprou o sistema de tradução automática SYSTRAN, criado nos Estados Unidos em 1968 por Peter Toma e utilizado inicialmente pela Força Aérea americana na tradução russo-inglês durante a Guerra Fria. Este período é considerado um ponto de viragem na história da tradução automática, graças a uma gradual confiança na investigação nesta área. No Japão e na Europa avançaram os estudos sobre o *transfer* como abordagem aos sistemas de tradução automática baseados em regras (RBMT, *Rule-based Machine Translation*). Acrescenta-se que nos anos entre 1970 e 1980 aumentou o interesse na tradução automática multilingue no sector técnico e comercial.

Nos anos 80 a tradução automática foi utilizada em vários países do mundo e aumentaram também as combinações linguísticas. Estes são os anos dos grandes projetos de tradução automática, como o EUROTRA, financiado pela Comunidade Económica Europeia. Neste período desenvolveu-se uma nova abordagem, a

interlíngua, baseada no conhecimento linguístico, para a criação de novos sistemas de tradução automática. São os anos dos *Knowledge-based Systems* (KBMT, sistemas baseados no conhecimento), que abriram o caminho para a ideia de que uma tradução de alta qualidade só pode ser feita através de uma plena compreensão do texto por parte do sistema de tradução automática, uma ideia presente, por exemplo, em Arnold *et al.,* (1994). Nesta década surge também a segunda geração de sistemas de tradução automática baseados em regras, os *Constraint-based Systems,* uma vez que as sequências de regras utilizadas na primeira geração eram excessivamente complicadas. Com os sistemas *Constraint-based* há uma simplificação das regras de análise, *transfer* e geração de (chamada também *síntese*). Os anos 80 foram um período crucial na história da tradução também porque no Japão surgiram os primeiros softwares comerciais de TA para computadores vendidos a nível mundial e, além disso, apareceram também as primeiras ferramentas de apoio à tradução. As empresas multinacionais começaram a investir na investigação no sector da tradução automática e, nestes anos, foram experimentados também os primeiros modelos de linguagem controlada utilizados pelas multinacionais e aplicados à tradução automática. Foi no final dos anos 80 que cessou o domínio dos sistemas baseados em regras, graças a novos métodos e estratégias impulsionados por exigências cada vez mais diversificadas. Em 1988 nasceu o primeiro sistema de tradução automática baseado em estatística e, paralelamente, foram criados no Japão os primeiros sistemas de tradução automática baseados em exemplos (EBMT, *Example-based Machine Translation*).

No início dos anos 90, os grupos de investigação japoneses concentraram-se na combinação entre sistemas baseados em *corpora* e sistemas baseados em regras, concebendo assim os primeiros sistemas de tradução híbridos. Em 1992 foi dado um ulterior impulso ao uso das linguagens controladas com o desenvolvimento de um sistema de tradução automática baseado em conhecimento linguístico combinado com o controlo dos *inputs* para a tradução multilingue de manuais técnicos (Nirenberg *et al.,* 1992). São os anos da globalização, anos em que aumentaram as vendas de *CAT Tools* (ferramentas de tradução assistida por computador) e a difusão gratuita dos primeiros tradutores online, como o Babelfish e o Google Translate.

Do início dos anos 2000 até hoje houve uma grande difusão a nível mundial sobretudo dos sistemas estatísticos, graças à elevada disponibilidade de *corpora*

paralelos, de ferramentas que podem ser encontradas gratuitamente online de alinhamento de textos, mas também de outras ferramentas computacionais para o processamento das línguas naturais. O tradutor automático mais utilizado atualmente é o Google Translate, sistema de tradução automática estatístico. Nestes últimos anos, aumentou o recurso à tradução automática, devido não só à necessidade de traduções multilingues por parte de várias empresas presentes em mercados multinacionais, mas também graças ao crescimento acelerado do uso da internet e das redes sociais. Além disso, procuram-se cada vez mais outras combinações linguísticas dada a importância crescente de línguas como o japonês, o chinês e o árabe. É na implementação destes pares linguísticos que muitos grupos de investigação estão atualmente a trabalhar.

2.2 PARADIGMAS DE TRADUÇÃO AUTOMÁTICA

Em termos gerais, um paradigma de tradução automática é um modelo de tradução que apresenta algumas caraterísticas, num certo sentido consideradas "revolucionarias", que definem o sistema de tradução e que permitem a classificação de outros sub-paradigmas. Há dois elementos cruciais que permitem a classificação dos paradigmas, e que são o recurso principal que o sistema de tradução utiliza: o conhecimento linguístico e os dados. Além disso, existem também hibridações destes dois paradigmas e, por isso, optou-se por classificar os paradigmas de tradução em três categorias:

1. Paradigmas orientados para o conhecimento linguístico;
2. Paradigmas orientados para os dados;
3. Paradigmas híbridos que combinam sistemas orientados para o conhecimento linguístico com sistemas orientados para os dados.

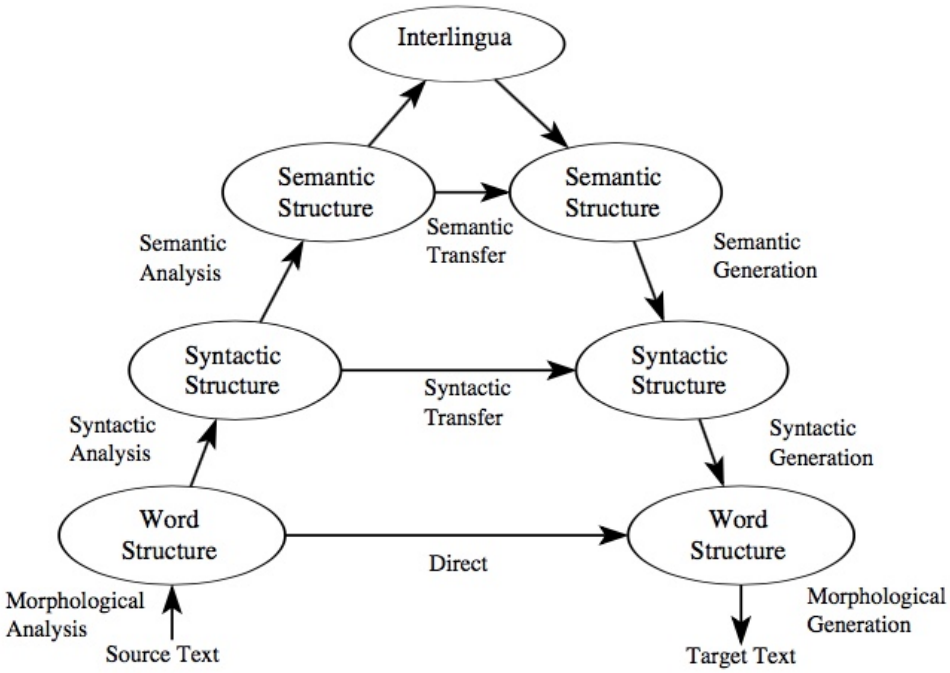
Esta classificação foi discutida pela primeira vez em 1992, durante a quarta edição da TMI (*The International Conference on Theoretical and Methodological Issues in Machine Translation*) em que os investigadores se dividiram em dois grupos, o da abordagem linguística à tradução automática e o da abordagem não linguística. Escolheu-se utilizar também a categoria dos paradigmas híbridos porque, hoje em dia, uma das áreas de investigação da TA é a dos sistemas que integram

características próprias dos sistemas orientados para o conhecimento linguístico e outras próprias dos sistemas orientados para os dados.

2.2.1 PARADIGMAS DE TRADUÇÃO AUTOMÁTICA ORIENTADOS PARA O CONHECIMENTO LINGUÍSTICO

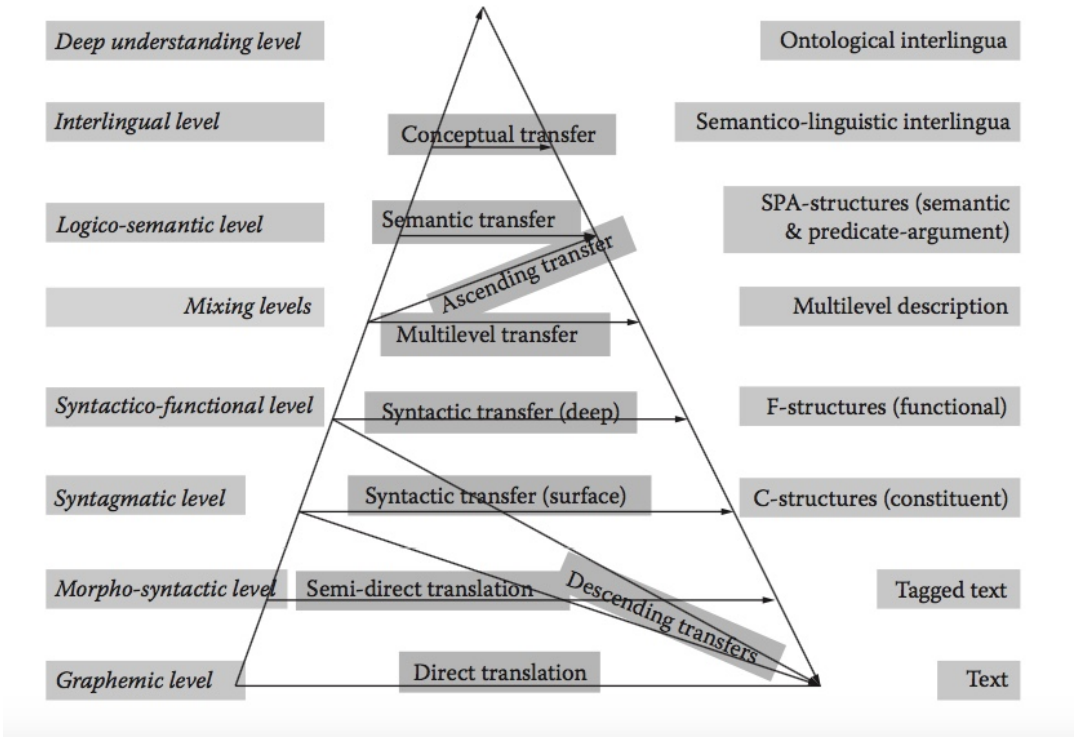
Os paradigmas de tradução automática orientados para o conhecimento linguístico têm fundamento nas investigações sobre a teoria linguística e utilizam restrições sintáticas, semânticas e lexicais para produzir traduções apropriadas na língua de chegada. Nas secções a seguir são apresentados os sistemas de tradução automática baseados em regras (RBMT, *Rule-based Machine Translation*), que hoje em dia constituem a abordagem mais utilizada apesar da crescente importância dos paradigmas orientados para os dados e dos híbridos. Mencionam-se também os sistemas baseados no léxico (LBMT, *Lexical-based MachineTtranslation*) e os sistemas baseados no conhecimento linguístico (KBMT, *Knowledge-based Machine Translation*).

No que diz respeito às abordagens à tradução automática dos paradigmas orientados para o conhecimento linguístico, fala-se de *transfer* e interlíngua, tratados na *subsecção 2.2.1*. É oportuno falar também da tradução direta, seguida nos primeiros "sistemas" de tradução automática, ou seja, os "sistemas" baseados em dicionários. Como ferramenta para a descrição da tradução direta e, mais a frente, do *transfer* e da interlíngua, introduz-se o Triângulo de Vauquois (1968):



Quadro 1. Versão simplificada do Triângulo de Vauquois (Extraído de Dorr *et al*., 1999).

Em termos gerais, o processo de tradução depende do nível de profundidade da análise: no lado esquerdo é representada a análise do texto de partida e no lado direito a geração do texto de chegada. A base do Triângulo representa os sistemas que operam só a análise morfológica, ou seja, a nível da palavra, do texto de partida para a geração morfológica do texto de chegada. Prosseguindo, na parte central são representados os sistemas que operam uma análise mais profunda das estruturas do texto de chegada: sintática e semântica. Estes sistemas, depois de ter efetuado esta análise e através do *transfer* sintático e semântico, geram o texto de chegada. Por último, o topo do Triângulo representa uma outra abordagem - interlíngua -, que utiliza uma representação indepentente à língua de partida e de chegada para a produção da tradução. O Quadro 1. mostra uma versão simplificada dos níveis de análise e, portanto, é importante mencionar a versão do Triângulo proposta por Bhattacharyya (2015):



Quadro 1.1. Triângulo de Vauquois (Extraído de Bhattacharyya, 2015:5).

Em conclusão e, observando também o Triângulo do Quadro 1.1., pode-se dizer que a níveis de análise mais profundos correspondem traduções melhores. Estas foram considerações gerais e preliminares e as três abordagens acima mencionadas são tratadas a seguir.

TRADUÇÃO DIRETA. A tradução direta não é considerada uma abordagem à tradução automática em sentido estrito, porque não há fases de análises (além da análise morfológica) e as traduções são do tipo "palavra a palavra". Estes "sistemas" são os "sistemas" baseados em dicionários (*Dictionary-based Machine Translation)* e são os mais primitivos (utilizados até ao fim dos anos 60) e a tradução é unidirecional e bilingue.

Na base do Triangulo (Quadro 1.) é ilustrada a tradução direta. Os "sistemas" baseados em dicionários traduzem sequências de palavras através da introdução de dicionários e são programados para a tradução de apenas um par linguístico, através da tradução direta do texto de partida no texto de chegada. São compostos por um dicionário bilingue, um *parser* que determina a estrutura gramatical do texto de partida, um programa de produção de textos que opera

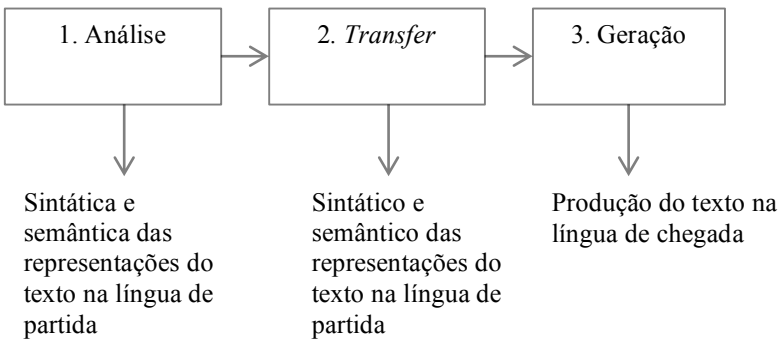
através do dicionário bilingue e uma gramática da língua de chegada para obter o texto final. A análise do texto de partida é estritamente morfológica, ao nível da palavra, portanto. Basicamente, o sistema substitui as sequências de palavras do texto de partida pelas sequências de palavras do texto de chegada, mesmo nos casos em que as duas línguas tenham estruturas diferentes. Por esta razão, a tarefa de tradução corresponde a uma única operação em que o dicionário existente no sistema armazena a informação linguística necessária sem o recurso a outros módulos. Como é evidente, estes "sistemas" não operam nenhuma análise semântica (ou muito raramente) e a análise sintática é básica. Só em alguns "sistemas" são integrados módulos para a reordenação sintática do texto na língua de chegada e, quando o "sistema" não tem módulos para reordenação, a leitura do texto de chegada resulta bastante complicada. São "sistemas" que requerem uma mínima informação linguística, e como os seus criadores são geralmente informáticos, torna-se impossível resolver problemas complexos como o da ambiguidade. Os únicos resultados positivos são obtidos apenas na tradução direta de duas línguas que podem ser consideradas "próximas" (como pode ser o caso da tradução da combinação linguística português-italiano), uma vez que os problemas relativos à ambiguidade estrutural e à ordem dos elementos da frase são mínimos. Ainda assim, os resultados estão longe do esperado. Esta abordagem apresenta vários problemas, como a falta de informação linguística, dificuldades na resolução de ambiguidades, dificuldades na leitura das traduções e a falta de uma abordagem de tipo modular: na introdução de novas entradas nos dicionários, os "sistemas" tornam-se poucos estáveis e a tradução resulta bastante perturbada. Além disso, os processos de tradução eram muito longos, as traduções de baixa qualidade e a manutenção dos sistemas muito cara. Estes foram os "sistemas" duramente criticados pelo relatório ALPAC de 1966 e esta abordagem foi lentamente abandonada nos anos a seguir, sobretudo na Europa e no Japão. Para concluir, resta mencionar que a abordagem direta era utilizada nas primeiras versões do sistema de tradução automática SYSTRAN na tradução bilingue russo-inglês e no projeto METEO no Canadá, na tradução da combinação inglês-francês de boletins meteorológicos.

RBMT: SISTEMAS DE TRADUÇÃO AUTOMÁTICA BASEADOS EM REGRAS. A ideia de base destes sistemas é a representação do conhecimento linguístico através de regras implementadas nos sistemas e

> "[...] success in practical MT involves defining a level of representation for texts which is abstract enough to make translation itself straightforward, but which is at the same time superficial enough to permit sentences in the various source and target languages to be successfully mapped into that level of representation" (Arnold *et al*., 1994:174).

O objetivo destes sistemas é converter as estruturas da língua de partida nas estruturas da língua de chegada, seguindo duas abordagens diferentes: *transfer* e interlíngua. Na análise destas abordagens é oportuno considerar o Triângulo de Vauquois[3]. Em termos gerais, o *transfer*, é operado a dois níveis, semântico e sintático. No vértice superior é representada a abordagem interlíngua.

TRANSFER. Os sistemas que utilizam o *transfer* como abordagem são os sistemas baseados em regras (RBMT) de segunda geração. O objetivo principal destes sistemas era obter um texto de chegada correto do ponto de vista sintático, transformando as representações do texto de partida em representações sintáticas próprias do texto de chegada. Este processo é definido por Bhattacharyya (2015) como processo ATG (análise-*transfer*-geração) e consta nas fases de análise do texto na língua de partida, na fase do *transfer* sintático e semântico, e na fase de geração em que se gera o texto na língua de chegada, como é possível observar no quadro a seguir:



Quadro 2. Processo ATG.

[3] *Cf*. Quadro 1.

Simplificando, a fase de análise acontece no lado esquerdo do Triângulo de Vauquois (Quadro 1.), a seguir o *input* é transferido na parte central em que é operado o *transfer* sintático e semântico. Acrescente-se que os sistemas que utilizam só o *transfer* sintático produzem traduções de qualidade inferior em comparação com os sistemas que integram também o *transfer* semântico, porque os níveis de análise não operam em profundidade. Por isso, usam-se as duas tipologias nas versões atuais dos sistemas de tradução automática que utilizam o *transfer.* O processo conclui-se com o *input* que chega ao lado direito do Triângulo, no qual é produzido o texto através de um dicionário da língua de chegada. Basicamente, na fase central do processo há regras de mapeamento entre a língua de partida e a língua de chegada, as quais operam desde a "superfície" do texto de partida e de chegada até às estruturas e às representações mais "profundas". Pode-se dizer que esta abordagem utiliza o conhecimento contrastivo das duas línguas em causa e cada fase do processo emprega dicionários específicos, nomeadamente o dicionário da língua de partida para a fase de análise, um dicionário bilingue na fase de *transfer* e um dicionário da língua de chegada para a produção do texto na fase de geração.

As traduções efetuadas por sistemas baseados no *transfer* produzem boas traduções se as regras forem completas e se o léxico bilingue cobrir o domínio de interesse. São sistemas capazes de resolver alguns dos problemas de ambiguidade do texto graças à análise sintática, a qual reconhece a categoria lexical das palavras do texto de partida. Por outro lado, estes sistemas utilizam regras complexas que variam em relação ao par linguístico utilizado, e por isso

> "A large set of transfer rules must be constructed for each source-language/target-language pair; a translation system that accommodates n languages requires n^2 sets of transfer rules" (Dorr *et al*., 1999:15).

Pode acontecer também que as regras não sejam completas e que não sejam suficientes para resolver os problemas de ambiguidade. Por isso, podem ocorrer erros na fase de análise, com consequente falta da fase de *transfer* e sem que se gere a tradução. Um outro ponto fraco é a análise pragmática, que juntamente com a análise sintática e semântica permitiria que o sistema produzisse traduções excelentes.

INTERLÍNGUA. As abordagens interlíngua são utilizadas na terceira geração dos sistemas de tradução automática baseados em regras (RBMT). A ideia central destes sistemas reside na capacidade de captar a representação do texto de chegada a partir da análise do texto de partida, independentemente da língua. De acordo com Jurafsky e Martin (2000), a interlíngua funciona como "intermediário" entre as línguas naturais, porque

> "An "interlingua" represents all sentences that mean the "same" thing in the same way, regardless of the language they happen to be in" (Jurafsky e Martin, 2000:812).

Dado que é uma representação comum às línguas naturais, não há a fase de *transfer* nem regras de *transfer* e a produção do texto passa apenas por duas fases: a análise (do texto de partida) e a geração (do texto de chegada). Basicamente, a interlíngua é uma representação "neutra" da linguagem. Como no *transfer*, também na interlíngua há regras de mapeamento entre as línguas, mas na interlíngua chega-se a um nível de representação que serve qualquer língua, sendo um nível de representação conceptual. Como já referido anteriormente, o processo consta de duas fases, a análise do texto de partida e a geração do texto de chegada:

Quadro 3. Processo interlíngua.

O Quadro 3. mostra as fases do processo de tradução que de forma geral pode ser assim resumido: na primeira fase, o texto na língua de partida é traduzido para a interlíngua e, na segunda fase, a interlíngua é traduzida para o texto de chegada. Na fase de análise são representados o conhecimento lexical, estrutural e discursivo do

texto de partida de maneira não ambígua, para que a representação interlíngua possa reagrupar as palavras na própria forma desambiguada para formar, sucessivamente, grupos de palavras chamados *multiwords*. Nesta fase há também a resolução de ambiguidades de tipo semântico e discursivo. Deste modo, o texto na língua de chegada é gerado pela interlíngua, neutra a respeito da língua de partida. Cabe acrescentar que, sendo a interlíngua independente das línguas, tem importância crucial sobretudo na tradução multilingue.

Os sistemas que se baseiam na interlíngua são modulares, o que significa que lhes possam ser adicionados módulos sem afetar e modificar as outras regras existentes, garantindo que não ocorram problemas de tradução. Por outras palavras, a adição de novos módulos aos sistemas modulares faz com que o sistema mantenha a sua estabilidade sem afetar as traduções.

Uma abordagem mais avançada de interlíngua é utilizada nos sistemas baseados no conhecimento linguístico (KBMT[4]), os quais têm conhecimento pragmático e semântico mais extensivo e têm a capacidade de "raciocinar" sobre conceitos diferentes (Quah, 2006:72). No estado da arte atual, os investigadores estão a estudar sistemas interlíngua que sejam adequados também para a tradução entre línguas que pertencem a grupos linguísticos diferentes, como por exemplo entre línguas asiáticas e línguas europeias.

A vantagem da abordagem interlíngua encontra-se no facto de a geração do texto de chegada ser dependente da representação universal comum às línguas da mesma "família". Isto quer dizer que o sistema não precisa de regras diferentes para a tradução de diferentes combinações linguísticas e as fases de análise e síntese (ou seja, de geração de textos) acontecem só uma vez e servem para línguas diferentes. No entanto, estes sistemas comportam também algumas desvantagens, nomeadamente a dependência da representação interlíngua em relação à sintaxe do texto de partida. A geração, por isso, é feita através desta representação e muitas vezes tem a forma de uma paráfrase e não de uma tradução (Dorr *et al.,* 1999).

[4] Abordagem *Knowledge-based* do sistema da Carneige Mellon University, o qual utiliza regras de mapeamento lexical e gramatical, conhecido como sistema KANT (*Knowledge-based Accurate Translation*).

Voltando aos sistemas RBMT e de acordo com Bhattacharyya (2015), nos sistemas RBMT que seguem o *transfer*, as regras são criadas por linguistas para cada etapa do processo ATG (análise-*transfer*-geração). Na fase de análise encontramos, entre outras, regras de análise morfológica, de *parsing* e de geração semântica para a resolução de diferentes tipos de ambiguidade. Na fase de *transfer,* o sistema utiliza um dicionário bilingue para o mapeamento de palavras e frases. Para concluir, na geração o sistema encontra a síntese morfológica das entradas do dicionário e opera uma ordenação sintática.

Nos sistemas baseados na interlíngua, a tradução é efetuada operando da base do Triângulo até à ponta para descer até à produção do texto na língua de chegada. Isto quer dizer que uma tradução puramente baseada na interlíngua não existe. A análise da língua de partida, por consequência, produz representações que funcionam também na língua de chegada. É esta representação que tem o nome de interlíngua e é suposto ser a representação comum a todas as línguas naturais.

Nos sistemas baseados no *transfer*, a análise acaba na fase intermédia representada no Triângulo e o nível de análise-*transfer*-geração é específico às línguas envolvidas na tradução, ao contrário do que acontece na interlíngua, que pode ser utilizada como representação comum a todas as línguas naturais. As regras, neste caso, são regras de *transfer* que têm de ser aplicadas na tradução de duas línguas que têm estruturas diferentes. É também necessário mencionar a diferença crucial entre sistemas baseados no *transfer* e sistemas baseados na interlíngua:

> "Transfer-based MT does not insist on complete disambiguation of the source sentence, interlingua-based MT does not have any transfer stage" (Bhattacharyya, 2015:177).

Para concluir a descrição dos paradigmas de tradução automática orientados para o conhecimento linguístico, resta ainda referir outros dois paradigmas, sobreponíveis com os RBMT: os baseados no léxico (LBMT) e os baseados no conhecimento linguístico (KBMT[5]). Nos primeiros, os sistemas são equipados com regras que ligam as entradas lexicais da língua de partida às entradas lexicais da

[5] A análise aprofundada dos sistemas LBMT e KBMT não cabe nos objetivos deste trabalho. Para uma leitura sobra a matéria veja-se Dorr *et al*., (1999), entre outros.

língua de chegada. Por outro lado, os sistemas KBMT concentram-se na veiculação da informação morfológica, semântica e sintática no léxico.

2.2.2 PARADIGMAS DE TRADUÇÃO AUTOMÁTICA ORIENTADOS PARA OS DADOS

Os paradigmas de tradução automática orientados para os dados, em termos gerais, não utilizam regras de mapeamento entre a língua de partida e de chegada e por isso não pressupõem conhecimento linguístico. As traduções são geradas através do uso de dados, entendidos neste caso como "material linguístico", algoritmos e cálculo da probabilidade. Nesta secção são apresentados os sistemas que se baseiam em *corpora*: os sistemas de tradução automática estatísticos (SMT, S*tatistical-based Machine Translation*) e os sistemas de tradução automática baseados em exemplos (EBMT, E*xample-based Machine Translation*).

ABORDAGEM BASEADA EM *CORPORA*. Esta abordagem é chamada *Corpus-based Approach*, ou seja, abordagem baseada em *corpora*. Os *corpora* (chamados também textos paralelos, *bitexts* ou *multitexts*) são constituídos por textos paralelos e traduções já existentes e podem ser bilingues ou multilingues. O conjunto de *corpora* constitui a base de textos utilizada pelo sistema de tradução automática. É preciso acrescentar que a utilidade dos *corpora* depende do estado em que se encontram disponíveis para o investigador e muitas vezes é necessário um processo de correção de erros, o que pode envolver custos muito elevados e de algum modo contaminar a "pureza" dos dados. Em alguns casos, após a correção dos erros, pode haver algumas divergências entre *corpora,* o que por sua vez poderá afetar os cálculos estatísticos na fase de alinhamento dos textos. Para concluir esta descrição preliminar, é importante mencionar que a tradução é vista neste caso como um *machine learning problem* (Lopez, 2008), porque consiste basicamente na aplicação de um algoritmo para a tradução de textos anteriormente traduzidos de modo a que a máquina seja capaz de traduzir outros textos.

Todos os paradigmas de tradução automática orientados para os dados seguem a abordagem baseada em *corpora* e a diferença está na aplicação de metodologias diferentes. Os sistemas SMT utilizam *corpora* e puros cálculos probabilísticos para a produção de traduções e, por outro lado, os sistemas EBMT utilizam os *corpora* para a extração de exemplos, ou seja fragmentos de frases, para

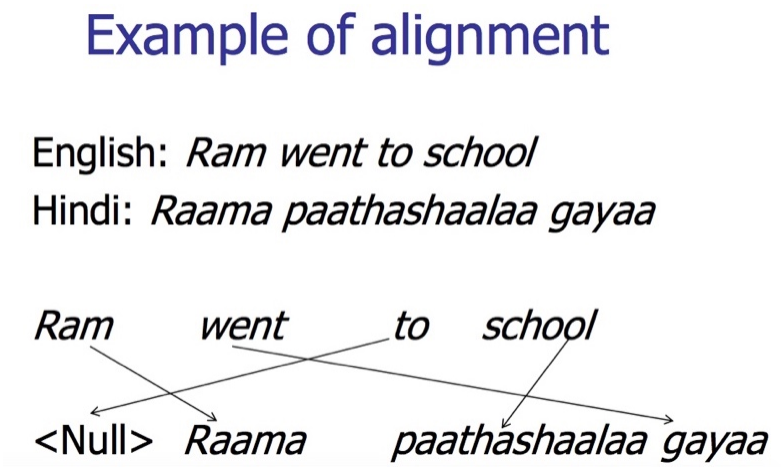
construir as memórias de tradução. A tradução, neste caso, não é produzida através de calculo probabilísticos, mas através de cálculos para a frequência de tradução.

SMT: SISTEMAS DE TRADUÇÃO AUTOMÁTICA BASEADOS EM ESTATÍSTICA. O interesse nos sistemas de tradução automática baseados em estatística começou a crescer com o processo de disseminação da informação em várias línguas através da internet aumentando, deste modo, o acesso a *corpora* bilingues e multilingues de textos e traduções. Um outro fator relevante é o interesse cada vez maior em informações escritas noutras línguas por parte de consumidores e investigadores, ou seja, o chamado processo de assimilação. Além disso, os SMT tornaram-se populares por não pressuporem nenhum tipo de conhecimento linguístico. Os SMT não são adaptados para um só par linguístico, dado que podem ser utilizados em todas as combinações linguísticas presentes nos *corpora* e as regras não têm de ser implementadas no sistema por parte de um especialista. Para dar uma definição melhor de SMT,

> "The term SMT can be understood in a narrow sense to refer to approaches which try to do away with explicitly formulating linguistic knowledge, or in a broad sense, to denote the application of statistically or probabilistically based techniques to parts of the MT tasks" (Arnold *et al*., 1994:190).

Os primeiros sistemas foram utilizados pela primeira vez em 1988 pela IBM no Parlamento Canadiano (The Canadian Hansard) na tradução de legislação através de um *corpus* bilingue francês-inglês.

No que respeita as fases do processo de tradução destes sistemas, nos SMT a tradução depende, em grande medida, de cálculos estatísticos que se baseiam em dois modelos: o modelo de tradução e o modelo da língua alvo. No primeiro, calcula-se a probabilidade da correspondência das palavras entre o texto de partida e o texto de chegada, no segundo calcula-se a probabilidade de as palavras estarem corretamente combinadas na língua de chegada e a sequência de palavras mais provável. Basicamente, o processo de tradução desenvolve-se em três fases: alinhamento, cálculo das correspondências e reordenação. No alinhamento, as frases, palavras e grupos de palavras são alinhados para encontrar as correspondências, como mostra o quadro a seguir:



Quadro 4. Exemplo de alinhamento estatístico (Extraído de Bhattacharyya, 2014:44).

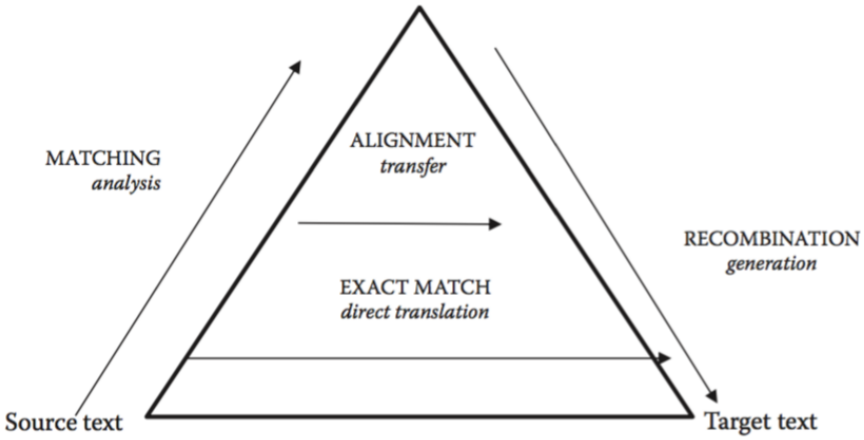
É preciso introduzir mais duas noções nesta etapa, a fertilidade e a distorção. A fertilidade de uma palavra do texto α do *corpus* é o numero de palavras que lhe correspondem no texto β. Já a distorção refere-se ao facto de a palavra do texto α e as que lhe correspondem no texto β não aparecerem na mesma posição (Quadro 4.). Por isso, os parâmetros que têm de ser calculados na fase de alinhamento são a probabilidade da fertilidade de cada palavra do texto α, as possibilidades de tradução dos pares de palavras e a probabilidade de distorção. Depois do alinhamento dos textos, são calculadas as correspondências entre palavras através da aplicação de algoritmos e cálculos probabilísticos. Na última fase, a da reordenação, são aplicados mecanismos que ordenam a frase segundo a estrutura sintática da língua de chegada para obter uma tradução correta.

EBMT: SISTEMAS DE TRADUÇÃO AUTOMÁTICA BASEADOS EM EXEMPLOS. O sistema aqui apresentado é o chamado EBMT, baseado em exemplos (*Example-based Machine Translation*, mas também *Case-based Machine Translation* e *Memory-based Machine Translation*). O desenvolvimento dos sistemas baseados em exemplos procurou dar resposta ao problema da procura contínua de correspondência entre palavras e termos que ocupa boa parte dos esforços da tradução humana, com o objetivo de encontrar a opção melhor entre língua de partida e língua de chegada, recorrendo a textos previamente traduzidos por outros tradutores. É esta a ideia central de Nagao (1984:173-180) no famoso *Translation*

by Analogy. Assim nasceram os primeiros EBMT, como alternativa aos métodos tradicionais RBMT (Hutchins e Somers, 1992:125). A tradução consiste em reconhecer a correspondência entre uma frase na língua de partida e a correspondente tradução contida num texto previamente traduzido, pratica chamada "Tradução por Analogia",

> "Our fundamental ideas about the translation are: (1) Man does not translate a simple sentence by doing deep linguistic analysis, rather, (2) Man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases (very often, into case frame units), then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference" (Nagao, 1984:178).

TRADUÇÃO POR ANALOGIA. Nesta metodologia, as regras de mapeamento entre língua de partida e língua de chegada são dispensadas em favor do processo de correspondência (*matching*) entre os exemplos armazenados nas memórias de tradução. A ideia central nos sistemas EBMT está em encontrar, através de um algoritmo para o alinhamento, a tradução mais próxima ao exemplo a traduzir. Graças a este fenómeno obtém-se um *template* de tradução que será melhorado através da tradução palavra a palavra. É evidente que o desempenho do sistema, como no caso dos SMT, depende de uma aplicação correta dos algoritmos para o alinhamento.



Quadro 5. Triângulo de Vauquois adaptado aos sistemas EBMT (Extraído de Bhattcharyya, 2015:195).

O Quadro 5. demonstra que é possível adaptar o Triângulo de Vauquois (Quadro 1.) aos sistemas EBMT, e permite explicar a "Tradução por Analogia" de Nagao. Bhattacharyya (2015) propõe a adaptação do Triângulo e justifica-a neste sentido:

> "Translation by deep linguistic analysis is nothing but doing translation at the tip of the Vauquois triangle. This entails processing the input source sentence at many natural language processing (NLP) layers, followed by elaborate natural language generation. Analogy, which is founded on computation of similarity, demands capturing common parts of sentences, called fragmental phrases, a process akin to phrase table construction in SMT. Translating fragmental phrases and putting the translations together is like SMT's decoding process" (Bhattcharyya, 2015:193).

A fase de análise encontra-se no lado esquerdo do Triângulo, através do *matching* (correspondência) dos fragmentos de frases do *input*. A seguir, o *transfer* é operado do texto de partida até o texto de chegada, para encontrar os segmentos alinhados presentes nas memórias de tradução. A fase de geração é o processo de recombinação, no qual juntam-se os segmentos para produzir o texto na língua de chegada. Para concluir, a base do Triângulo representa a tradução direta, que acontece no caso fortuito de encontrar o exato correspondente na língua de chegada. É oportuno acrescentar que na "Tradução por Analogia", os algoritmos para o alinhamento diferem dos algoritmos dos sistemas SMT (algoritmos para o cálculo

da probabilidade), porque se baseiam no conceito de semelhança de textos e têm de respeitar dois elementos fundamentais: medir a semelhança para classificar os textos em função da semelhança e da dissemelhança e uso de redes léxico-conceptuais[6], que fornecem os recursos necessário para medir tal semelhança.

Graças às considerações feitas com o auxílio do Triângulo, pode-se resumir o processo de tradução dos sistemas EBMT, que consiste basicamente em três etapas: correspondência, alinhamento e recombinação. Na primeira fase, os exemplos são selecionados a diferentes níveis linguísticos e extraídos do banco de dados de exemplos. Cada exemplo é composto por um par de textos de dimensões arbitrárias em duas línguas diferentes dos quais um é a tradução do outro. Depois da seleção dos exemplos, o sistema encontra as várias correspondências e armazena os exemplos úteis para a tradução. Para encontrar as correspondências é crucial a noção de cálculo da distância, na qual é calculada a proximidade entre os exemplos armazenados, numa hierarquia de termos e conceitos que são fornecidos por um thesaurus o por redes léxico-conceptuais. Assim, o sistema calcula a distância entre o *input* e os vários exemplos graças à hierarquia do thesaurus (Arnold *et al*., 1994:188). Na fase do alinhamento, o sistema identifica os segmentos contidos nos exemplos que correspondem ao *input* e que vão ser utilizados na tradução, através da aplicação de algoritmos. Nas última fases, recombinação e reordenação, o sistema recombina e reordena os segmentos em unidades de tradução.

Os sistemas EBMT e SMT são orientados para os dados, mas, mesmo assim, apresentam algumas diferenças. Nos sistemas EBMT a ausência da probabilidade é evidente e, de consequência, os algoritmos não servem para calcular a probabilidade mas a semelhança de um fragmento do texto de partida com os exemplos armazenados nas memórias de tradução. Como nos SMT é presente o alinhamento, mas neste caso é utilizado para encontrar os "candidatos" de tradução melhores no banco de dados de exemplos, escolhidos através da correspondência sintática e semântica. Neste sentido, os EBMT são mais próximos aos RBMT. Por outro lado os SMT baseiam-se no puro cálculo probabilístico.

[6] A Wordnet é um exemplo de rede léxico-conceptual. Para o português, veja-se a rede léxico-conceptual desenvolvida no Centro de Linguística da Universidade de Lisboa pelo CLG – Grupo de Computação do Conhecimento Léxico-Gramatical, disponível em http://www.clul.ul.pt/wn/.

2.2.3 PARADIGMAS HÍBRIDOS

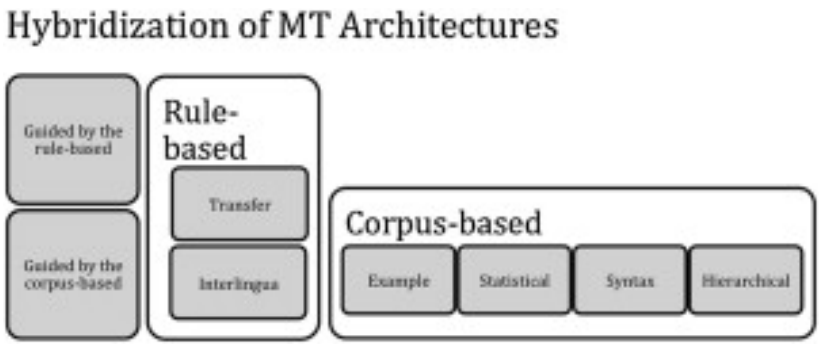
Nesta secção são apresentadas diferentes metodologias de hibridização dos paradigmas de tradução automática, com particular atenção à hibridização orientada por sistemas RBMT, como é o caso da versão 7.0 de 2009 do sistema SYSTRAN.

A expansão dos paradigmas de tradução automática orientados para o conhecimento linguístico e para os dados serviu também como área de investigação para encontrar os limites de cada um destes sistemas. É neste sentido que na última década surgiu a exigência de ultrapassar os limites dos sistemas orientados para o conhecimento linguístico e dos sistemas orientados para os dados, com a introdução dos paradigmas de tradução híbridos. No que diz respeito aos limites, os sistemas RBMT têm natureza dedutiva e são baseados em regras linguísticas. Estes sistemas armazenam os resultados de tradução, mas não reutilizam os segmentos precedentemente traduzidos, tornando mais difícil a adaptação a novas áreas. Por outro lado, os sistemas orientados para os dados têm natureza indutiva: as regras são derivadas diretamente de um conjunto de exemplos extraídos de textos já traduzidos e novas regras são introduzidas com novos exemplos. Um outro limite dos sistemas RBMT é a produção de resultados poucos consistentes depois da introdução no sistema de novas regras, além dos custos bastante elevados. Os sistemas orientados para os dados são bastante flexíveis no processamento das frases, mesmo que não estejam bem formadas, mas podem produzir resultados pouco satisfatórios na tradução de frases mais compridas, além de apresentar uma certa lentidão na fase de processamento. Estas ideias foram discutidas nas três edições do HyTra Workshop (*Workshop on Hybrid Apporaches to Translation*), onde linguistas, engenheiros e cientistas da computação se reuniram para construir um sistema de tradução híbrido de sucesso, tomando como ideia central a combinação dos pontos de força dos sistemas RBMT dos paradigmas orientados para os dados.

Face ao que é a matéria central deste trabalho, dá-se uma breve perspetiva sobre as principais estratégias de hibridização. Os sistemas híbridos podem ser classificados e caraterizados a partir da fonte de informação, ou seja, o

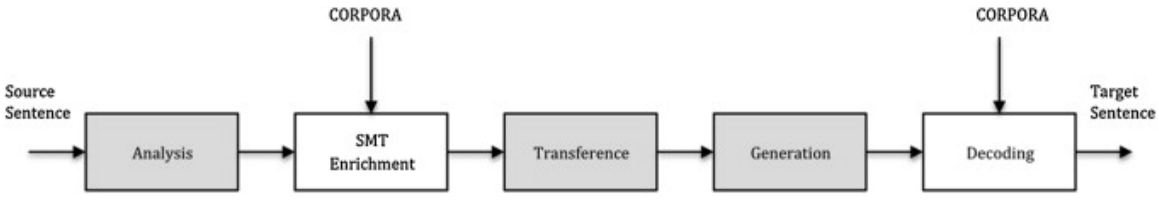
conhecimento linguístico e os dados. Neste sentido, é possível reconhecer três principais categorias de hibridização[7] (Quadro 6.):

1. Hibridização orientada por sistemas RBMT;
2. Hibridização orientada por sistemas baseados em *corpora*;
3. Hibridização entre SMT e EBMT, em que memórias de tradução são integradas com a introdução da tradução de sequências de palavras traduzidas por um tradutor humano, com componentes próprias dos sistemas EBMT e algoritmos para o alinhamento dos sistemas SMT.



Quadro 6. Tipologias de hibridização (Extraído de Costa-jussà e Fonollosa, 2015).

HIBRIDIZAÇÃO ORIENTADAS POR SISTEMAS RBMT. De acordo com Costa-jussà e Fonollosa (2015), é possível reconhecer três abordagens diferentes na hibridação orientada por sistemas RBMT.



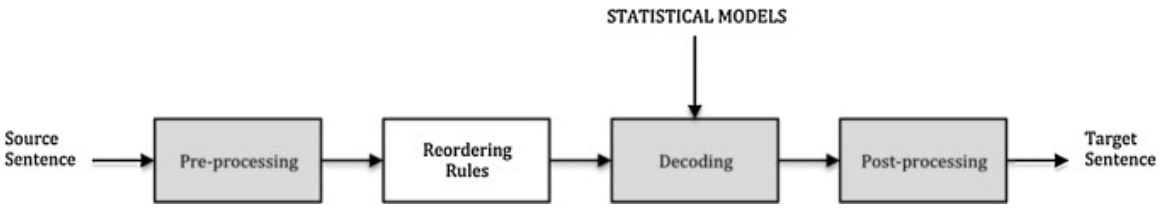
Quadro 7. Hibridização orientada por sistemas RBMT (*Ibdiem*).

[7] Para uma leitura aprofundada sobre a matéria veja-se, entre outros, España-Bonet e Costa-jussà (2016:1-26).

Uma primeira abordagem consiste na utilização de um *corpus* para a construção do sistema RBMT. O motivo principal do uso desta abordagem está na redução do tempo e dos custos de tradução. É uma abordagem simples que inclui também a melhoria dos dicionários com a introdução de frases e exemplos extraídos dos *corpora* e de entradas da BabelNet e do Wiktionary. As abordagens mais complexas utilizam regras de *transfer* e constroem modelos de seleção do léxico a partir dos *corpora*.

Além desta abordagem, há também uma outra em que ferramentas baseadas em *corpora* são utilizadas para o melhoramento do *output* produzido por um sistema RBMT, através da introdução de modelos de linguagem. É neste sentido que se desenvolveu recentemente a investigação conduzida pela empresa SYSTRAN na construção de um sistema estatístico de inferências para substituir o modelo de *transfer* dos sistemas RBMT.

HIBRIDIZAÇÃO ORIENTADA POR SISTEMAS BASEADOS EM *CORPORA*. Nesta abordagem os sistemas híbridos podem ser incorporados com regras ou podem também ser combinados sistemas SMT e EBMT.



Quadro 8. Hibridização orientada por sistemas SMT (*Ibidem*).

Há basicamente duas formas para a integração de regras nos sistemas SMT: na primeira, a inclusão de regras é realizada nas fases de pré e pós edição dos textos; na segunda, são incorporados dicionários no modelo nuclear e é introduzida a informação das regras dos sistemas RBMT para o melhoramento do alinhamento estatístico.

Voltando aos sistemas híbridos, cabe mencionar que têm diversas aplicações, de que são exemplo a tradução da fala ou a integração nos sistemas de tradução

assistida por computador. A lista, claramente, não é exaustiva, dado que é uma área de investigação bastante recente e ainda em fase de desenvolvimento.

Acrescente-se também que nos sistemas híbridos permanece o problema da tradução de terminologia, herdade dos sistemas RBMT, em que é preciso procurar o equivalente exato na língua de chegada. Este problema pode ser ultrapassado através da integração dos sistemas SMT, nos quais a tradução é vista como um problema de aprendizagem da máquina. É o próprio utilizador que personaliza o sistema e treina o sistema. A empresa SYSTRAN foi a primeira, em 2009, a utilizar sistemas de tradução automática deste tipo.

Além do recurso a sistemas híbridos para compensar as carências dos paradigmas orientados para o conhecimento e dos paradigmas orientados para os dados, há outras duas abordagens a mencionar: os sistemas de tradução automática interativa e os *multiengine systems*[8].

2.2.4 SISTEMAS DE TRADUÇÃO AUTOMÁTICA ONLINE

Hoje em dia, a par dos sistemas de tradução automática disponíveis no mercado, existem também sistemas de tradução automática online que, na maioria dos casos, são versões gratuitas dos softwares de tradução automática, como é o caso do SYSTRANet, versão online do sistema de tradução automática SYSTRAN. Os sistemas de tradução automática online tornaram-se populares por causa da crescente procura de traduções em tempo real nas redes sociais, blogues e sites para uma rápida troca de informação. Estes sistemas funcionam principalmente em dois sentidos: podem ser utilizados por parte de um utilizador específico ou podem ser integrados nos sites como motores de tradução, como nos casos do Facebook, Twitter e Instagram ou outras redes sociais que oferecem traduções em tempo real de conteúdos e comentários.

O sistema de tradução automática online mais conhecido a nível mundial é o Google Translate da Google Inc., lançado em 2006 e que utilizou o software SYSTRAN, baseados em regras, até 2007. A partir de outubro de 2007, o Google Translate deixou de usar as tecnologias SYTRAN baseadas em regras e lançou o novo Google Translate, baseado em estatística. O Google Translate traduz textos, frases presentes em imagens, sites, vídeos em tempo real e discursos orais a partir

[8] Para uma leitura mais aprofundada veja-se Quah (2006) e Hutchins (2010:29-70), entre outros.

de 103[9] pares linguísticos e, além do sistema de tradução automática disponível na Internet, a Google Inc. disponibiliza também aplicações para os sistemas operativos Android e iOS, além de disponibilizar igualmente um API (*Application Browser Interface*) para a construção de softwares. A par do Google Translate, estão disponíveis na Internet outros sistemas de tradução automática como BabelFish Yahoo![10], Promt-online[11], WorldLingo[12] e SYSTRANet[13], versão gratuita do sistema de tradução automática SYTRAN, apresentado na secção a seguir.

2.3 O SISTEMA SYSTRAN

A empresa SYSTRAN foi fundada em 1968 por Peter Toma e é uma das primeiras na área da tradução automática com primeira sede em La Jolla, California. A SYSTRAN nasceu de uma das primeiras experiências realizadas na área da tradução automática, em 1954 na Georgetown University com o apoio da IBM, e é uma das poucas empresas de tradução automática que sobrevive ao relatório ALPAC de 1966. A empresa tinha como objetivo a tradução de documentos do russo para o inglês durante a Guerra Fria, utilizados pela Força Aérea americana sob o patrocínio da Foreign Technology Division. Durante este primeiros anos, o sistema de tradução automática SYSTRAN era baseado em dicionários (*Dictionary-based Machine Translation*) e apesar de a qualidade das traduções ser pouco elevada, era ainda assim suficiente para a compreensão dos textos.

O ponto de viragem na investigação da empresa SYSTRAN foi em 1975, ano em que foi proposto à Comissão da Comunidade Europeia (CEC) um protótipo de sistema de tradução automática para a tradução da combinação linguística inglês-francês e, a partir de 1976, para a tradução do francês para inglês e do inglês para italiano. É nestes anos que a Comissão começa a utilizar o SYSTRAN como sistema de tradução automática promovendo deste modo a tradução entre outras combinações linguísticas, disponíveis a partir de 1981.

[9] Dados atualizados de Julho de 2016 (https://en.wikipedia.org/wiki/Google_Translate).

[10] Disponível em https://www.babelfish.com.

[11] Disponível em www.online-translator.com.

[12] Disponível em www.worldlingo.com.

[13] Disponível em www.systranet.com/translate.

Em 1986, a empresa foi vendida à família Gachot e a sede foi transferida para Paris, ainda que mantendo em atividade a sede original em La Jolla. É neste ano que tem lugar a primeira World Systran Conference, organizada pela Comissão da Comunidade Europeia. Foi a única conferencia na área da tradução automática dedicada unicamente a um sistema de tradução.

No que diz respeito às abordagens de tradução automática, podem ser identificadas três ao longo da história do SYSTRAN: tradução direta, a abordagem baseada em regras que utiliza o *transfer* e a mais recente hibridação das componentes baseadas em regras com sistemas de tradução automática estatísticos. Na primeira "geração", o sistema de tradução automática SYSTRAN baseava-se em dicionários bilingues da língua de partida e da língua de chegada, gerando o texto de acordo com o processo de geração de textos dos sistemas baseados em dicionários (veja-se a *secção 2.2.1.*). No sistema SYTRAN baseado em regras (RBMT) de segunda "geração", os dicionários bilingues continuavam a ser a componente principal: o *Main Stem Dictionary*, com as entradas lexicais bilingues, as descrições morfológicas, sintáticas e semânticas, os marcadores semânticos e a tradução da forma equivalente de cada entrada nas línguas de chegada; e o *Multi-word Contextual Dictionary*, que fornecia os dados necessários para a análise de uma entrada lexical conforme o contexto. Apesar de os dicionários serem a componente principal, a geração do texto na língua de chegada era efetuada através do processo ATG (análise-*transfer*-geração; veja-se a *secção 2.2.1.)*. Estes sistemas possuíam um elevado grau de modularidade, uma vez que a fase de inserção de novos módulos linguísticos não afetava os módulos já existentes para outros pares linguísticos e os programas de análise e síntese eram independentes de uma combinação de línguas:

> "The system has been designed in order to be more modular. The modularity means than we can extract each component from the system and use it for other purposes" (Senellart *et al.,* 2001:3).

Na terceira "geração" do sistema de tradução automática SYSTRAN são combinadas as caraterísticas dos sistemas baseados em regras com os métodos estatísticos da abordagem baseada em *corpora*. A ideia que está na base desta

hibridização foi proposta por Senellart[14] em 2008: *Can we relearn an RBMT system*?, ou seja "Podemos reaprender um sistema RBMT?", em que Senellart analisa os pontos a favor das novas tecnologias baseadas em *corpora*, mais competitivas do que o sistema SYSTRAN baseado em regras, dado que os *corpora* têm os recursos necessários para a tradução de um texto que pertença a uma área especifica. Neste sentido, a equipa da SYSTRAN começa a trabalhar na hibridação do "antigo" sistema SYSTRAN, puramente baseado em regras, com as técnicas estatísticas,

> "We call this system "SYSTRAN Relearnt" because, as far the translation model is concerned, this system is a statistical model of the rule-based engine" (Dugast *et al*., 2008:175).

Assim, em 2009, com a versão 7.0 do software SYSTRAN, nasceu o novo SYSTRAN híbrido, o primeiro deste tipo, que na versão 8.0 traduz a partir de 130 combinações linguísticas. A particularidade deste sistema de tradução reside no facto de traduzir textos e de aprender diretamente das traduções efetuadas, de forma automática. Isto quer dizer que os utilizadores podem "treinar" e "ensinar" o sistema de tradução a traduzir textos numa determinada área com o objetivo de melhorar a qualidade das traduções e diminuir os custos.

SYSTRANET. SYSTRANet é a versão online gratuita do sistema de tradução automática SYSTRAN e pode ser utilizado com ou sem a criação de uma conta. Na versão que pode ser utilizada sem conta, o utilizador dispõe de um serviço de tradução online de textos, documentos pessoais e também de páginas web. As traduções podem ser visualizadas diretamente na interface do site ou podem ser recebidas por e-mail. Na versão sem conta, são 36 as combinações linguísticas disponíveis. Por outro lado, SYSTRANet permite a criação de uma conta gratuita para aceder a serviços de tradução extra, como o uso de dicionários especializados, a possibilidade de criar um dicionário pessoal, a tradução de documentos que mantêm a formatação do ficheiro original e a possibilidade de traduzir conteúdos

[14] Jean Senellart, CEO da SYSTRAN SA.

RSS[15] a partir de 52 combinações linguísticas. Além destes serviços extra, o sistema dispõe de todas as outras funcionalidades disponíveis sem a criação de conta.

O site do SYSTRANet disponibiliza também um manual para os utilizadores que contém sugestões sobre a utilização dos sistemas e sobre as regras a serem seguidas durante a fase de tradução, consideradas nas secções a seguir na criação de regras gerais para o controlo da língua portuguesa na tradução automática para italiano.

[15] *Really Simple Syndication*, formato para a distribuição de conteúdos nas páginas web.

3. AS LINGUAGENS CONTROLADAS

Neste capítulo é tratada a noção de linguagem controlada, avançando uma definição da própria e explicando a relevância que atualmente tem na investigação científica.

Na *secção 3.1,* à luz dos critérios de legibilidade e de traduzibilidade, são abordadas duas tipologias de linguagem controlada: as orientadas para os humanos e as orientadas para as máquinas.

Na *secção 3.2* é explicado o processo de criação de regras de linguagem controlada, mais especificamente de regras proscritivas e prescritivas.

Na *secção 3.3* são tratadas, de maneira mais aprofundada, as linguagens controladas aplicadas à tradução automática.

A noção de linguagem controlada (CNL, *Controlled Natural Language*) nasce da ideia de simplificar a estrutura de uma língua natural para que um falante não nativo a possa utilizar de maneira mais fácil e eficaz. É neste sentido que nos anos 30 foi criado o *Basic English*, o qual estabelecia uma variedade "mínima" do inglês para a utilização por parte de falantes não nativos, tendo em vista uma comunicação mais fácil e, consequentemente, mais rápida. O *Basic English* baseava-se na utilização de poucas palavras (aproximativamente 75.000) e no uso de estruturas frásicas simples e não ambíguas, sendo por isso considerado o primeiro exemplo de linguagem controlada. Isto porque, geralmente, a linguagem controlada é definida como um subgrupo de uma língua natural, que impõe simplificações, o uso restrito do vocabulário e outros tipos de restrições. Em geral, uma linguagem controlada é

> "[...] a subset of natural language with artificially restricted vocabulary, grammar and style" (Kaji, 1999:37).

Por outras palavras, uma linguagem controlada é composta por uma série de regras que operam a nível lexical, sintático e estrutural criadas pelo ser humano para ir ao encontro de um objetivo específico. A questão principal, neste sentido, é tornar a comunicação o quanto menos ambígua possível, o que constitui o objetivo principal da investigação nesta área. É neste aspeto que se encontra a diferença crucial entre uma língua natural e a linguagem controlada: na linguagem controlada as restrições impostas a nível lexical, sintático e semântico visam eliminar totalmente, se possível, qualquer tipo de ambiguidade. Por isso, pode dizer-se que o objetivo

principal das linguagens controladas consiste na redução da ambiguidade e da complexidade dos textos, limitando o modo como a informação é expressa, procurando melhorar a comunicação. Além da eliminação ou redução das ambiguidades, um outro elemento que define as linguagens controladas é objetivo ao qual se destinam. Neste sentido, podem ser aplicadas na criação de textos para facilitar a comunicação ou para a redação de textos a serem traduzidos por um sistema de tradução automática, questões que são aprofundadas nas secções a seguir. É importante mencionar também as diversas aplicações das linguagens controladas, como a representação do conhecimento, a produção de textos técnicos, a simplificação de uma língua natural e, para concluir, o controlo do desempenho de um sistema de tradução automática, questões que são igualmente abordadas nas secções que se seguem.

O uso das linguagens controladas implica várias vantagens e desvantagens. A vantagem geral, como referido anteriormente, está no facto de a manipulação do texto o tornar mais "compreensível" para os homens e para as máquinas, através da redução das ambiguidades, das formas homonímias, das sinonímias e da complexidade lexical. Isto comporta uma maior consistência textual e uma maior uniformidade no uso de estruturas frásicas e no uso da terminologia. É neste sentido que a tradução automática beneficia da aplicação das linguagens controladas: maior rapidez nas traduções e redução dos custos. Uma outra vantagem que as linguagens controladas trazem, consiste no facto de poderem ser utilizadas como base para o ensino ou a aprendizagem de uma língua estrangeira, como no caso do Português Controlado[16] criado pelo CLG da Universidade de Lisboa. Por outro lado, a desvantagem principal consiste na memorização, por parte do autor do texto, de regras que podem ser muito complexas, com uma consequente demora na redação.

[16] *Cf.* Marrafa *et al*., (2011).

3.1 LEGIBILIDADE E TRADUZIBILIDADE[17]

É oportuno fazer algumas considerações preliminares sobre duas abordagens diferentes à linguagem controlada. É através da escolha de uma destas abordagens que se definem as caraterísticas de cada tipologia de linguagem controlada, em função do objetivo ao qual se destinam. De acordo com Clark *et al.* (2009), é possível reconhecer duas escolas diferentes no que diz respeito à abordagem utilizada na construção de uma linguagem controlada: naturalista e formalista. Na primeira, a interpretação da linguagem controlada é tratada como uma forma simplificada de uma dada língua natural, onde permanecem as ambiguidades, se bem que em número menor, com o objetivo de tornar a interpretação desta língua mais fácil. Na abordagem formalista, por outro lado, as interpretações da linguagem controlada são especificações da "língua base" natural que torna a linguagem controlada num tipo de linguagem de programação bem definido e mais fácil de utilizar em comparação com a "língua base". Citando Marrafa *et al.*, é possível resumir as duas abordagens do seguinte modo:

> "[...] "naturalist" approaches, which view controlled languages as sets of restrictions on the existing structures and lexicon of a given natural language, stating which structures and lexical items are not to be used; and "formalist" approaches, which view controlled languages as sets of vocabulary and rules to form utterances in a given natural language, determining the lexicon allowed as well as the syntactic and interpretation rules allowed" (Marrafa *et al.,* 2012:153).

À luz destas considerações, é possível classificar as linguagens controladas em função do objetivo ao qual se destinam e portanto, neste sentido, reconhecem-se duas orientações principais no uso das linguagens controladas, ou seja, as orientadas para os humanos (HOCL, *Human-oriented Controlled Language*) e as orientadas para as máquinas (MOCL, *Machine-oriented Controlled Language*), que respondem respetivamente aos critérios de legibilidade e de traduzibilidade.

No que diz respeito às HOCL, o objetivo é melhorar a legibilidade, a compreensibilidade e a consistência dos textos para a comunicação entre humanos.

[17] *Cf.* Reuther (2003:124-132).

As HOCL tiveram uma aplicação particularmente importante no âmbito do comércio internacional (*Basic English*[18]) e na indústria, especialmente na criação de manuais técnicos (ASD, *Simplified Technical English*[19]) e na escrita de advertências sobre o uso de certas máquinas (Airbus Warning *Language*[20]). De acordo com Marrafa *et al.,* (2012), esta tipologia de linguagem segue a abordagem naturalista, de mais fácil compreensão e utilização por um humano e na qual permanecem algumas das ambiguidades próprias das línguas naturais. Por outro lado, nas chamadas MOCL, um dos objetivos é tornar o texto "compreensível" e por isso processável por um sistema de tradução automática seguindo os "critérios de traduzibilidade" (Reuther, 2003). Neste caso, as regras de linguagem controlada podem comportar uma degradação do *input* para que o *output* seja gramatical e, consequentemente, a leitura por parte de um humano pode tornar-se mais complicada. As informações têm de ser específicas e as instruções são dadas para que uma máquina, através de processos computacionais, consiga "compreender" a informação. É por isso que, por exemplo, cada entrada lexical tem de incluir o necessário para a gestão da terminologia, como detalhes sobre as categorias sintáticas ou as datas de criação e de modificação. As MOCL seguem a abordagem formalista, porque utilizam um único sentido e uma única interpretação aceitável, com uma consequente melhor "compreensão" e utilização por uma máquina. Esta tipologia de linguagem controlada é de mais difícil utilização por parte do humano e, muitas vezes, depende de ferramentas sofisticadas para o seu uso. Também neste caso as aplicações são múltiplas e este particular tipo de linguagem controlada pode ser aplicada à redação de documentos traduzíveis através de um sistema de tradução automática (KANT[21]), na representação e aquisição do conhecimento (ACE[22],

[18] *British American Scientific International Commercial*. Linguagem controlada criada por Charles Key Ogden em 1930 que visa à simplificação da língua inglesa para a comunicação.

[19] ASD STE-100, *Simplified Technical English*. Linguagem controlada para a produção de documentação na industria aerospacial, variante simplificada do inglês.

[20] Linguagem controlada desenvolvida em 1998 pela Airbus para a criação de documentação técnica.

[21] CTE, *Caterpillar Technical English*. Linguagem controlada criada para o sistema de tradução automática KANT e desenvolvida pela Mellon Carneige University junto com a empresa Caterpillar Inc.

[22] ACE, *Attempto Controlled English*. Linguagem controlada para a representação do conhecimento, desenvolvida pela Universidade de Zurique.

PENG[23], CPL[24]) e na construção de redes semânticas (ACE *View*[25], *Rabbit Lite Natural Language*).

As HOCL e as MOCL têm, como é óbvio, algumas caraterísticas em comum, como por exemplo a limitação do comprimento das frases e imposição do uso de determinadas estruturas frásicas. Por outro lado, é possível encontrar um ponto de divergência no que diz respeito à forma como as regras são escritas. No caso das linguagens orientadas para os humanos, as regras *podem* ser computacionalmente intratáveis e intencionalmente vagas, enquanto no caso das linguagens orientadas para as máquinas, *devem* ser precisas e computacionalmente tratáveis.

3.2 CONCEPÇÃO DE UMA LINGUAGEM CONTROLADA

A partir das definições de linguagem controlada propostas por Kittredge (2003), a qual afirma que a linguagem controlada é uma versão "restrita" de uma língua natural, e por Arnold (1995), em que a linguagem controlada é vista como uma forma de uso da língua em que se opera o controlo sobre a gramática e o léxico, é possível extrapolar dois elementos cruciais para a criação de uma linguagem controlada, independentemente do objetivo ao qual se destina: uma determinada língua natural e as restrições impostas sobre a gramática e o léxico da mesma. De acordo com Kuhn (2013), a linguagem controlada baseia-se numa língua natural, chamada "língua base" (*base language*), da qual difere em virtude das restrições lexicais, sintáticas e semânticas. Além disso, preserva todas as caraterísticas da própria "língua base", para que os utilizadores possam perceber, de forma intuitiva, o texto. Dado que se trata de uma língua "construída" e, por isso, explicitamente definida, não é produto de um processo natural implícito do ser humano. Quanto às restrições impostas sobre a gramática e o léxico, de acordo com Mitamura e Nyberg (1995), são aplicadas a três níveis diferentes: lexical, em que são criadas regras para a eliminação da ambiguidade e para a seleção de palavras e termos próprios de um

[23] PENG, *Processable English*. Linguagem controlada para a representação do conhecimento.

[24] CPL, *Computer Processable English*. Linguagem controlada para a representação do conhecimento, desenvolvida pela Boeing Research Technology.

[25] ACE *View, Attempto Controlled English View*.

determinado contexto; sintático, em que as regras operam nos constituintes frásicos; e, por fim, a nível estrutural, com regras textuais e pragmáticas.

No que diz respeito à tipologia de regras e de acordo com Somers *et al*., (2003), distinguem-se duas abordagens, nomeadamente a abordagem proscritiva e a abordagem prescritiva. Na abordagem proscritiva, basicamente, as regras descrevem as estruturas não permitidas que, numa fase sucessiva de análise, são comparadas com o *input*. Neste processo são detetadas as estruturas que não são permitidas sem a necessidade de especificar, de forma exaustiva, as estruturas permitidas. Esta abordagem ignora alguns problemas que podem surgir, com a consequente produção de um *output* não adequado. Na abordagem prescritiva as regras descrevem as estruturas permitidas e, normalmente, são criadas de raiz graças a um trabalho mais intensivo, dado que requerem uma definição por cada estrutura linguística permitida. No caso em que esta gramática prescritiva seja implementada num sistema computacional, pode acontecer que cada frase seja analisada para verificar se é permitida e se respeita as regras da própria gramática. Nesta abordagem, a análise das estruturas é mais aprofundada e é mais difícil que o *output* seja inapropriado. Por outro lado, é provável que haja algumas estruturas frásicas que são ignoradas na definição original da linguagem mas que são consideras necessárias (Somers *et al*., 2003:252-253). Para concluir, as regras distinguem-se normalmente entre regras gerais, que permitem a eliminação do maior número de ambiguidades e que podem ser utilizadas para o controlo de línguas diferentes, e regras específicas, normalmente criadas para o controlo de uma determinada língua.

3.3 LINGUAGEM CONTROLADA PARA A TRADUÇÃO AUTOMÁTICA

> *MT is potentially one of the most interesting computational application of CL. If a CL and MT system are attuned to each other, MT of texts written in that CL can be much more efficient and effective, requiring far less – or ideally even no – human intervention* (Somers *et al*., 2003:254).

Um dos objetivos da tradução automática é garantir uma tradução de alta qualidade e, para que isso aconteça, em muitos casos é necessária a intervenção humana na fase de pré-edição do texto, o que requer a simplificação e a redução das estruturas ambíguas no texto original ou a redação do próprio texto em linguagem controlada.

Uma das necessidades de ter traduções de alta qualidade nasce da internacionalização de muitas empresas em áreas do mundo em que é falada mais do que uma língua e, o controlo do texto de partida insere-se nesta perspetiva. Isto porque muitas empresas multinacionais têm, em primeiro lugar, necessidade de reduzir os custos e os tempos de tradução, mas têm também interesse em traduções multilingues, dada a importância a nível linguístico e económico de determinadas áreas do mundo. É por esta razão que muitas empresas desenvolveram as próprias linguagens controladas, como é o caso da Caterpillar Technical English, criada pela Caterpillar Inc. em conjunto com a Carneige Mellon University, para a tradução automática através do sistema KANT. Este é um exemplo útil para demonstrar que se trata de uma área de investigação fértil, capaz também de criar parcerias entre empresas e universidades. Neste sentido, as empresas internacionais produzem e traduzem a sua própria documentação: os manuais técnicos são escritos na linguagem controlada desenvolvida pela própria empresa e são traduzidos através de um sistema de tradução automática específico. Como referido anteriormente, isto comporta a redução dos custos e dos tempos de tradução e visa garantir uma consistência terminológica e uma consequente melhor qualidade da tradução.

Os critérios para a aplicação de uma linguagem controlada à tradução automática são vários, dado que se trata de uma tradução de alta qualidade para a disseminação de informação, ou seja, trata-se de uma tradução publicável. Primeiro, os autores devem conhecer as regras da linguagem controlada e têm de ser "bem treinados" se não disponibilizam de *softwares* e ferramentas para a redação de textos. Além disso, o domínio tem de ser bem definido, para que a terminologia não seja ambígua e seja o mais consistente possível. Se o autor utiliza ferramentas para a redação de textos, então tem de utilizar também *checkers*, ou seja "corretores" que lhe permitam corrigir o texto conforme às regras, caso contrário é necessário ter muito cuidado na releitura do texto antes de proceder com a tradução.

Na aplicação da linguagem controlada, há duas abordagens à tradução automática diferentes: a tradução automática para linguagens controladas "vagamente" definidas e a tradução automática para linguagens controladas "estritamente" definidas. No primeiro caso, as especificações da linguagem controlada não são muito precisas, como no caso da PACE, linguagem controlada criada pela Perkins Engine LTD. A caraterística fundamental desta linguagem consiste na simplificação de documentos que podem ser utilizados por falantes não

nativos de inglês e num léxico de 2.500 palavras e 10 regras para a redação. Em contraste, nas linguagens controladas "estritamente" definidas encontram-se especificações formais da sintaxe. Isto é um argumento muito interessante para a tradução automática dado que, graças à escolha de restrições impostas de forma apropriada, é possível garantir uma tradução de alta qualidade, idealmente sem intervenção humana na pós-edição:

> "The CL itself is to be designed in such a way that user involvement is limited to the phase of document creation. Subsequent translation should fully automatically produce grammatically correct target-language expressions that are acceptable as translations and that require no (or, at worst, minimal) post-editing" (Somers *et al.,* 2003:256).

Existem também linguagens controladas que funcionam só com um determinado tipo de sistema de tradução automática, como é o caso do *Multinational Customized English*, desenvolvido para a Xerox Corporation e aplicável ao sistema de tradução automática SYSTRAN. As funções desta linguagem controlada são múltiplas, nomeadamente a eliminação de ambiguidades no texto *input*, uma melhor qualidade do *output*, uma rápida produção de documentos técnicos em várias línguas e uma leitura facilitada do texto *input*. Pode-se dizer que a aplicação do *Multinational Customized English* traz diversas vantagens, como a produção de boas traduções a baixo custos, entregas dentro dos prazos e melhoramento da comunicação, respondendo, portanto, aos dois critérios precedentemente mencionados, a legibilidade e a traduzibilidade.

No caso das linguagens controladas orientadas para as máquinas, as regras diferem ligeiramente das regras da linguagem controlada orientada para os humanos. Primeiro, é preciso considerar os fenómenos da língua de partida e de chegada. Neste sentido, é possível que o controlo da linguagem empobreça o *input* e que a frase controlada seja agramatical ou pouco fluente para um falante nativo mas que, ainda assim, o sistema de tradução automática consiga produzir um *output* aceitável. Isto é possível porque na fase de criação destas regras são considerados também os fenómenos de processamento da linguagem natural envolvidos no sistema de tradução automática. No que diz respeito ao léxico, ao contrário do que acontece nas linguagens controladas orientadas para os humanos, não há imposições quanto ao número de palavras a utilizar, pois os computadores

conseguem memorizar um número maior de palavras. É importante também dizer que, não obstante o maior número de palavras, os léxicos controlados são constituídos por listas de palavras aprovadas e não aprovadas, sobretudo em certas áreas de especialização.

A aplicação da linguagem controlada à tradução automática comporta também o desenvolvimento de outras ferramentas para a redação de textos destinados à tradução, dada a dificuldade de memorização e de utilização destas regras por parte dos autores. Por isso, há sistemas de linguagem controlada, nomeadamente *checkers* para a gramática e o vocabulário, sistemas para a autoria de textos escritos em linguagem controlada, sistemas para a autoria interativos e memórias de linguagem controlada.

Uma linguagem controlada, para ser eficaz e para ser utilizada ao longo do tempo precisa de manutenção, sobretudo porque a terminologia das áreas técnicas muda continuamente e precisa de ser atualizada. Como já referido anteriormente, a terminologia deve ser consistente. Neste sentido, graças à colaboração de vários autores que utilizam ao mesmo tempo a mesma linguagem controlada, é necessário um processo bem definido de manutenção. Primeiro, os autores têm de reportar os problemas encontrados no uso da terminologia ou da gramática controlada, que têm de ser analisados e resolvidos por especialistas. Este trabalho consiste na avaliação e na revisão periódicas dos problemas do texto de partida e de chegada. Depois da anotação e resolução dos problemas terminológicos, o *checker* tem de ser implementado com a inclusão da nova terminologia, conjuntamente com a implementação da terminologia na língua de chegada no sistema de tradução automática. O mesmo processo aplica-se também às regras sintáticas e semânticas. Outros problemas ligam-se à manutenção da terminologia, como o estabelecimento de um método para a criação de linguagens controladas utilizáveis em determinadas áreas e uma maneira para aumentar as funções e a precisão de ferramentas para a autoria de textos em linguagem controlada. Uma chave para resolver estes problemas encontra-se no uso das técnicas de processamento da linguagem natural próprias dos sistemas de tradução automática baseados em *corpora,* através da análise de textos que pertencem a uma área específica, para resolver problemas relacionados com a ambiguidade e a consistência terminológica:

> "Beyond being used for studying the vocabulary, a corpus will also play essential roles in designing an acceptable and effective controlled language. For example, corpus-based word-sense disambiguation will help us specify approved and unapproved meanings of polysemous words. Moreover, the capability of controlled- language authoring tools for detecting ambiguities can be greatly improved by using knowledge extracted from the corpora of domains" (Kaji, 1999:39).

Atualmente, os investigadores estão a trabalhar em novas ferramentas e em novos sistemas de linguagem controlada aplicáveis à tradução automática. Um exemplo disso pode ser a criação de sistemas que rescrevem "automaticamente" o texto em linguagem controlada. Neste caso, é o sistema que aplica as regras, escolhe a terminologia adequada e muda as estruturas das frases quando o autor não escreve respeitando as regras. A desambiguação é feita sem intervenção humana, requerida só para a releitura do texto de modo a verificar se ocorreram erros de outro tipo. Estes sistemas poderiam ajudar a aumentar a produtividade e a reduzir os problemas. Os investigadores estão a trabalhar também em sistemas que traduzam automaticamente o texto na língua base para um texto em linguagem controlada, dado que pode acontecer que as regras sejam muito complicadas.

No capítulo que se segue, é apresentado um fragmento de português controlado para a tradução automática para italiano, sendo o sistema de tradução automática utilizado para o efeito o SYSTRANet, disponível online. Na criação deste fragmento, primeiro, são analisadas as especificidades do português e do italiano com o objetivo de identificar as estruturas mais problemáticas no que diz respeito ao modo, à modalidade, ao tempo e ao aspeto. De seguida, são dadas justificações para o controlo e, por fim, são analisados os resultados da tradução automática obtidos através do controlo do *input*. O fragmento de linguagem controlada criado segue duas diretivas principais: do ponto de vista da abordagem, como não foi possível desenvolver uma linguagem controlada de ampla cobertura, escolheu-se criar regras para o controlo no âmbito do modo, da modalidade e do aspeto recorrendo a regras "estritamente" definidas. Quanto às regras, são de carácter proscritivo, mas especificam também as estruturas que têm de ser utilizadas, para fornecer uma alternativa à proscrição. Isto quer dizer que se baseiam numa determinada língua natural, neste caso o português, e definem as estruturas que não são permitidas na redação do texto e deixam indicações claras no que diz respeito às estruturas a utilizar.

4. PORTUGUÊS CONTROLADO

Nas secções a seguir, são apresentados e analisados fenómenos linguísticos que colocam problemas de tradução automática, em particular os que decorrem das especificidades do italiano e do português no uso de modos verbais em frases subordinadas, bem como na expressão da modalidade, do tempo e do aspeto. Antes de aprofundar o estudo destas questões, são analisados exemplos que serviram como base para o estabelecimento de regras gerais de linguagem controlada aplicáveis na combinação linguística português-italiano. Como já referido anteriormente, o sistema de tradução automática utilizado para o efeito é o SYSTRANet. Na criação deste conjunto de regras, foram seguidas as sugestões para a redação de textos que o próprio sistema fornece no *Help Center*[26], as regras gerais de linguagem controlada criadas pelo CLG - Grupo de Computação do Conhecimento Léxico-Gramatical - do Centro de Linguística da Universidade de Lisboa para o par linguístico português-inglês[27], que também serviu como base para a criação das regras específicas igualmente propostas neste trabalho.

4.1 REGRAS GERAIS

Graças à ajuda do site e às regras de linguagem controlada criadas pelo CLG do Centro de Linguística da Universidade de Lisboa, foi possível estudar e traduzir as frases do *corpus* para criar regras gerais aplicáveis ao par linguístico português-italiano. O site do SYSTRANet aconselha a tradução de textos curtos que tenham frases curtas e simples. Para testar o sistema, foi introduzido um texto bastante longo no tradutor e foi verificada a tradução para italiano. Como a tradução não resultou correta, foi preciso segmentar e simplificar o texto:

[26] O *Help Center* do site do SYSTRANet, na secção *How to improve translation quality?* disponibiliza algumas sugestões para a redação de textos que podem ser traduzidos através do próprio sistema de tradução. Disponível em: http://www.systranet.com/systranet-help/help-improve-translation-quality.

[27] *Cf.* Marrafa *et al.*, (2012:152-166).

(1a) Estão todos entre as dez espécies desta lista elaborada por um grupo de especialistas internacionais e que ontem foi divulgada pelo International Institute of Species Explorations do ESF, College of Environmental Science and Forestry de Nova Iorque, para celebrar o dia de nascimento, a 23 de maio, de Carolus Linnaeus, que no século XVIII criou a moderna taxonomia -[28] a classificação das espécies.

(1b) Sono tutti tra le dieci specie di questa lista elaborata da un gruppo di specialisti internazionali e che ieri è stato rivelato dall'International Institute of Species Explorations di ESF, collegio of Environmental scienza and Forestry, di New York, per celebrare il giorno di nascita, il 23 maggio, di Carolus Linnaeus, che tra il secolo XVIII ha creato la tassonomia moderna - la classificazione delle specie.

LC: (1c) Estão todos entre as dez espécies desta lista elaborada por um grupo de especialistas internacionais. A lista foi divulgada ontem pelo International Institute of Species Explorations do ESF[29] de Nova Iorque, para celebrar o dia de nascimento de Carolus Linnaeus, no dia 23 de maio[30]. Linnaeus, no curso do século XVIII, criou a taxonomia moderna, ou seja, a classificação das espécies.

(1d) Sono tutti tra le dieci specie di questa lista elaborata da un gruppo di specialisti internazionali. La lista è stata rivelata ieri dall'International Institute of Species Explorations di ESF di New York, per celebrare il giorno di nascita di Carolus Linnaeus, il 23 maggio. Linnaeus nel corso del secolo XVIII, ha creato la tassonomia moderna, cioè, la classificazione delle specie.

[28] Simplificação da pontuação.

[29] Eliminação da designação completa do instituto por causa da ambiguidade lexical *college/collegio*.

[30] Em italiano utiliza-se a expressão *il giorno,* razão por qeu no controlo o português *a dia* foi substituído pela tradução literal do italiano *o dia*.

Como é possível observar no exemplo, as frases foram simplificadas. Foi modificada também a pontuação, para que o sistema consiga processar melhor a informação sem cometer erros de tradução. Em consequência, a tradução resultante do texto controlado não revela problemas de gramaticalidade.

Nas frases a seguir são analisados exemplos de má ortografia, muito frequente sobretudo em textos que podem ser encontrados online nos blogues e nas redes sociais, em que muitas vezes aparece a opção de tradução automática que pode ser efetuada pelos utilizadores. Vejam-se os exemplos:

(2a) O Miguel é um rapaz muito perguiçoso.
(2b) *Miguel è un giovane molti perguiçoso.
LC: (2c) O Miguel é um rapaz muito preguiçoso.
(2d) Miguel è un giovane molto pigro.

(3a) O Bruno repara sempre nos promenores.
(3b) *Bruno ripara sempre in promenores.
LC: (3c) O Bruno nota[31] sempre os pormenores.
(3d) Bruno osserva sempre i dettagli.

Nos exemplos (2a) e (3a) observa-se que o sistema não tem estas sequências (*perguiçoso* e *promenores*) no dicionário e obviamente não as traduz. Antes de começar a traduzir um texto através de um tradutor automático, é preciso sempre controlar a ortografia e eventualmente reescrever corretamente as palavras em que ocorram erros. No bloco de exemplos seguintes, o nome próprio *Rui* é interpretado como a terceira pessoa do singular do presente do indicativo do verbo *ruir*:

[31] O verbo *reparar* é substituído por *notar algo*, como em italiano o verbo *riparare* é equivalente de *reparar, consertar, remendar, restaurar, arranjar*. *Riparare* in Italiano |Português [em linha]. Porto: Porto Editora, 2003-2016. [consult. 2016-06-14 16:26:17]. Disponível na Internet: http://www.infopedia.pt/dicionarios/italiano-portugues/riparare.

(4a) O rui deve estar a escrever o relatório.

(4b) *Si crolla deve essere scrivere la relazione.

LC: (4c) Talvez o Rui está a escrever[32] o relatório

(4d) Forse Rui sta scrivendo la relazione.

Neste caso é preciso escrever o nome com maiúscula - (4c), para que o sistema o tome como nome próprio, como é possível observar em (4d).

É frequente encontrar frases que incluem constituintes nominais sem o determinante expresso, como no exemplo a seguir:

(5a) Ø Televisões, Ø imprensa escrita e Ø debates na rádio são palco de reflexões de especialistas.

(5b) *Televisioni, della stampa scritta e dei dibattiti nella radio sono scena di riflessioni di specialisti.

LC: (5c) As televisões, a imprensa[33] e os debates na rádio são palco de reflexões de especialistas.

(5d) Le televisioni, la stampa ed i dibattiti nella radio sono scena di riflessioni di specialisti.

Em português é possível que haja constituintes nominais sem determinante em casos nos quais em italiano podem ocorrer os partitivos, como é possível observar em (5a) e (5c), o que causa problemas de tradução automática uma vez que neste contexto o sistema não faz uma seleção adequada dos determinantes obrigatórios em italiano. Neste caso, são utilizados partitivos italianos e por isso em (5c) é preciso incluir sempre determinantes para uma tradução gramatical.

Uma outra regra geral de linguagem controlada consiste em evitar sempre a utilização de expressões com sentido figurado, como no exemplo seguinte:

[32] Para o controlo veja-se a *regra 20.1* do Anexo, p. 133.

[33] O adjetivo *escrita* foi eliminado porque redundante e causa problemas de tradução.

(6a) Por este andar, o Rui deve ser ministro antes dos trinta.

(6b) *Questo piano, Rui devono essere ministro prima dei trenta.

LC: (6c) Se continuar assim, é provável que o Rui seja ministro antes dos trinta[34].

(6d) Se continua così, è probabile che Rui sia ministro prima dei trenta.

No exemplo acima, a expressão idiomática *por este andar*, presente em (6a), foi substituída por uma expressão semanticamente equivalente, com sentido literal, ou seja, *se continuar assim*, - (6c). Isto porque o sistema não dispõe de informação sobre o sentido figurado das expressões, traduzindo literalmente cada palavra. A mesma regra foi aplicada ao exemplo que se segue:

(7a) O João está em maus lençóis.

(7b) *João é in cattivi panni.

LC: (7c) O João está numa situação complicada.

(7d) João è in una situazione complicata.

A expressão idiomática *estar em maus lençóis*, presente em (7a), foi substituída por uma expressão semanticamente equivalente e com sentido literal, *estar numa situação complicada* - (7c), pelo que - (7d) - é gramatical.

Importa dizer que na tradução das frases do *corpus* surgiram outros fenómenos linguísticos que são analisados nas secções a seguir.

[34] Para o controlo veja-se a *regra 20.2* do Anexo, p. 133.

4.2 REGRAS ESPECÍFICAS

Nas secções seguintes são tratadas especificidades da língua portuguesa em matéria de modo, modalidade e aspeto que põem problemas de tradução automática. Nas frases que constituem o *corpus* traduzido através do sistema de tradução automática SYSTRANet, ocorreram outros problemas que não cabem nos objetivos deste trabalho, mas que, embora não de forma exaustiva, são abordados nas notas de rodapé.

Na *secção 4.2.1* são analisadas especificidades relativas ao modo, divididas entre frases finitas e não finitas. No que diz respeito às frases finitas, são tratadas as frases completivas, as temporais e as construções condicionais. Na categoria das frases não finitas são analisadas as frases completivas sujeito, as restritivas, as concessivas, as temporais e as causais. Por último, as frases imperativas.

Na *secção 4.2.2*, relativa à modalidade, no que diz respeito à modalidade epistémica, vai ser analisado o verbo modal *dever* e será também analisado o uso de *ter* + *de* na expressão da modalidade deôntica e da modalidade de capacidade interna.

Na última secção, a *4.2.3*, relativa a tempo e aspeto, são analisados o pretérito perfeito simples e o *passato prossimo*, o aspeto progressivo, as expressões verbais *ir* + *gerúndio* e *andar a* + *infinitivo*.

4.2.1 MODO

Nesta secção apresenta-se um estudo contrastivo de fenómenos da língua portuguesa e da língua italiana que determinam variações no emprego dos modos verbais em diferentes tipos de frases subordinadas. A partir deste estudo, foram criadas regras de linguagem controlada para evitar modos verbais e complementadores que podem ser problemáticos na fase de tradução automática.

O modo, tal como o aspeto e o tempo, é uma categoria linguística que integra a flexão verbal em ambas as línguas e está fundamentalmente relacionada com a expressão de diferentes modalidades. Embora o português e o italiano disponham dos mesmos tipos de modo, registam-se especificidades que colocam problemas à tradução automática.

4.2.1.1 FRASES FINITAS

As frases finitas (em italiano tradicionalmente chamadas *subordinate esplicite*) são frases cujo verbo se encontra conjugado em modos finitos, nomeadamente indicativo, conjuntivo e condicional. Nesta secção são estudadas especificidades da língua portuguesa e da italiana no que diz respeito à variação do uso dos modos verbais finitos nas orações completivas, temporais e condicionais, com o objetivo de criar regras para o controlo do português para a tradução automática.

COMPLETIVAS. A frase completiva é uma frase subordinada que constitui um argumento de um dos núcleos lexicais da frase superior (Mateus *et al.,* 2003:595). As frases completivas podem ser de verbo, nome ou adjetivo. Nos exemplos serão consideradas só as construções completivas de verbo. Em português, nas frases completivas, o modo indicativo é selecionado por verbos epistémicos, percetivos, declarativos, entre outros, sendo aqui considerados apenas os primeiros. São verbos que exprimem conhecimento e crença forte, como *achar*, *acreditar*, *crer* e *pensar* e na língua portuguesa selecionam o modo indicativo. Em italiano, contrariamente, estes verbos selecionam o conjuntivo. Recorrendo ao sistema SYSTRANet, a tradução automática das frases completivas que contêm os verbos *acreditar*, *crer* e *pensar*, não coloca problema de gramaticalidade, como é possível observar no quadro a seguir:

Português	**Italiano**
O João *acredita* que a Maria tem razão	João *crede* che Maria abbia ragione
O João *crê* que a Maria tem razão	João *crede* che Maria abbia ragione
O João *pensa* que a Maria tem razão	João *crede* che Maria abbia ragione

Quadro 9. Tradução obtida através do sistema de tradução SYSTRANet das frases completivas com os verbos *acreditar*, *crer* e *pensar*.

O sistema não consegue, contudo, traduzir corretamente o verbo *achar* com o indicativo na completiva, traduzido por *trovare* (equivalente de *encontrar*). Veja-se o exemplo a seguir:

(1a) Acho que é uma boa ideia.

(1b) *Trovo che è una buona idea.

LC: (1c) Eu[35] penso que seja uma boa ideia.

(1d) Penso che sia una buona idea.

Na frase (1a) o sistema de tradução automática traduz o verbo *achar* pelo verbo italiano *trovare*, equivalente de *encontrar*. Por esta razão, nas completivas que selecionam o verbo *achar* no sentido de *pensar*, o verbo *achar* foi substituído pelo verbo *pensar*, conjugado no conjuntivo - (1c), e foram feitos outros testes para o controlo:

(2a) Acho que não é uma coisa justa.

(2b) *Trovo che non è una cosa giusta.

LC: (2c) Eu[36] penso que não seja uma coisa justa.

(2d) Penso che non sia una cosa giusta.

(3a) Eles acham que é melhor estudar na biblioteca.

(3b) *Trovano che è migliore studiare nella biblioteca.

LC: (3c) Eles[37] pensam que seja melhor estudar em[38] biblioteca.

(3d) Pensano che sia migliore studiare in biblioteca.

Nos exemplos observa-se que as traduções (2d) e (3d), que resultam das frases controladas (2c) e (3c), são gramaticais porque o verbo *achar* foi substituído pelo verbo *pensar*. Portanto, pode concluir-se que para o controlo destas construções é preciso substituir o verbo *achar* pelo verbo *pensar*, conjugado no modo conjuntivo.

[35] A realização do sujeito serve para a desambiguação de *penso* (verbo/substantivo). Com o sujeito nulo o sistema reconhece o verbo como substantivo, *penso*, equivalente de *fasciatura* em italiano.

[36] *Cf.* nota 35.

[37] *Cf.* nota 35.

[38] O determinante do SN que integra o locativo (*na* ≡ *em* + *a*) é eliminado. Em italiano, no complemento locativo, é preferível utilizar apenas a proposição.

Nas secções que se seguem, são analisadas as frases adverbiais finitas que funcionam como adjunto adverbial de outras frases e são introduzidas por conjunções subordinativas (Cunha e Cintra, 1998:406). Subsequentemente, são criadas regras de linguagem controlada para as frases temporais e para as construções condicionais, porque a língua italiana e a língua portuguesa selecionam modos verbais diferentes na formação destas frases.

TEMPORAIS. Uma diferença entre português e italiano encontra-se nos tempos verbais do modo conjuntivo, que em português são: presente, pretérito imperfeito, pretérito perfeito composto, pretérito mais-que-perfeito composto, futuro simples e futuro composto. Por seu turno, o sistema verbal italiano compreende só o presente, o *imperfetto*, o *passato*, e o *trapassato* do conjuntivo, não tendo formas de futuro (simples nem composto) do conjuntivo. Na língua portuguesa, nas frases temporais em que o evento da frase subordinada é posterior ao da subordinante, usa-se o futuro do conjuntivo, ao contrário do que acontece em italiano, em que se usa o presente do indicativo. Para o controlo, foi analisado o exemplo seguinte:

(1a) A Ana vai morar em Paris quando concluir o curso.
(1b) *Anne vivrà ∅ Parigi quando concluderà il corso.
LC: (1c) A Ana vai morar em Paris quando conclui o curso.
(1d) *Anne* vivrà a Parigi quando conclude il corso.

Na frase (1a) pode observar-se que o verbo da temporal é conjugado no futuro do conjuntivo, *concluir*, traduzido em (1b) no presente do indicativo, *concluderà*. É preciso acrescentar que em (1b) o nome *Ana* é traduzido para francês, *Anne*, fenómeno impossível de controlar. Além disso, no controlo, o futuro do conjuntivo da temporal de (1a) foi substituído pelo presente do indicativo em (1c), com consequente resultado gramatical em (1d). Foi feito também um outro teste de tradução para o mesmo fenómeno:

(2a) Vamos ao cinema quando eles saírem do trabalho.

(2b) *Andiamo al cinema quando usciranno dal lavoro.

LC: (2c) Vamos ao cinema quando eles saem do trabalho.

(2d) Andiamo al cinema quando escono dal lavoro.

Como é possível observar no exemplo, a tradução - (2d) - que resulta do controlo - (2c) - é gramatical. É possível concluir que para o controlo das frases temporais com o verbo conjugado no futuro do conjuntivo é preciso substituir esta forma pelo presente do indicativo.

CONDICIONAIS. As construções condicionais são constituídas por duas frases que têm entre si uma relação de dependência semântica: a frase condicional é a frase de cujo conteúdo proposicional depende o conteúdo proposicional da frase principal. A diferença na formação de construções condicionais entre português e italiano está no emprego dos modos verbais da frase principal, dado que em português é admitido o condicional, simples ou composto, o imperfeito do indicativo e o pretérito mais-que-perfeito composto do indicativo. Em italiano, por outro lado, é admitido só o condicional, simples ou composto.

Para o controlo, nos casos em que se verifica o imperfeito ou o pretérito mais-que-perfeito do indicativo na frase principal, é preciso substituir o indicativo pelo condicional, respeitando as restrições da *consecutio temporum*:

Simultaneidade	
Frase principal	Condicional simples
Frase condicional	Pretérito imperfeito do conjuntivo

Quadro 10. Relação de simultaneidade.

Anterioridade	
Frase principal	Condicional composto
Frase condicional	Pretérito mais-que-perfeito composto do conjuntivo

Quadro 11. Relação de anterioridade.

O uso destes tempos verbais na formação de uma construção condicional é possível também em português, portanto, respeitando estas regras, o controlo resulta eficiente. Primeiro foram analisados os casos em que entre o evento da frase principal e o da condicional há uma relação de simultaneidade (Quadro 10.), veja-se o exemplo:

(1a) Se chovesse, ia de carro.
(1b) *Se piovesse, andava da automobile.
LC: (1c) Se chovesse, iria em[39] carro.
(1d) Se piovesse, andrebbe in automobile.

Na frase (1a) o imperfeito do indicativo da frase principal foi substituído pelo condicional simples em (1c), respeitando as restrições da *consecutio temporum* de simultaneidade (Quadro10.). Isto porque - (1b) - é agramatical, dada a presença do imperfeito do indicativo na frase principal. Consequentemente, depois do controlo, a frase (1d) não envolve questões de gramaticalidade.

Foram controladas também as construções condicionais em que o evento da frase condicional é anterior ao evento da frase principal (Quadro 11.). Veja-se o exemplo que se segue:

[39] A preposição *de* foi substituída pela preposição *em*, porque em italiano a expressão correta é *andare in macchina*, ou seja, *ir em carro.*

(2a) Se não tivesse cuidado de mim, hoje tinha estado sem casa.
(2b) *Se non avesse cure di me, oggi era stato senza casa.
LC: (2c) Se eu não me fosse[40] tomado cuidado[41] de mim, hoje eu[42] teria estado sem casa.
(2d) Se non mi fossi preso cure di me, oggi sarei stato senza casa.

O verbo da frase principal de (2a) é conjugado no pretérito mais-que-perfeito composto do indicativo, pelo que - (2b) - é agramatical. Por conseguinte, em (2c) o enunciado em português foi controlado utilizando os tempos verbais que exprimem uma relação de anterioridade (Quadro 11.) e o pretérito mais-que-perfeito composto do indicativo da frase principal foi substituído pelo condicional composto em (2c). O mesmo controlo foi aplicado ao exemplo que se segue:

(3a) Se ela tivesse chegado a tempo, ela tinha visto o filme.
(3b) *Se fosse arrivata tempestivamente, aveva visto il film.
LC: (3c) Se ela tivesse chegado em tempo[43], ela teria visto o filme.
(3d) Se fosse arrivata in tempo, avrebbe visto il film.

Como se verifica, o pretérito mais-que-perfeito composto do indicativo da frase principal de (3a) foi substituído pelo condicional composto em (2c), dado que - (3b), tradução automática de (3a), é agramatical. A frase (3d), ou seja, o resultado da tradução automática de (3c), resulta, portanto, gramatical.

[40] Na formação do tempo composto, foi preciso substituir o auxiliar *ter* pelo auxiliar *ser*, dado que na língua italiana o auxiliar dos tempos compostos dos verbos reflexivos é *essere.*

[41] Problema na tradução do verbo *cuidar*. Em italiano a expressão que equivale a *cuidar* é *prendersi cura*, ou seja, literalmente em português *tomar cuidado*. Mesmo assim permanece um problema de tradução, porque traduz *cuidado* no plural, ou seja, *cure*.

[42] Realização do sujeito. Sem a realização do sujeito, o sistema traduz o verbo na terceira pessoa do singular.

[43] A expressão *a tempo* é substituída por *em tempo*, a fim de se obter *in tempo* na tradução, dado ser a expressão adequada no contexto.

4.2.1.2 FRASES NÃO FINITAS

Na língua portuguesa, uma frase não finita é um tipo de frase subordinada que não se inicia por um complementador e que tem o verbo numa das formas nominais, ou seja, no infinitivo, no gerúndio ou no particípio.

> (1a) Todos nós havemos de morrer; *basta* estarmos vivos (Cunha e Cintra, 1998:408).

A frase (1a), cujo verbo se encontra sublinhado, não é introduzida por complementador, nem o verbo se apresenta numa forma finita. É assim uma frase não finita (de infinitivo flexionado). A frase (1a) pode ser equiparada a - (1b), abaixo:

> (1b) Todos nós havemos de morrer, *basta que* estejamos vivos (Cunha e Cintra, 1998:409).

As duas frases, (1a) e (1b), são, portanto, equivalentes.

Em italiano, utilizando as definições tradicionais, dá-se o nome de *implicita* a este tipo de subordinada, que pode ser ou não introduzida por um complementador (Serianni, 2010:547), encontrando-se o verbo numa das formas nominais:

> (2a) Penso *di* fare presto (*Ibidem*).

Geralmente, o infinitivo e o gerúndio podem ser utilizados quando o evento da frase não finita é simultâneo ou anterior ao evento da principal e, por outro lado, a relação de posterioridade entre frase principal e não finita é expressa pelo particípio passado, que se encontra na frase não finita. A frase (2a), no entanto, pode ser comparada com a frase que se segue:

> (2b) Penso *che* farò presto (*Ibidem*).

A frase (2b) é uma completiva finita introduzida por *che*, cujo verbo se encontra no futuro simples do indicativo. Também em italiano, como em português pelos exemplos de (1a) e (1b), as frases (2a) e (2b) são equivalentes.

Nas secções seguintes são apresentadas regras para o controlo do modo infinitivo nas completivas sujeito, restritivas, concessivas, temporais e causais, de impossível realização em italiano por razões que se prendem com a co-referência dos sujeitos.

COMPLETIVAS SUJEITO. As frases completivas sujeito exercem a função de sujeito da frase. Para o controlo, veja-se o exemplo a seguir:

(1a) É importante estudares na biblioteca.
(1b) È importante studiare nella biblioteca.
LC: (1c) É importante que estudes em[44] biblioteca.
(1d) È importante che studi in biblioteca.

Como é possível observar no exemplo, em (1c) a frase não finita foi substituída pela correspondente forma finita introduzida por *que*, com o verbo no conjuntivo. Isto porque - (1a) - não encontra correspondência em (1b) no que respeita ao sujeito da completiva. Em português o sujeito da infinitiva é interpretável a partir da flexão, o que não se verifica em italiano, face à não existência de infinitivo flexionado. Para efeitos de confirmação, veja-se o exemplo seguinte:

(2a) É injusto eles serem castigados.
(2b) È ingiusto essere punito.
LC: (2c) É injusto que eles sejam castigados.
(2d) È ingiusto che siano puniti.

Mais uma vez, - (2a) - não encontra correspondência em (2b), sendo o sujeito da infinitiva na terceira pessoa do plural. Por esta razão, foi aplicado o controlo acima referido, pelo que a frase (2d) é gramatical. Concluiu-se, portanto, que é preciso substituir a frase não finita pela correspondente forma finita introduzida por *que* e com o verbo no conjuntivo.

[44] *Cf.* nota 38.

RESTRITIVAS. Em italiano a construção infinitiva é possível só no caso das frases adjetivas restritivas[45] introduzidas pela preposição *da* (Dardano e Trifone, 1995:469). Veja-se a equivalência:

PT: Não tenho nada *para* comer.
IT: Non ho niente *da* mangiare.

O caso acima não colocou problemas de tradução automática, dado que ambas as línguas utilizam o modo infinitivo. O problema da tradução foi encontrado, por outro lado, no caso de frases restritivas não finitas introduzidas pela preposição *a,* que não encontram correspondência em italiano, sendo necessário, para a obtenção dos resultados esperados, utilizar a correspondente forma finita, introduzida por *che*. Para o controlo, veja-se o exemplo que se segue:

> (2a) O SCIgen foi criado em 2005 por[46] investigadores a trabalharem no Instituto de Tecnologia de Massachusetts (MIT, sigla em inglês), nos Estados Unidos.
> (2b) *SCIgen è stato creato nel 2005 Ø ricercatori lavorare nell'Istituto di Tecnologia di Massachusetts (MIT, iniziale in inglese) negli Stati Uniti.
> LC: (2c) O SCIgen foi criado em 2005 pelos[47] investigadores que trabalhavam no Instituto de Tecnologia do Massachusetts (MIT, sigla em inglês), nos Estados Unidos.
> (2d) SCIgen è stato creado nel 2005 dai ricercatori che lavoravano nell'Istituto di Tecnologia del Massachusetts (MIT, iniziale in inglese), negli Stati Uniti.

[45] Embora não caiba nos objetivos deste trabalho, há quem considere estas frases como finais e não como adjetivas restritivas.

[46] Problema na tradução da forma passiva. A diferença entre as duas línguas está no uso da preposição: em português utiliza-se *por* e em italiano *da.*

[47] Inclusão do determinante no SN agente na construção passiva, neste caso *por* + [$_{SNagente}$DET N] é substituído por *da* + [$_{SNagente}$DET N].

Como é possível observar, - (2b), resultado de tradução automática de - (2a) - é agramatical, dada a ocorrência do verbo no infinitivo (*lavorare*) e a omissão do determinante (*Ø ricercatori*). Por esta razão, a frase restritiva não finita de (2a) foi substituída pela correspondente frase finita introduzida por *que*, com o verbo no indicativo - (2c). Consequentemente, a frase (2d) não envolve problemas de gramaticalidade.

Nas secções que se seguem, são tratadas as frases adverbiais não finitas, ou seja frases que são introduzidas por uma expressão prepositiva, que desempenha uma função adverbial relativamente à subordinante. Como os outros tipos de frase, também as frases adverbiais podem ocorrer na forma não finita e na forma finita quer em português quer em italiano, ainda que no caso da língua italiana, geralmente, só seja possível a ocorrência de uma frase adverbial não finita quando o sujeito da principal é co-referente do sujeito da subordinada, como acontece nos outros tipos de frases anteriormente analisados.

CONCESSIVAS INTRODUZIDAS POR *APESAR DE*. Utilizando a definição tradicional, a frase concessiva exprime um evento que contrasta com o evento da subordinante. Em italiano, as frases concessivas não finitas são constituídas pela construção *pur* + *gerúndio* só se o sujeito da principal e da subordinada são co-referentes. Em todos os outros casos, em italiano é preciso utilizar a frase finita. Em português a frase concessiva não finita é introduzida por *apesar de/não obstante,* com o verbo no infinitivo flexionado. O sistema de tradução automática SYSTRANet não consegue traduzir o infinitivo flexionado, razão por que foi necessário controlar os enunciados em português.

Para o controlo, primeiro foram consideradas as concessivas introduzidas por *apesar de,* com o verbo no infinitivo flexionado:

(1a) Apesar de estar triste, ela continua a sorrir.

(1b) *Nonostante essere triste, continua a sorridere.

LC: (1c) Embora ela esteja triste, ela continua a sorrir.

(1d) Benché sia triste, continua a sorridere.

Observa-se que - (1b), resultado de tradução automática de - (1a), envolve problemas de gramaticalidade, dada a coocorrência de *nonostante* com o verbo no infinitivo (*essere*) em vez de no conjuntivo (*sia*). Em (1c) pode observar-se que a frase concessiva não finita introduzida por *apesar de,* com o verbo no infinitivo flexionado, foi substituída pela concessiva finita introduzida por *embora*, com o verbo no conjuntivo, - (1c). Contudo, acrescenta-se que é necessário garantir a co-referência com a realização do sujeito na concessiva. O mesmo controlo foi aplicado ao exemplo que se segue:

(2a) Apesar de ter chorado, sorriu a todos os convidados.
(2b) *Nonostante avere pianto, ha sorriso a tutti gli ospiti.
LC: (2c) Embora ele tenha chorado, ele sorriu a todos os convidados.
(2d) Benché abbia pianto, ha sorriso a tutti gli ospiti.

Verifica-se que - (2b) - é agramatical porque, mais uma vez, é possível observar a coocorrência de *nonostante* com o verbo no infinitivo (*avere pianto*). Em (2c) foi aplicado o controlo proposto no exemplo anterior. Consequentemente, a frase (2d) não envolve problemas de gramaticalidade.

CONCESSIVAS INTRODUZIDAS POR *NÃO OBSTANTE*. Como referido anteriormente, em português as frases concessivas podem ser introduzidas por *não obstante,* com o verbo no infinitivo flexionado:

(1a) Não obstante ser ainda jovem, conquistou posições invejáveis.
(1b) *Tuttavia essere ancora giovane, ha conquistato posizioni invidiabili.
LC: (1c) Embora ele ainda seja jovem, conquistou posições invejáveis.
(1d) Benché ancora sia giovane, ha conquistato posizioni invidiabili.

No exemplo observa-se que - (1b), resultado da tradução automática de - (1a), é agramatical porque *tuttavia* seleciona o verbo no infinitivo (*essere*) e não no conjuntivo (*sia*). Também neste caso, o controlo - (1c) - foi feito através da substituição da concessiva não finita introduzida por *não obstante*, com o verbo no infinitivo flexionado, pela concessiva finita introduzida por *embora*, com o verbo conjugado no conjuntivo. Em - (1d) - observa-se uma tradução gramatical para italiano.

TEMPORAIS INTRODUZIDAS POR *AO*. As frases temporais exprimem uma relação temporal entre a frase principal e a subordinada. As relações temporais que este tipo de subordinada expressa são relações de anterioridade, simultaneidade e posterioridade. As frases temporais introduzidas por *ao,* com o verbo no infinitivo flexionado*,* exprimem uma relação de simultaneidade entre o evento da principal e o da temporal. Em italiano, em termos gerais é possível ter uma frase temporal não finita só quando o sujeito da principal e o da temporal são co-referentes. Assim, é possível ter o verbo no gerúndio[48] na temporal só quando os sujeitos da principal e da temporal são co-referentes, como abaixo se evidencia. Para o controlo, foi feito um primeiro teste de tradução utilizando o gerúndio na temporal:

(1a) Ao ver a estátua, senti uma das maiores emoções da minha vida.
(1b) *Vedere Ø statua, hanno sentito una delle più grandi emozioni della mia vita.
LC: (1c) Vendo a estátua, senti uma das maiores emoções da minha vida.
(1d) Vendo la statua, ho sentito una delle più grandi emozioni della mia vita.

Em (1d) é possível observar que o verbo no gerúndio não é traduzido corretamente, dado que em italiano a forma correta seria *vedendo* (equivalente de "*vendo*" em português). Isto é porque *vendo*, em português, é uma forma ambígua entre o

[48] No italiano antigo, era possível utilizar o *gerundio assoluto*, no caso de sujeitos não co-referentes (Serianni 2010:609).

gerúndio de *ver* e a primeira pessoa do singular do presente do indicativo de *vender*. O sistema, por esta razão, interpreta-o como primeria pessoa do singular do presente do indicativo de *vender*. Segue-se que, para evitar qualquer tipo de ambiguidade, é preciso controlar o enunciado em português e é preciso utilizar a forma finita apropriada, i.e, introduzida por *quando*, com o verbo no indicativo. Para o controlo, foi retomado o exemplo anterior:

(1a) Ao ver a estátua, senti uma das maiores emoções da minha vida.
(1b) *Vedere Ø statua, hanno sentito una delle più grandi emozioni della mia vita.
LC: (1c) Quando vi a estátua, eu[49] senti uma das maiores emoções da minha vida.
(1d) Quando ho visto la statua, ho sentito una delle più grandi emozioni della mia vita.

Em (1a) observa-se que a frase temporal introduzida por *ao*, com o verbo no infinitivo flexionado não é traduzida corretamente em (1b), dada a ausência do determinante *(vedere Ø statua*). Em (1c) é utilizada a correspondente forma não finita, introduzida por *quando* e com o verbo no indicativo. Em (1d) pode observar-se que a tradução não envolve problemas de gramaticalidade. Para efeitos de confirmação, o mesmo controlo foi aplicado ao exemplo que se segue:

(2a) Ao rever o amigo, deu-lhe um longo beijo.
(2b) *Alla revisione l'amico, gli ha dato un lungo bacio.
LC: (2c) Quando reviu o amigo, deu-lhe um longo beijo.
(2d) Quando ha rivisto l'amico, gli ha dato un lungo bacio.

Verifica-se que - (2b) - não é gramatical porque o verbo *rever* é traduzido pelo substantivo *revisione* (equivalente de "*revisão*"). Depois de ter aplicado a regra

[49] Realização do sujeito. Sem a realização do sujeito, o sistema traduz o verbo na terceira pessoa do plural.

para o controlo, em (2c), observa-se uma tradução correta, em (2d). Veja-se também o exemplo:

(3a) Ao ir à universidade, encontrei a Joana.
(3b) *Andare all'università, ho trovato a Joana.
LC: (3c) Quando eu[50] ia à universidade, encontrei a Joana.
(3d) Quando andavo all'università, ho trovato Joana.

Também no exemplo acima - (3b) - envolve problemas de gramaticalidade na tradução da temporal não finita, em que o verbo é deixado no infinitivo (*andare*). Mais uma vez, em (3c) foi aplicada a regra para o controlo das frases temporais não finitas (simultaneidade), pelo que (3d) resulta gramatical.

TEMPORAIS INTRODUZIDAS POR *ATÉ*. Um outro caso em que é preciso criar regras de linguagem controlada é o das frases temporais não finitas introduzidas por *até,* em que o evento expresso na principal ocorre num intervalo de tempo cujo limite superior coincide com o limite inferior do intervalo de tempo em que ocorre o evento da temporal. Também em italiano é possível ter este tipo de construção utilizando *prima di/ fino/ finché* com o verbo no *infinitivo* só se os sujeitos da frase principal e da temporal são co-referentes. De modo a obter uma tradução gramatical para italiano foi criada uma regra de linguagem controlada em que é preciso utilizar a frase finita introduzida por *até que*, com o verbo no conjuntivo. Veja-se o exemplo:

(1a) A Maria vai esperar até eu chegar.
(1b) *Maria aspetterà fino a me arrivare.
LC: (1c) A Maria vai esperar até que eu chegue.
(1d) Maria aspetterà fino a che io arrivo.

No exemplo acima é possível observar que - (1b) - é agramatical, dado que o sistema de tradução traduz o verbo no infinitivo (*arrivare*), entre outros problemas

[50] Realização do sujeito. Sem a realização do sujeito, o sistema traduz o verbo na terceira pessoa do singular.

de gramaticalidade. Um problema ligado a este facto é também a ocorrência do pronome pessoal com função de complemento, *me,* em vez do pronome pessoal sujeito *io*. Em (1c) foi aplicada a regra para o controlo, utilizando a frase finita introduzida por *até que* e o verbo no conjuntivo, com consequente tradução gramatical, - (1d). Para efeitos de confirmação, veja-se ainda o resultado do teste seguinte:

(2a) Não vais sair até concluíres o trabalho.
(2b) *Non uscirai fino a concludere il lavoro.
LC: (2c) Não vais sair até que concluas o trabalho.
(2d) Non uscirai fino a che concludi il lavoro.

No exemplo acima o único problema de tradução que ocorreu foi a tradução da frase temporal não finita e, depois de aplicar o controlo, em (2c), a tradução resulta gramatical, - (2d).

TEMPORAIS INTRODUZIDAS POR *DEPOIS DE*. Um outro grupo de frase temporais não finitas é constituído pelas frases que são introduzidas por *depois de,* em que o evento da temporal ocorre num intervalo de tempo anterior ao da principal. Como nos outros casos, em português pode-se utilizar o infinitivo flexionado quando quer o sujeito da temporal e o da principal sejam co-referentes quer tenham referência disjunta, como é possível observar nos exemplos abaixo. Em italiano tal não é possível, portanto o sistema de tradução SYSTRANet não consegue traduzir de maneira gramatical este tipo de temporal. Para o controlo dos enunciados em português, foi feito um primeiro teste de tradução utilizando *depois + que*, com o verbo conjugado no modo indicativo:

(1a) Depois de o António ter estacionado o carro, os amigos vieram ter com ele.

(1b) *Dopo António avere parcheggiato l'automobile, gli amici sono venuti ad avere con lui.

LC: (1c) Depois que o António estacionou o carro, os amigos andaram desde ele[51].

(1d) *In seguito che Antonio ha parcheggiato l'automobile, gli amici sono andati da lui.

O resultado do controlo - (1d) - não é aceitável porque em italiano *in seguito che Antonio ha parcheggiato* é agramatical, dada a má formação de *in seguito che*. Por esta razão, foi feito um outro teste em que foi substituída a frase temporal não finita pela correspondente finita, introduzida por *depois de* + *que*. Veja-se o exemplo:

(1a) Depois de o António ter estacionado o carro, os amigos vieram ter com ele.

(1b) *Dopo António avere parcheggiato l'automobile, gli amici sono venuti ad avere con lui.

LC: (1c) Depois de que o António tem estacionado o carro, os amigos andaram desde ele.

(1d) Dopo che António ha parcheggiato l'automobile, gli amici sono andati da lui.

No exemplo acima é possível observar que a aplicação do controlo, em (1c), produz resultados satisfatórios - (1d), com a tradução da frase temporal não finita a não envolver problemas de gramaticalidade. Vejam-se ainda os resultados do teste seguinte:

[51] A expressão *ir ter com* foi substituída pela correspondente tradução literal para italiano *andar desde alguém*.

(2a) Ambos tiveram morte imediata depois de o condutor ter perdido o controlo do carro.

(2b) *I due hanno avuto decesso imediato dopo il conducente avere perso il controllo dell'automobile.

LC: (2c) Ambos tiveram morte imediata depois de que o condutor perdeu o controlo do carro.

(2d) I due hanno avuto decesso imediato dopo che il conducente ha perso il controllo dell'automobile.

Como se pode observar, também aqui os resultados produzidos são satisfatórios - (2d), pelo que se concluiu que as frases temporais não finitas introduzidas por *depois de*, com o verbo no infinitivo flexionado, devem ser substuídas pelas correspondentes finitas, introduzidas por *depois de* + *que* e com o verbo no indicativo.

CAUSAIS. As frases causais denotam a causa do evento da principal. Em português, as frases causais não finitas podem ser introduzidas por *por*, com o verbo no infinitivo. Em italiano, as causais não finitas em que o sujeito da principal e da causal são co-referentes são introduzidas por *per*, com o auxiliar no infinitivo e o verbo no particípio passado. No controlo, a causal não finita introduzida por *por*, com o verbo no infinitivo flexionado, é substituída pela causal finita introduzida por *porque,* com o verbo no indicativo. Cabe dizer que foram criadas duas regras de linguagem controlada no que diz respeito à co-referência dos sujeitos da frase principal e da frase causal. Na primeira, controlam-se as causais cujo sujeito é co-referente com o sujeito da principal, na segunda, por outro lado, controlam-se as causais cujo verbo não é co-referente com o verbo da principal. No primeiro caso, foi considerado o exemplo que se segue:

(1a) O Rui não obteve bons resultados por não ter estudado.

(1b) *Rui non ha ottenuto buoni risultati di non avere studiato.

LC: (1c) O Rui não obteve bons resultados porque não estudou.

(1d) Rui non ha ottenuto buoni risultati perché non ha studiato.

Como se verifica, a preposição *por*, em (1a), é traduzida pela preposição *di* - (1b) - e o verbo *estudar* é deixado no infinitivo, pelo que - (1b) - é agramatical. Por esta razão, a frase infinitiva foi substituída pela correspondente forma finita (no caso, com o verbo no pretérito perfeito simples) introduzida por *porque* - (1c). A tradução (1d) resulta gramatical. Contudo, para efeitos de confirmação, foi feito mais um teste de tradução aplicando a mesma regra:

(2a) A Maria ficou em casa por estar doente.

(2b) *Maria è rimasta a casa essere malato.

LC: (2c) A Maria ficou em casa porque estava doente.

(2d) *Maria è rimasta a casa perché era *malato.*

Não ocorreram problemas na tradução da frase causal finita. O único erro encontrado foi de concordância entre o sujeito, *Maria*, e o adjetivo *malato* (em vez de *malata*) em (2d). Como se pode observar na frase (2c), na frase causal o sujeito não é realizado. Razão por que foi feito um outro teste de tradução com a realização do sujeito na frase causal:

(3a) A Maria ficou em casa porque ela estava doente.

(3b) *Maria è rimasta a casa perché ∅ era *malato*.

O resultado obtido, uma vez mais, apresenta um erro de concordância entre sujeito e adjetivo, portanto concluiu-se que o erro resulta de um mau desempenho do sistema não suscetível de controlo no contexto em causa.

Nos exemplos a seguir, o sujeito da frase principal e o da subordinada não são co-referentes, portanto, em italiano é impossível ter uma frase causal não finita introduzida por *por*, com o verbo no infinitivo flexionado. No controlo, a frase causal não finita introduzida por *por*, com o verbo no infinitivo flexionado, é substituída pela causal finita introduzida por *dado que*, com o verbo no indicativo:

(4a) Eu gosto do meu pai por ser carinhoso e inteligente.
(4b) *Amo mio padre essere affettuoso ed inteligente.
LC: (4c) Eu gosto do meu pai, dado que é carinhoso e inteligente.
(4d) Amo mio padre, dato che è affettuoso ed intelligente.

Na frase (4a) aparece a expressão verbal *gostar* + *de*, em concreto *gosto de*, equivalente em italiano a *mi piace*. Neste caso, temos uma divergência estrutural, que envolve diferentes posições dos argumentos nas construções *eu gosto de/mi piace*. O verbo *gostar* é semanticamente menos forte do que o verbo *amare*, que ocorre em (4d), mas este foi o único tipo de controlo possível para que a tradução resultasse gramatical.

4.2.1.3 FRASES IMPERATIVAS

Escolheu-se incluir o controlo de frases imperativas nas regras de linguagem controlada porque foram encontrados vários problemas de tradução automática.

Na definição tradicional, o modo imperativo é um modo verbal finito que exprime a modalidade deôntica (relacionada com a ordem, podendo assumir significados de domínios similares, tais como pedido, convite, conselho, ...). O modo imperativo é usado em frases principais, coordenadas e absolutas, e não nas frases subordinadas. Em português só há a segunda pessoa do singular e a segunda pessoa do plural e, nas outras pessoas, é substituído pelo conjuntivo.

Nas secções seguintes são analisadas frases imperativas afirmativas e frases imperativas negativas, organizadas em função da variação em pessoa. No que diz respeito à desambiguação, foi analisada também a terceira pessoa do plural do conjuntivo (que no caso vertente adquire valor exortativo) para obter frases imperativas na segunda e na terceira pessoas do plural.

SEGUNDA PESSOA DO SINGULAR. Como já referido anteriormente, em português o modo imperativo só tem a segunda pessoa do singular e do plural e nas outras pessoas é substituído pelo conjuntivo. Em italiano, por outro lado, a segunda pessoa do imperativo é igual à segunda pessoa do presente do indicativo e, por isso, o sistema de tradução SYSTRANet não consegue traduzir corretamente as frases na segunda pessoa do singular do imperativo. Para o controlo, é preciso substituir a segunda pessoa do singular do imperativo pela segunda pessoa do singular do presente do indicativo, como demostrado no exemplo:

(1a) Faz o trabalho!
(1b) Fa il lavoro!
LC: (1c) Fazes o trabalho!
(1d) Fai il lavoro!

A frase (1b) é ambígua porque em italiano pode ser interpretada como terceira pessoa do singular do presente do indicativo. Por isso, para o controlo, a segunda pessoa do singular do imperativo é substituída pela segunda pessoa do singular do presente do indicativo, como se pode observar em (1c), pelo que - (1d) - não envolve problemas de gramaticalidade.

TERCEIRA PESSOA DO SINGULAR. Um outro problema de tradução é posto pela terceira pessoa do singular, expressa em português pelo conjuntivo, como é possível observar no exemplo:

(1a) Durma bem!
(1b) Dorme bene!
LC: (1c) Que durma bem!
(1d) Che dorma bene!

A frase (1b) é gramatical, porque o sistema interpreta e traduz o verbo na terceira pessoa do singular do presente do indicativo (*dorme*). Para o controlo do enunciado em português, é preciso forçar a interpretação imperativa e a terceira pessoa do singular do presente do conjuntivo da frase imperativa é precedida por *que* em (1c).

Foram feitas tentativas de tradução da frase *espero que ele durma bem*, cuja tradução é mais natural para um falante nativo de italiano, mas o sistema traduz o verbo *esperar*[52] pelo verbo *attendere*. Um outro problema de tradução foi encontrado na frase seguinte:

(2a) Lave a roupa.
(2b) Lava l'abito.
LC: (2c) Que ele lave as roupas[53].
(2d) Che ∅ lava gli abiti.

O problema ocorreu na tradução da terceira pessoa do singular do conjuntivo, em (2c), traduzida pela segunda pessoa do singular do imperativo - (2d). No controlo, foram feitos testes quer do português para italiano quer do italiano para português para tentar encontrar uma solução para o problema. A primeira tentativa de controlo foi feita do português para italiano:

(3a) Tu lavas a roupa.
(3b) *Tu lave l'abito.

Neste exemplo, foi utilizada a segunda pessoa do singular do presente do indicativo porque em italiano é igual à terceira pessoa do singular do presente do conjuntivo (*tu lavi/che egli lavi*), incluindo a realização do sujeito. O sistema, contudo, não consegue traduzir corretamente o verbo *lavar* na segunda pessoa do singular do presente do indicativo e reconhece *lavas* como substantivo plural, traduzido em italiano por *lave*, razão por que - (3b) - é agramatical. Foram ainda feitos outros testes de tradução do italiano para português:

(4a) Che egli lavi gli abiti.
(4b) *Que Ø lava os vestuários.

[52] O verbo *esperar* é ambíguo, mas em italiano há verbos diferentes para as diferentes interpretações de *esperar* - *attendere/aspettare* e *sperare*. No caso vertente, o sistema deveria selecionar *sperare*.
[53] No controlo, *roupa* foi substituído por *roupas*.

Neste exemplo, foi utilizada a terceira pessoa do singular do conjuntivo que no caso vertente adquire valor exortativo. Podemos observar a omissão do sujeito em (4b), presente em italiano - (4a), e a ocorrência do verbo na segunda pessoa do singular (*lava*), pelo que - (4b) - é agramatical.
Com vista à confirmação da consistência das restrições a adoptar procedeu-se a novo teste, que se apresenta abaixo:

(5a) Tu lavi gli abiti.
(5b) *Você lavas os vestuários.

Neste caso, é utilizada a segunda pessoa do presente do indicativo, sendo, em italiano, igual à terceira pessoa do presente do conjuntivo. O resultado deste controlo envolve problemas de gramaticalidade, dado que o sujeito de terceira pessoa do singular ocorre com o verbo conjugado na segunda pessoa do singular (*você lavas*).

Concluiu-se que a impossibilidade do controlo é consequência do mau desempenho do sistema de tradução automática aparentemente não resultante de razões linguísticas da ordem das aqui relevantes.

SEGUNDA PESSOA DO PLURAL. Uma outra questão importante é o uso em italiano da segunda pessoa do plural na conjugação verbal. Em português só se usa em alguns dialetos do Norte, enquanto em italiano faz parte do uso comum da língua. Por esta razão, foi preciso criar uma regra para a desambiguação da terceira pessoa do plural, por forma a obter-se a segunda pessoa do plural. No primeiro teste de tradução, utilizou-se a segunda pessoa do plural do imperativo:

(1a) Façam o trabalho rapidamente!
(1b) Fanno il lavoro rapidamente!
LC: (1c) Fazei o trabalho rapidamente!
(1d) Fanno il lavoro rapidamente!

O controlo proposto em (1c) não resulta eficaz, dada a ocorrência em (1d) do verbo *fare* na terceira pessoa do plural do presente do indicativo. A língua italiana, na

formação da segunda pessoa do pluarl do imperativo, seleciona a segunda pessoa do plural do presente do indicativo. Por esta razão, foi feito um outro controlo em que há que substituir a terceira pessoa do plural do presente do conjuntivo da frase imperativa pela segunda pessoa do plural do presente do indicativo. Veja-se o exemplo seguinte:

(2a) Façam o trabalho rapidamente!
(2b) Fanno il lavoro rapidamente!
LC: (2c) Façais o trabalho rapidamente!
(2d) Fate il lavoro rapidamente!

A frase (2b) está na terceira pessoa do plural e para desambiguar o verbo é preciso aplicar a regra acima mencionada, ou seja, substituir a terceira pessoa do plural do presente do conjuntivo da frase imperativa pela segunda pessoa do plural do presente do conjuntivo. Em (2d), como se observa, o resultado é gramatical.

TERCEIRA PESSOA DO PLURAL. Para o controlo foram utilizados os mesmos exemplos do controlo da segunda pessoa do plural:

(1a) Façam o trabalho rapidamente!
(1b) Fanno il lavoro rapidamente!
LC: (1c) Que façam o trabalho rapidamente!
(1d) Che facciano il lavoro rapidamente!

Como é possível observar no exemplo, no controlo foram inseridas alterações que envolvem a introdução de *que* e a realização do sujeito. Para obter uma tradução gramatical - (1d) - é preciso a realização do sujeito da frase imperativa em (1c). Isto porque o sujeito da frase (1a) é interpretado pelo sistema de tradução automática na terceira pessoa do plural do presente do indicativo.

No *corpus* analisado para a tradução da terceira pessoa do plural das frases imperativas, foi encontrado um problema no exemplo que se segue:

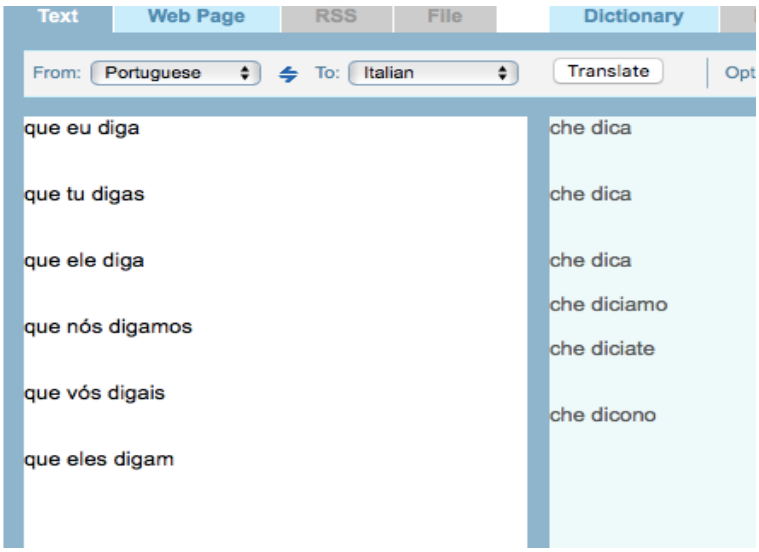
(2a) Digam a verdade!

(2b) Dicono la verità!

LC: (2c) Que eles digam a verdade!

(2d) *Che dicono la verità!

O problema ocorreu na tradução de *digam* da frase (2a), conjugado na segunda pessoa do plural do presente do conjuntivo, traduzido pela terceira pessoa do plural do presente do indicativo, *dicono* em (2d). O controlo foi feito utilizando a mesma regra do exemplo anterior, mas neste caso a tradução da terceira pessoa do plural do verbo *dizer* - (2d) - envolve problemas de tipo gramatical, porque o sistema traduz o conjuntivo *digam,* pela terceira pessoa do plural do presente do indicativo precedida por *che*, *che dicono*, estrutura mal formada. Para resolver o problema, foram feitos outros testes de tradução em que o verbo *dizer* foi conjugado no presente do conjuntivo e a seguir foi traduzido:



Quadro 12. Conjugação e tradução do verbo *dizer* no presente do conjuntivo.

Como se observa no Quadro 12., o sistema não traduz corretamente a terceira pessoa do plural. Pode-se chegar à conclusão de que este erro é uma consequência do mau desempenho do sistema de tradução automática SYSTRANet eventualmente relacionada com a análise morfológica.

SEGUNDA PESSOA DO SINGULAR EM FRASES IMPERATIVAS NEGATIVAS. É preciso fazer algumas considerações sobre as frases imperativas negativas, porque as duas línguas, na formação deste tipo de frase, selecionam modos verbais diferentes, como no caso da segunda pessoa, em que o italiano seleciona o modo infinitivo e o português o conjuntivo. Veja-se o exemplo:

(1a) Não bebas café.
(1b) Non bevi caffè.
LC: (1c) Não beber café.
(1d) Non bere caffè.

A tradução da frase (1a), ou seja, (1b), não é agramatical, porque o sistema de tradução automática traduz o verbo na segunda pessoa do presente do indicativo. O problema está na tradução do modo verbal e, para forçar a interpretação imperativa, é preciso controlar o enunciado e substituir o conjuntivo pelo infinitivo. O mesmo controlo, para efeitos de confirmação, foi aplicado ao exemplo que se segue:

(2a) Não fumes, faz mal à saúde!
(2b) Non fumi, fa male alla salute!
LC: (2c) Não fumar, faz mal à saúde!
(2d) Non fumare, fa male alla salute!

No exemplo acima ocorre o mesmo problema que se verifica com (1a), sendo necessário forçar a interpretação imperativa. Depois de ter aplicado a regra para o controlo acima mencionada, a tradução, em (2d), resulta gramatical.

SEGUNDA PESSOA DO PLURAL EM FRASES IMPERATIVAS NEGATIVAS. Nos exemplos que se seguem é controlada a terceira pessoa do plural para obter uma frase imperativa negativa com o verbo na segunda pessoa do plural, pelo que se propõe este controlo para a desambiguação da terceira pessoa do plural. No primeiro controlo, a terceira pessoa do plural do presente do conjuntivo foi substituída pela segunda pessoa do plural do presente do conjuntivo:

(1a) Não fumem, faz mal à saúde!

(1b) Non fumano, fa male alla saulte!

LC: (1c) Não fumeis, faz mal à saúde!

(1d) Non fumavate, fa male alla salute!

A frase (1b), tradução automática de - (1a), não envolve problemas de gramaticalidade, sendo *fumano* a terceira pessoa do plural do presente do indicativo. O objetivo é obter uma frase imperativa, por esta razão, em (1c) o verbo na terceira pessoa do plural do presente do conjuntivo foi subsituído pela segunda pessoa do plural do presente do conjuntivo, sendo esta a forma selecionada na língua portuguesa. Contudo, o resultado de tradução automática (1d) não é eficaz, dada a ocorrência de *fumavate* na segunda pessoa do plural do imperfeito do indicativo. Por esta razão, foi feito um outro controlo e como regra de linguagem controlada estabeleceu-se que se deve substituir a terceira pessoa do plural do presente do conjuntivo pela segunda pessoa do plural do presente do indicativo, sendo esta a forma selecionada em italiano:

(1a) Não fumem, faz mal à saúde!

(1b) Non fumano, fa male alla salute!

LC (1c) Não fumais, faz mal à saúde!

(1d) Non fumate, fa male alla salute!

A frase (2b) não envolve problemas ligados à gramaticalidade, sendo o verbo conjugado no modo indicativo. Em (2c) a terceira pessoa do plural do presente do conjuntivo é substituída pela segunda pessoa do plural do presente do indicativo, pelo que - (2d) é gramatical. Para efeitos de confirmação, veja-se:

(2a) Não mintam!

(2b) Non mentiscono[54]!

LC (2c) Não mintais!

(2d) Non mentite!

[54] Forma antiga da conjugação do verbo *mentire.* Esta forma caiu em desuso e no italiano moderno é preferível utilizar a forma *mentono*.

(3a) Não chorem!

(3b) Non piangono!

LC (3c) Não chorais!

(3d) Non piangete!

Consequentemente, como se observa no exemplo, os resultados obtidos com este controlo são satisfatórios, pelo que as frases (2d) e (3d) são gramaticais.

Na fase de teste e de tradução observou-se que ocorreram problemas para o controlo de frases imperativas negativas com o verbo *falar*:

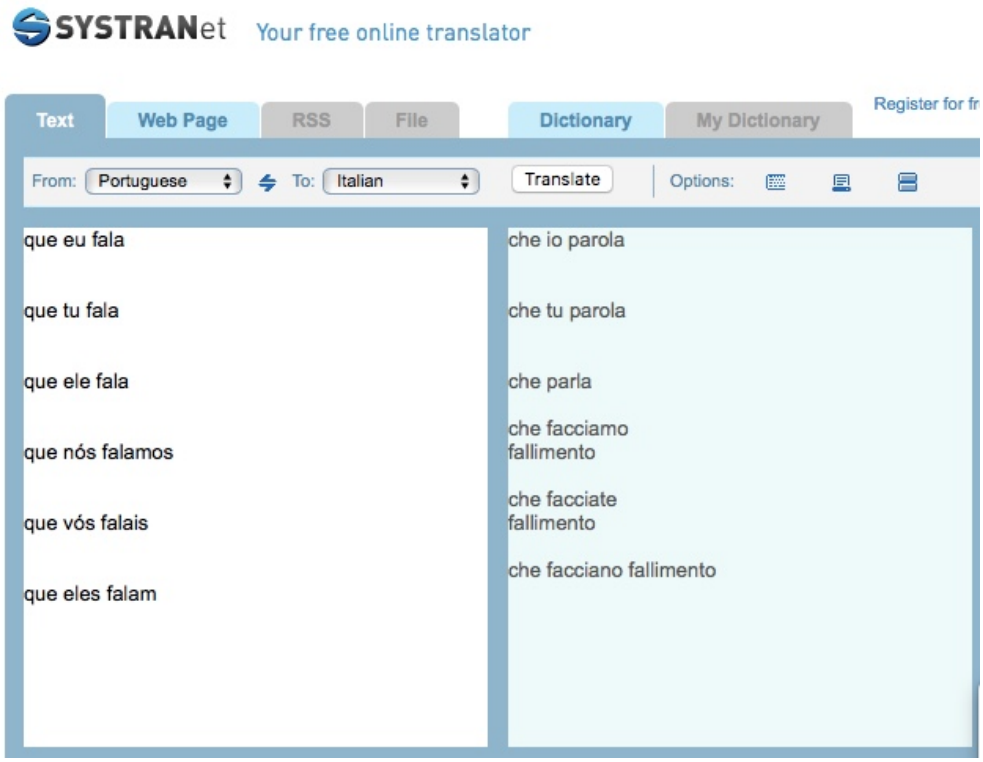
(6a) Não falem!

(6b) Non fanno fallimento!

LC: (6c) Não falais!

(6d) Non fate fallimento!

Para o controlo, o verbo foi conjugado na segunda pessoa do plural do presente do indicativo. O modo verbal é traduzido corretamente, o problema ocorre na seleção do próprio verbo, dado que o sistema seleciona o a expressão *fare fallimento*, equivalente de *falir* e não de *parlare* ("*falar"* em português). Isto porque *falais* é também a forma da segunda pessoa do plural do presente do conjuntivo de *falir*. Para testar o sistema, o verbo *falir* foi conjugado no presente do conjuntivo e a seguir foi traduzido:



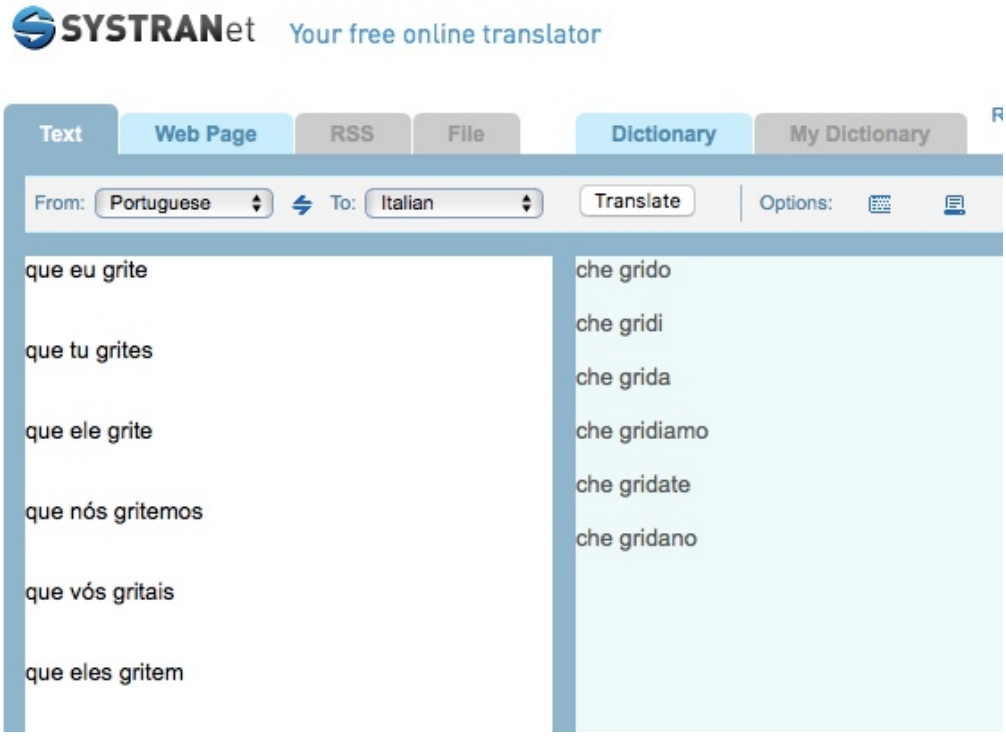
Quadro 13. Conjugação e tradução do verbo *falir* no presente do conjuntivo.

No Quadro 13. observa-se que o verbo *falir* não é conjugado corretamente no presente do conjuntivo e concluiu-se que a impossibilidade de obter uma tradução correta é uma consequência do mau desempenho do sistema de tradução automática a nível não captável nos parâmetros desta investigação.

TERCEIRA PESSOA DO PLURAL EM FRASES IMPERATIVAS NEGATIVAS. Na subsecção anterior foi proposto um controlo para segunda pessoa do plural do imperativo em frases imperativas negativas, ou seja, foi desambiguada a terceira pessoa do plural, com base nos dados apresentados. Nesta subsecção, por outro lado, procura-se controlar a terceira pessoa do plural do conjuntivo em frases imperativas negativas. Partamos do seguinte exemplo:

(1a) Não gritem!
(1b) Non gridano!
LC: (1c) Que não gritem!
(1d) *Come non gridano!

No exemplo, em (1c) foram introduzidos o complementador *que* e a realização do sujeito. Isto, porque a frase (1b) é gramatical, dado que o verbo é traduzido na terceira pessoa do plural do presente do indicativo, mas, é preciso forçar a leitura imperativa. O resultado de tradução automática - (1d) - não é satisfatório dado que o controlo produz um resultado agramatical, porque *que* é traduzido como equivalente de *como*, ou seja, *come*. Optou-se, portanto, por conjugar e traduzir no sistema de tradução automática o verbo *gritar* no presente do conjuntivo:



Quadro 14. Conjugação e tradução do verbo *gritar* no presente do conjuntivo.

No Quadro 14. observa-se que a segunda e a terceira pessoa do plural não são conjugadas no presente do conjuntivo, mas no presente do indicativo, pois as formas corretas são *che voi gridiat*e e *che loro gridino*, respetivamente. Por estas razões, procurou-se encontrar um outro controlo:

(2a) Não gritem!

(2b) Non gridano!

LC: (2c) Ordeno que não gritem!

(2d) Comando che non gridino!

A a terceira pessoa do plural do presente do conjuntivo de (2b) é precedida por *ordeno que*. O controlo resulta eficiente, sendo o verbo *gridare* de (2d) conjugado no presente do conjuntivo. Veja-se também o exemplo:

(3a) Não gritem!
(3b) Non gridano!
LC: (3c) Peço que não gritem!
(3d) Chiedo che non gridino!

Neste caso, a terceira pessoa do pluar do presente do conjuntivo é precedida por *peço que* em (3c), sendo - (3d) - gramatical. Optou-se por propor dois controlos para oferecer a possibilidade de interpretar o a frase imperativa quer como ordem - (2d) - quer como pedido - (3d).

4.2.2 MODALIDADE

Nesta secção são estudadas as diferenças no emprego de verbos modais e de outros itens lexicais na expressão da modalidade epistémica e deôntica em português e em italiano que colocam problemas de tradução automática. De seguida, são analisados e traduzidos alguns exemplos que servem como base para a criação de regras de linguagem controlada.

Em termos gerais, a modalidade é a expressão da atitude do falante (crença, esperança, obrigação, ...) no que diz respeito ao conteúdo das proposições. A modalidade pode ser expressa através de *lexical clues* (ou *modal triggers*): adjetivos, advérbios, morfologia (modo e tempo) da flexão verbal, em particular. Tradicionalmente, entre os diferentes tipos de modalidade, é dada uma maior relevância do ponto de vista linguístico a modalidade epistémica e a modalidade deôntica. A modalidade epistémica diz respeito a informação de natureza probabilística, ou seja, tem que ver com:

> "[...] il grado e la natura dell'impegno alla verità di ciò che si asserisce (che può essere verificato, probabile, possibile o falsificato)" (Tucci, 2005:2).

Isto quer dizer que o valor epistémico de um enunciado depende dos processos de conhecimento, crença e juízo avaliativo do falante, que coloca o evento denotado pelo enunciado numa escala de probabilidade. Por outro lado, a modalidade deôntica diz respeito à permissão ou obrigatoriedade de envolvimento no evento.

Cabe dizer que, dada a natureza deste trabalho, foram utilizadas as definições tradicionais de modalidade[55].

Neste trabalho, à luz das especificidades do português e do italiano, apresentam-se regras para a modalidade epistémica, deôntica e para a modalidade de capacidade interna. Por último, importa dizer que a modalidade de capacidade externa não foi analisada porque não apresenta problemas de tradução automática.

4.2.2.1 MODALIDADE EPISTÉMICA

A modalidade epistémica, em termos gerais, é a modalidade respeitante à expressão de probabilidade enformada por crença ou juízo avaliativo do falante e normalmente está relacionada com a expressão de diferentes graus de certeza sobre um determinado facto. Citando Palmer,

> "[...] with epistemic modality speakers express their judgements about the factual status of the proposition" (Palmer 1986:8).

Em português, os verbos que exprimem a modalidade epistémica são os verbos *dever* e *poder*, embora também *ter + de* e *ser capaz de* possam surgir com esta leitura em determinados contextos. Consideremos alguns exemplos ilustrativos retirados de Mateus *et al*., (2003:249):

(1) O Jorge *pode* ter chegado há minutos.
(2) O Jorge *deve* ter chegado há minutos.
(3) O Jorge *tem de* ter chegado há minutos.
(4) O Jorge *é capaz de* ter chegado há minutos.

Nas frases (1) e (2) é possível observar os verbos *poder* e *dever* em contextos epistémicos e, por outro lado, *ter + de*, em (3), é considerado epistémico só para a

[55] Para outras perspetivas ver Nuyts e Van der Auwera (2016), entre outros.

interpretação em que a afirmação decorre de um pressuposto por parte do locutor que encontra fundamento no contexto situacional. Noutros termos, o locutor exprime a sua "quase certeza" com base no seu conhecimento de que o Jorge chega sempre a horas. À frase (4) também pode ser dada uma leitura epistémica, dado o uso do presente do indicativo (*Ibidem*).

Na língua italiana, os verbos que exprimem a modalidade epistémica são os verbos *dovere* ("*dever*") e *potere* ("*poder*"), podendo o primeiro comutar com a expressão *è probabile che* ("*é provável que*") e o segundo com *è possibile che* ("*é possível que*"):

(5) Parlava bene, doveva essere una persona istruita ("*Falava bem, devia ser uma pessoa instruída*").
(6) Posso essermi sbagliato ("*Posso ter-me enganado*").

A frase (5) é equivalente de "*parlava bene, è probabile che fosse una persona istruita*" ("*falava bem, é provável que fosse uma pessoa instruída*") e, por outro lado, - (6) - é equivalente de "*è possibile che mi sia sbagliato*" ("*é possível que me tenha enganado*"). Serianni (2010) afirma que é nestes contextos que os verbos modais *dovere* e *potere* adquirem valor epistémico.

***TALVEZ* + INDICATIVO.** Nesta secção propõe-se uma regra para o controlo do modal *dever* em contextos epistémicos, sendo que o verbo *dovere* é ambíguo entre modalidade epistémica e modalidade deôntica. Por esta razão, no controlo, foi considerado "*ser capaz de*". Veja-se, o exemplo:

(1a) Atualmente, deve ser o pintor mais admirado.
(1b) Attualmente, deve essere il pittore più ammirato.
LC: (1c) Atualmente, é capaz de ser o pintor mais admirado.
(1d) *Attualmente, è capace di essere il pittore più ammirato.

O controlo testado em (1c) não resulta eficaz, sendo *è capace di essere* uma estrutura mal formada, uma vez que *è capace* deveria ocorrer com *che + conjuntivo*. Por esta razão, com o objetivo de encontrar um controlo com resultados adequados,

optou-se por substituir *dever* + *infinitivo* por *talvez* + *indicativo*. Considere-se o exemplo:

(2a) Atualmente, deve ser o pintor mais admirado.
(2b) Attualmente, deve essere il pittore più ammirato.
LC: (2c) Atualmente, talvez é o pintor mais admirado.
(2d) Attualmente, forse è il pittore più ammirato.

Antes de mais, é oportuno dizer que em italiano o advérbio *forse*, equivalente de "*talvez"*, seleciona sempre o modo indicativo, ao contrário do que acontece em português, em que *talvez* seleciona sempre o modo conjuntivo. Recorde-se que o objetivo da linguagem controlada é controlar o *input* para obter um *output* aceitável na língua de chegada e, como é possível observar no exemplo acima, foi preciso degradar a aceitabilidade do *input* (*talvez* + *indicativo*). Por esta razão, no controlo, *dever* + *infinitivo* é substituído por *talvez* + *indicativo*. Isto porque a frase (2b), resultado da tradução automática de (2a), não envolve problemas de gramaticalidade mas de ambiguidade, dado que *dever* pode adquirir um significado quer epistémico quer deôntico. Por esta razão, para evitar qualquer tipo de ambiguidade, *dever* + *infinitivo* é substituído por *talvez* + *indicativo*, envolvendo também uma mudança no uso do modo (sendo, em português, *talvez* + *conjuntivo*). O mesmo controlo foi aplicado ao exemplo que se segue:

(3a) O Rui deve estar a escrever o relatório.
(3b) *Rui deve essere scrivere la relazione.
LC: (3c) Talvez o Rui está a escrever o relatório.
(3d) Forse Rui sta scrivendo la relazione.

Como se observa, - (3b), resultado da tradução automática de - (3a), apresenta problemas de gramaticalidade, dado que o verbo modal *dever* não pode coocorrer com o verbo *essere* (equivalente, neste contexto, de "*estar*") e o verbo *scrivere,* ("*escrever*") ambos conjugados no infinitivo. Por esta razão, ao enunciado em português foi aplicado o controlo *talvez* + *indicativo* com consequente tradução gramatical em (3d). Observe-se também o exemplo:

(4a) O João está atrasado, deve ter perdido o comboio.

(4b) João è ritardato, deve aver perso il treno.

LC: (4c) O João está em atraso[56], talvez tem perdido o comboio.

(4d) João è in ritardo, forse ha perso il treno.

Mais uma vez, ocorreram problemas de ambiguidade em (4b). Neste caso, a ambiguidade é resultado da ocorrência do verbo *deve* ("*deve*") com o verbo *avere* (equivalente de "*ter*"), tornando ambígua a leitura. Foi aplicada a regra *talvez* + *indicativo* em (4c), pelo que - (4d) - resulta gramatical. Para efeitos de confirmação, foi considerado o exemplo que se segue:

(5a) Ela estuda bem, deve passar o ano.

(5b) *Studia bene, deve passare Ø.

LC: (5c) Ela estuda bem, talvez passa o ano.

(5d) Studia bene, forse passa l'anno.

No exemplo, (4b) é agramatical, dada a omissão do objeto. O problema de tradução não se colocou na tradução do verbo *dever* mas na tradução de *passar o ano*. Por esta razão, foi feito o teste que se segue:

Quadro 15. Tradução de *passar o ano*.

[56] Ambiguidade lexical. O sistema traduz a expressão *em atraso* com o adjetivo italiano *ritardato*. A expressão foi substituída por uma outra expressão, *estar em atraso.*

Verifica-se que, fora do contexto, a expressão *passar o ano* é traduzida de forma correta. No controlo foi preciso forçar a interpretação, como se observa no exemplo:

(5a) Ela estuda bem, deve passar o ano.
(5b) *Studia bene, deve passare Ø.
LC: (5c) Ela estuda bem, talvez passa o ano escolar.
(5d) Studia bene, forse passa l'anno scolastico.

Em (5c) foi aplicada a regra de controlo e o verbo *dever* + *infinitivo* foi substituído por *talvez* + *indicativo* e a interpretação foi forçada adicionando *escolar*, pelo que - (5d) - resulta gramatical.

DEVER. Propõe-se uma outra regra para o controlo de *dever* epistémico. Neste caso, retomando Serianni (2010), *dever* + *infinitivo* é substituído por *é provável que* + *conjuntivo*. Como referido anteriormente, também neste caso está envolvida uma mudança no uso do modo. Veja-se o exemplo:

(1a) Por este andar[57], o Rui deve ser ministro antes dos trinta.
(1b) *Questo piano, Rui devono essere ministro prima dei trenta.
LC: (1c) Se continuar assim, é provável que o Rui seja ministro antes dos trinta.
(1d) Se continua così, è probabile che Rui sia ministro prima dei trenta.

Em (1b), resultado da tradução automática de (1a), a agramaticalidade resulta da coocorrência do sujeito da terceira pessoa do singular, *Rui*, com o verbo conjugado na terceira pessoa do plural, *devono* (equivalente de "*devem*"). Cabe acrescentar que, em italiano, *dovere* é utilizado principalmente em contextos deônticos, ou seja,

[57] Ambiguidade lexical. Como se observa, *andar* é traduzido para italiano como *piano*, ou seja, um andar de um edifício. Por esta razão, toda a expressão *por este andar* foi substituída pela expressão *se continuar assim*. Há que evitar expressões idiomáticas na língua de partida para que a tradução resulte gramatical na língua de chegada.

quando a realização de uma ação é vista como obrigatória ou necessária podendo comutar com a expressão "*è probabile che*" ("*é provável que*") (Serianni, 2010:396). No controlo em (1c), *dever* é substituído por *é provável que* + *conjuntivo*, com consequente tradução gramatical em (1d). Isto, porque o verbo *dever* em português e o verbo *dovere* em italiano podem ser utilizados com valor epistémico mas, neste caso, em italiano é preferível utilizar a construção *è probabile che*, porque contém explicitamente o adjetivo de valor epistémico *probabile*. O mesmo controlo foi aplicado também ao exemplo que se segue:

(2a) Ela passou o ano, deve estudar bem.
(2b) *È passato l'anno, deve studiare bene.
LC: (2c) Ela superou[58] o ano, é provável que tenha estudado bem.
(2d) Ha superato l'anno, è probabile che abbia studiato bene.

Também em (2a) não é necessário, mas é provável que *ela tenha passado o ano porque estou bem*, porque podem estar envolvidos outros fatores. Por esta razão, *dever* foi substituído por *é provável que* + *conjuntivo* em (2c) e o resultado da tradução automática - (2d) - não envolve anomalias de tipo gramatical.

4.2.2.2 MODALIDADE DEÔNTICA

A modalidade deôntica baseia-se na noção de obrigação e exprime a atitude do falante perante ações que podem ser obrigatórias, permitidas ou proibidas. Em termos gerais, diz respeito às circunstâncias externas (pessoais, regras sociais ou normas...) que obrigam, permitem ou proíbem o participante a envolver-se na situação (Mateus *et al*., 2003:248). Citando Palmer,

> "[...] although Deontic modality stems from some kind of external authority such as rules or the law, typically and frequently the authority is the actual speaker, who gives permission to, or lays an obligation to the addressee" (Palmer 1986:10).

[58] Ambiguidade lexical. O verbo *passar* foi substituído por *superar*, equivalente de *superare*.

Em português, os verbos utilizados para a expressão da modalidade deôntica são os verbos *poder*, *dever* e a expressão verbal *ter* + *de*. Foram retirados de Mateus *et al*., (2003:249) alguns exemplos ilustrativos:

(1) Tu <u>podes</u> / o Rui <u>pode</u> sair já.
(2) Tu <u>deves</u> / o Rui <u>deve</u> já sair já.
(3) Tu <u>tens de</u> / o Rui <u>tem de</u> sair já.

Nas frases acima, *poder*, *dever* e *ter* + *de* são utilizados na expressão da modalidade deôntica, ou seja, em caso de permissão ou obrigação direta ou relatada (*Ibidem*).

Utilizando as definições tradicionais, em italiano, os verbos que exprimem a modalidade deôntica são os verbos *dovere* e *potere*, sendo que o primeiro é utilizado para exprimir obrigação e o segundo permissão (Serianni, 2010:396):

(4) <u>Devi</u> essere onesto con lei ("*<u>Tens de</u> ser honesto com ela*").
(5) Si <u>può</u> sapere perche non mi rispondi al telefono? ("*<u>Pode-se</u> saber por que não me atendes o telefone?"*).

Em (4) o modal *dovere* exprime uma obrigação e é equivalente a "*hai il dovere di essere onesto con lei*" ("*tens o dever de ser honesto com ela*") enquanto, por outro lado, o modal *potere* em (5) exprime uma permissão, sendo equivalente a "*è legittimo chiedere perché non mi rispondi al telefono*" ("*é legítimo saber porque não me atendes o telefone*"). Serianni (2010:396) afirma que dado que *dovere* e *potere* podem comutar com expressões que ocorrem em contextos deônticos (obrigação e permissão), eles próprios são portadores de "sentido deôntico".

TER + DE. O verbo modal *dever* tem paradigma defetivo e não é conjugado nos tempos perfeitos, que são substituídos pelas expressões verbais *ter* + *de/ter* + *que* e que ocorrem em contextos deônticos porque, de acordo com Campos (1998), a definição de valor epistémico bloqueia a sua combinação com tempos gramaticais perfeitos (Campos, 1998:127). Em italiano, por outro lado, o paradigma do verbo modal *dovere* é completo. Para o controlo da expressão verbal *ter* + *de*, veja-se o exemplo:

(1a) Ela teve de sair mais cedo para não perder o avião.

(1b) *Ha dovuto uscire prima per non perdere l'aereo.

LC: (1c) Ela é[59] devida sair mais cedo para não perder o avião.

(1d) È dovuta uscire prima per non perdere l'aereo.

No exemplo acima é possível observar que (1b), resultado da tradução automática de (1a), envolve problemas de gramaticalidade, dada a ocorrência do auxiliar *avere* ("*ter*") na formação do *passato prossimo* (*ha dovuto*). Neste caso, em italiano, o verbo *dovere* seleciona o verbo *essere* ("*ser*") como auxiliar, razão por que (1b) resulta agramatical. No controlo - (1c) -, a expressão verbal *ter* + *de* foi substituída pelo verbo *dever* conjugado no pretérito perfeito composto do indicativo, pelo que (1d) é gramatical.

4.2.2.3 MODALIDADE DE CAPACIDADE INTERNA

Lyons (1977), distingue entre dois tipos diferentes de modalidade deôntica: a modalidade deôntica subjetiva e a modalidade deôntica objetiva. O estudo de Lyons foi retomado por Verstraete (2001), o qual diz que a modalidade deôntica é de um só tipo, mas que é composta por uma vertente subjetiva e uma vertente objetiva. A vertente subjetiva envolve uma atitude do enunciador perante a necessidade ou permissividade de uma determinada ação enquanto a vertente objetiva descreve apenas a existência de uma necessidade de envolvimento de atitude do enunciador. A diferença entre estes dois tipos diferentes de vertentes está na fonte da modalidade: na vertente objetiva, a fonte da modalidade é externa ao próprio enunciador e não está relacionada com ele unívoca e diretamente; na vertente subjetiva, a fonte é interna ao próprio enunciador. Verstraete vai, no essencial, ao encontro de Palmer (1986), ou seja, ao encontro da ideia de que a modalidade deôntica tem a característica de ter duas tipologias diferentes de participação por parte do falante: a participação interna (*internal capacity*) que corresponde à vertente subjetiva, e a participação externa (*external capacity*), que corresponde à

[59] Em italiano os tempos compostos utilizam como auxiliar quer o verbo *ter* quer o verbo *ser,* e não há uma regra específica que determine a escolha do auxiliar. Neste caso, o verbo *dovere* seleciona *essere*, equivalente de *ser*, como auxiliar e, razão por que, foi necessário degradar o *input* e substituir o verbo *ter* (*tem devido*) pelo verbo *ser*.

vertente objetiva. Neste trabalho optou-se por controlar a modalidade de capacidade interna, pois a de capacidade externa não colocou problemas de tradução automática. Veja-se o exemplo:

(1a) A: Deves ir visitar o Presidente.
B: Não devo, tenho de ir. Eu prometi.
(1b) A: *Devi andare Ø visitare il Presidente.
B: *Non devo, ho. Ho promesso.
LC: (1c) A: Deves ir a[60] visitar o presidente.
B: Não devo, vou[61] próprio. Eu prometi.
(1d) A: Devi andare a visitare il presidente.
B: Non devo, vado proprio. Ho promesso.

No exemplo acima, a modalidade deôntica é de tipo subjetivo e no específico é uma capacidade interna ao próprio falante, interpretação que o *prometer* força. Ocorreram dois problemas na tradução de *dever/ter* + *de*, presentes em - (1a). O primeiro encontra-se na tradução do modal *dever* que, como referido nas secções acima, tem paradigma defetivo e na conjugação dos tempos perfetivos, na língua portuguesa, utiliza-se a expressão verbal *ter* + *de*. Um outro problema de tradução é colocado pela ocorrência do verbo *dever* e de *ter* + *de* na mesma frase (de difícil tradução para italiano também para um tradutor humano). Em italiano, para obter o mesmo tipo de modalidade expressa através do verbo *prometer*, é necessário utilizar o advérbio *próprio* (equivalente de *mesmo* em português). É importante dizer que *próprio* em português é um adjetivo, mas em italiano pode ser quer adjetivo quer advérbio e portanto, neste caso, é necessário utilizar *próprio* com função de advérbio, como é possível observar em (1c). Na primeira fase de tradução e de controlo, foi utilizado o advérbio *mesmo*, mas o sistema SYSTRANet não conseguiu traduzir corretamente a frase e traduziu *mesmo* para o italiano *anche*

[60] Foi preciso adicionar a preposição *a* na construção da frase declarativa não finita *prometi que o iria visitar* introduzida pelo verbo *prometer,* porque em italiano nas construções em que o verbo rege o infinitivo e um outro complemento, temos a construção Aux *andare* + a + V infinitivo. O mesmo acontece em alguns dialetos do português, por exemplo: "*Vou a fazer o jantar*".

[61] Introdução da conjugação do verbo *ir* na primeira pessoa do singular do presente do indicativo.

(equivalente de "*também"*em português). Por outro lado, se traduzido isoladamente e fora do contexto, a tradução resulta correta. Para concluir, podem observar-se os resultados da aplicação do controlo em (2d), que resulta gramatical.

Acrescente-se que em português é utilizada a expressão verbal *ter de* em casos de reforço do valor modal, quando se pretende exprimir um valor modal mais forte:

> "Por vezes, num mesmo enunciado, dá-se o reforço gradual do valor modal que incide sobre relações predicativas que se sucedem, quer essas relações predicativas sejam semanticamente equivalentes, quer se construam como complementares linguísticos umas das outras" (Campos, 1998:130).

Concluiu-se que nos casos em que na mesma frase estejam presentes *dever* e a expressão verbal *ter* + *de*, é preciso controlar o português e substituir a expressão verbal *ter* + *de* pelo adjetivo *próprio*.

4.2.3 TEMPO E ASPETO

Nesta secção são analisadas as categorias verbais de tempo e aspeto, e mais especificadamente o diferente uso de tempos e expressões verbais na expressão do perfetivo, do imperfetivo e do progressivo em português e em italiano, tendo em vista a criação de regras de linguagem controlada, face aos objetivos deste trabalho.

A categoria de tempo diz respeito à localização dos eventos no eixo do tempo, com referência ao momento da enunciação ou a um tempo de referência em geral explicitamente expresso, sendo que a forma mais comum de marcar essa localização é feita através dos tempos verbais (Mateus *et al*., 2003:129), ou, mais precisamente, da flexão verbal, que em português e em italiano é também portadora de informação aspetual e, por isso, a distinção entre tempo e aspeto pode ser feita morfologicamente. Podem distinguir-se três tempos gramaticais que se articulam nos seguintes intervalos: presente, passado e futuro,

> "[...] permitindo-nos falar de uma relação de anterioridade, simultaneidade ou posterioridade do tempo relativamente a um momento escolhido como o de referência e que normalmente é o da enunciação" (Mateus *et al*., 2003:130).

Como acima referido, o tempo divide-se em três intervalos: presente, passado e futuro, localizados em relação ao momento da fala (F), que corresponde ao ponto da enunciação. O ponto do evento (E) diz respeito ao tempo do acontecimento descrito pela frase (Mateus *et al.,* 2003:131) e o tempo de referência (R) serve como ponto a partir do qual se pode colocar o evento descrito. Por outro lado, o aspeto,

> "[...] fornece informações sobre a forma como é perspetivada ou focalizada a estrutura temporal interna de uma situação descrita pela frase, em particular, pela sua predicação" (Mateus *et al*., 2003:129).

Isto quer dizer que o aspeto não tem que ver com a colocação do evento num intervalo de tempo, respeita antes à forma como o evento se desenrola num dado intervalo de tempo. Tradicionalmente, quando se fala de aspeto é preciso falar também de *Aktionsart*, ou seja, de aspeto lexical, que designa o valor aspetual do próprio verbo. O aspeto, nas línguas como o português e o italiano, é gramatical e é realizado através de morfemas flexionais, enquanto a *Aktionsart* tem natureza lexical (Mateus *et al*., 2003:133). Citando Mateus *et al*.,

> "A distinção entre aspecto gramatical e aspecto lexical (ou *Aktionsart*) foi introduzida pelos Neogramáticos no século XIX para dar conta da diferença entre, por um lado, o tipo de situação e, por outro, certos efeitos produzidos por afixos (em particular, prefixos) nas línguas eslavas. Com efeito, nestas línguas e noutras, certas informações como *concluído*, *terminado*, *em curso*, por exemplo, são obtidas através de afixos ou de outros morfemas distintos que veiculam o texto" (*Ibidem*).

Não requerendo este trabalho distinções aspectuais de granularidade muito fina, são apenas tidos em conta os tipos de eventos primitivos adoptados em Marrafa (1993)[62]:

> "[...] três tipos de eventos primitivos: *estados*, *processos* e *transições*. Informalmente, um *estado* (E) é definido como um evento atómico, não avaliado a qualquer outro, uma *transição* (T) como um evento avaliado relativamente a outro evento, e um *processo* (P) como uma sequencia de eventos idênticos" (Marrafa, 1993:27).

[62] Sobre esta matéria ver também Pustejovsky (1995).

Do ponto de vista temporal, um evento

> "[...] é visto como um conjunto de períodos que pode incluir um subperíodo inicial, um subperíodo interno e um subperíodo final, assumindo-se, de acordo com Zangona (1993), que o que determina a "partição" temporal do evento - ou, por outras palavras, as suas caraterísticas aspectuais - são as mudanças de estado de cada argumento" (Marrafa, 1993:29).

Como referido anteriormente, as regras de linguagem controlada aqui apresentadas são criadas em função de variações aspetuais e por isso também do uso de tempos verbais, na expressão do aspeto perfetivo, em que tradicionalmente um evento é dado como concluído, do aspeto imperfetivo, em que não há delimitação do intervalo de tempo em que o evento ocorre, e do aspeto progressivo, que exprime eventos que estão a decorrer. Tradicionalmente, na expressão do aspeto perfetivo em português utiliza-se o pretérito perfeito simples e em italiano o *passato remoto* e o *passato prossimo*, enquanto na expressão do aspeto imperfetivo utiliza-se o imperfeito em ambas as línguas. À luz destas variações foram criadas as regras para o controlo dos enunciados português abaixo apresentadas.

PRETÉRITO PERFEITO SIMPLES E PASSATO PROSSIMO. Em termos gerais, na maioria das línguas românicas, as formas perfeitas compostas do verbo são utilizadas na expressão do aspeto perfetivo e, tradicionalmente, referem-se a eventos passados concluídos, localizados no eixo temporal num ponto próximo ao momento da fala. Em português, a semântica do pretérito perfeito composto é diferente da semântica do mesmo tempo verbal das outras línguas românicas, porque exprime a duração e a iteração de uma situação (Squartini, 1998:152). Por razões terminológicas, é importante acrescentar que, como descrito na subsecção anterior, o que Squartini denomina "*situation*" é referido neste trabalho por "*evento*". Tenham-se em consideração os seguintes exemplos:

(1) Tenho estudado imenso desde que decidi fazer o exame (Squartini, 1998:152);
(2) Ultimamente tenho comido pouco;
(3) *Tenho comido aqui umas vez/duas vezes (Squartini, 1998:152).

Nos exemplos acima, o tempo verbal exprime iteratividade num intervalo de tempo com o limite inferior definido e o limite superior aberto. Em (1) não se regista qualquer incompatibilidade porque não há nenhuma expressão adverbial aspetual; em (2) a ocorrência de *ultimamente* não envolve qualquer problema de gramaticalidade, porque é compatível com o valor aspetual referido acima; (3) é agramatical porque a expressão adverbial aspetual denota pontualidade e, em consonância com isso, intervalos de tempo fechados. A partir destas considerações, Squartini conclui que,

> "Unlike other Romance languages, in Portuguese the Perfect cannot refer to really past situations, not even when these are located in the recent past, or interpreted as experiential, as "hot news", or as triggering a Reference Time Reading of the Speech Time, or in hodiernal contexts, and in all these cases only the Simple Past can be used" (Squartini, 1998:153).

Por estas razões pode dizer-se que em português o pretérito perfeito composto tem forma imperfetiva ou que tem uma forma perfetiva com caraterísticas imperfetivas (Squartini, 1998:157):

> "The major requirement in that the CP [Compound Past] should refer to a durative or iterative situation, starting in the past and continuing up to the Speech Time. This implements the so-called inclusive meaning of the perfect, in which the event is seen as still ongoing at the Reference Time (obviously coinciding with the Speech Time, in the case of the Present Perfect), while nothing is presupposed regarding what follows it" (Squartini e Bertinetto, 1995:408).

Na língua italiana, tradicionalmente, fala-se do uso de duas formas de perfeitos: o *passato prossimo* e o *passato remoto*. É muito difícil definir as funções destes dois tempos, dada a dificuldade em estabelecer quais são as relações que ocorrem entre *passato prossimo* e *passato remoto*. Os próprios termos "*passato prossimo*" e "*passato remoto*" são muito discutidos, pois tradicionalmente o evento do *passato prossimo* desenvolve-se num intervalo de tempo próximo ao momento da fala e, por outro lado, o evento do *passado remoto* desenvolve-se num passado que não tem qualquer ligação com o momento da fala (Dardano e Trifone, 2005:355). Na verdade, estas definições são muito discutidas porque o evento

expresso pelo *passato remoto* pode ser mais recente do que o evento expresso pelo *passato prossimo*:

(4) Quattro anni fa andai a Londra.
(5) Dio ha creato il mondo (Serianni, 2010:471).

Na frase (4) o verbo *andare* é conjugando no *passato remoto*, *andai*, e, por outro lado, em (5) *creare* é conjugado no *passato prossimo*, *ha creato*. É evidente que o evento de (5), expresso através do *passato prossimo*, é anterior ao evento de (4), expresso pelo *passato remoto*. Em termos gerais e utilizando as definições tradicionais, pode dizer-se que o *passato remoto* denota um evento anterior ao momento da fala ou um evento que não tem qualquer ligação, objetiva ou psicológica, com o momento da fala. Por outro lado, o *passato prossimo* denota um evento do passado mas que não é necessariamente anterior ao momento da fala (*Ibidem*). Por último, cabe dizer que a língua italiana utiliza o *passato prossimo* na expressão do aspeto perfetivo, que indica um evento como concluído. É importante acrescentar que na língua italiana os falantes nativos preferem o uso do *passato prossimo* para se referir a ações do passado, mas o *passato remoto* é mais utilizado, para os mesmos efeitos, nas regiões do sul e na Toscana como forma dialetal (Serianni, 2010:472; Dardano e Trifone, 2005:355).

Nos exemplos que seguem é analisada a expressão do aspeto perfetivo através do uso do pretérito perfeito simples, casos em que em italiano é preciso utilizar o *passato prossimo* como forma não marcada. Normalmente, o sistema de tradução automática SYSTRANet traduz corretamente o pretérito perfeito simples para *passato prossimo* e no *corpus* utilizado como base para a criação de regras de linguagem controlada o único caso em que foi preciso controlar o pretérito perfeito simples é apresentado no exemplo a seguir:

(1a) Hoje de manhã fui ao supermercado.

(1b) *Ho questa mattina estate al supermercato.

LC: (1c) Hoje de manhã tenho andado[63] ao supermercado.

(1d) Questa mattina sono andato al supermercato.

Em (1a) ocorre o verbo *ir* conjugado no pretérito perfeito simples (*fui*), traduzido por *ho questa mattina estate* em (1b), estrutura mal formada, sendo *estate* equivalente do português *verão*. É possível observar também que o verbo *ir* é traduzido pelo verbo *avere* (equivalente de "*ter*") conjugado no presente do indicativo. Por esta razão, em (1c) o verbo *ir* foi substituído pelo verbo *andar*, conjugado no pretérito perfeito composto, pelo que - (1d) - resulta gramatical.

Por outro lado, põe-se o problema da tradução do pretérito perfeito simples por *passato remoto*, como é possível observar no exemplo:

(2a) Carducci nasceu em 1835.

(2b) *Carducci nato nel 1835.

Antes de mais, é importante mencionar uma caraterística importante do italiano moderno, ou seja,

> "In molti casi il grado di attualità di un evento trascorso è legato alla sua dislocazione sull'asse del tempo. Si è portati a rivivere più intensamente un fatto recente che non un fatto accaduto parecchio tempo fa. Caratteristica, nell'italiano moderno, l'opposizione tra «è nato» (detto di un vivente) e «nacque» (detto di chi è morto): «Alberto Abrasino è nato nel 1930» / «Giovanni Verga naque nel 1840»" (Serianni 2010:471-472).

A "oposição" de que fala Serianni pode ser vista na frase (2a), em que o *passato remoto* é utilizado para se referir à data de nascimento do poeta Giosuè Carducci. Por esta razão, em (2b) é preciso ter o verbo *nascere* (equivalente de "*nascer*") conjugado no *passato remoto*. Como é possível observar em (2b), o verbo *nasceu* é

[63] No primeiro teste de tradução o verbo *ir* foi conjugado no pretérito perfeito composto, obtendo como tradução "*stamattina sono passato al supermercato*". Neste caso, o verbo *ir* é traduzido por *passare* (equivalente de "*passar*") e, por esta razão, no controlo, foi preciso utilizar o verbo *andar* para obter o verbo *andare* (equivalente de "*ir*") conjugado no pretérito perfeito composto em (1d).

traduzido pelo particípio passado *nato* (equivalente de "*nascido*"), o que leva a pensar que o sistema reconhece o tempo passado do verbo mas que não tem a informação necessária para o traduzir corretamente. Para efeitos de confirmação foi feito mais um teste:

> (3a) Durante a guerra, os inimigos destruíram os antigos castelos da cidade.
> (3b) Durante la guerra, i nemici hanno distrutto gli antichi castelli della città.

Neste caso, - (3b) - é ambíguo, porque o verbo no pretérito perfeito simples de (3a) é traduzido pelo *passato prossimo*. Este resultado não pode ser considerado agramatical porque em italiano é muito comum a utilização do *passato prossimo* em vez do *passato remoto*, mas é preferível o uso do último porque, como referido anteriormente, é um evento ocorrido no passado que não tem qualquer ligação com o momento da fala (Dardano e Trifone, 2005:355). Neste contexto, o controlo vista a tradução do pretérito perfeito simples pelo *passato remoto* resulta sem sucesso.

ASPETO PROGRESSIVO. Na variedade europeia da língua portuguesa o aspeto progressivo pode ser expresso pela construção verbal *estar a + infinitivo* e pode ter uma morfologia perfetiva, ao contrário do que acontece na língua italiana. De facto, a construção *stare + gerúndio* teve morfologia perfetiva na língua italiana até ao seculo XIX e é considerada agramatical no italiano moderno (Squartini, 1998:73-74). No português europeu, é possível encontrar uma frase em que ocorra *estar a + infinitivo*, com o auxiliar no pretérito perfeito simples, de impossível tradução para italiano por causa de diferenças semânticas entre as duas línguas. Citando Squartini,

> "In Italian the Progressive has specialized as an aspectual imperfective marker denoting a situation as on-going at a given time. Consistent with such a requirement, it is excluded in any context in which the situation is simply durative and not visualized as on-going at a given time, independently of its combination with perfective morphology" (Squartini, 1998:76).

Para o controlo, considere-se o exemplo:

(1a) Estive a ler um romance muito interessante.
(1b) *Attacca a leggere una romanza molto interessante.
LC: (1c) Tenho lido uma novela[64] muito interessante.
(1d) Ho letto un romanzo molto interessante.

Em (1a) ocorre a construção *estar a + infinitivo,* com o verbo *estar* conjugado no pretérito perfeito simples, traduzido em (1b) por *attacca a leggere*, estrutura mal formada. Em (1a) a construção progressiva com o verbo *estar* conjugado no pretérito perfeito simples foi eliminada e substituída pelo verbo conjugado no pretérito perfeito composto, - (1c). Isto porque não foi possível manter o aspeto progressivo no controlo e optou-se por manter a morfologia perfetiva do verbo.

A construção progressiva, em português, pode coocorrer com o advérbio de tempo *ontem* e com o auxiliar *estar* conjugado no pretérito perfeito simples*,* agramatical em italiano. Isto porque o advérbio indica um evento concluído (*ontem*) que não pode ser expresso através de uma construção progressiva, que tradicionalmente indica eventos que ainda estão a decorrer. Para o controlo, veja-se o exemplo:

(2a) Ontem estive a trabalhar todo o dia.
(2b) *Ieri attacca a lavorare tutto il giorno.
LC: (2c) Ontem tenho trabalhado todo o dia.
(2d) Ieri ho lavorato tutto il giorno.

No caso da frase (2a) ocorre um advérbio temporal, *ontem*, que localiza temporalmente o evento, dado como concluído, num intervalo de tempo fechado. A frase (2a) é também durativa porque está presente a expressão *todo o dia,* que occorrendo com *ontem* induz a interpretação de que "*estar a trabalhar*" se desenvolveu durante um dia inteiro, findo o qual o evento está concluído. Em

[64] Ambiguidade lexical. O substantivo *romance* é traduzido por *romanza.* No dicionário SYSTRANet foi pesquisado do italiano para português o substantivo *romanzo*, cujo resultado foi *novela.* Por esta razão, no controlo, substituiu-se *romance* por *novela*.

português é aceitável uma frase deste tipo porque o pretérito perfeito simples (*estive*) pode ser utilizado para eventos durativos (*todo o dia*), ao contrário do italiano, em que o *passato remoto* pode ser utilizado só para eventos perfetivos e por isso não durativos. É importante também acrescentar que em italiano a construção progressiva é utilizada só em contextos imperfetivos. Por esta razão, no controlo do enunciado em português, é preciso utilizar o verbo principal no pretérito perfeito composto do indicativo, que na língua italiana é utilizado para descrever eventos concluídos num passado próximo do momento da fala. Por estas razões, a frase (1d) é gramatical. Importa recordar que, ao contrário do que acontece em italiano e em outras línguas românicas, em português o pretérito perfeito composto (equivalente do *passato prossimo* italiano) é utilizado com função iterativa para descrever a repetição de um ato ou a sua continuidade até ao presente, ao momento da fala (Cunha e Cintra, 1998:326).

A construção *estar a + infinitivo* em português pode coocorrer também com o advérbio *sempre* (equivalente de "*always*" em inglês):

> "[...] the Portuguese forms, both European *estar a + infinitive* and Brazilian *estar a + gerund*, occur in contexts admitted in Spanish and barred in Italian, such as for instance in combination with the adverbial *always* denoting the continuous, often hyperbolic, duration of a given situation, or in durative delimited situations, or with negative Imperative" (Squartini, 1998:114).

Por estas razões, optou-se por fazer o seguinte controlo:

(3a) Aquele menino está sempre a discutir com os outros.
(3b) *Quel ragazzo sta sempre discutendo con gli altri.
LC: (3c) Aquele menino discute sempre com os outros.
(3d) Quel ragazzo discute sempre con gli altri.

Na frase (3a), *estar a + infinitivo* ocorre com o advérbio *sempre*, agramatical em italiano. Por esta razão, a frase (3b), resultado da tradução automática de - (3a) - envolve problemas de gramaticalidade. Para o controlo do enunciado em português, é preciso utilizar o presente do indicativo do verbo principal que, ocorrendo com o advérbio *sempre*, denota iteratividade. O mesmo controlo foi aplicado ao exemplo que se segue:

(4a) Eles estão sempre a dizer a mesma coisa.

(4b) *Stanno sempre dicendo la stessa cosa.

LC: (4c) Eles dizem sempre a mesma coisa.

(4d) Dicono sempre la stessa cosa.

Também no exemplo observa-se que - (4b) - é agramatical, pelas razões mencionadas. Para o controlo, foi aplicada a regra acima referida, como se pode observar em (4c), pelo que - (4d) - é gramatical.

ASPETO PROGRESSIVO E FRASES IMPERATIVAS NEGATIVAS. Em italiano não se podem utilizar frases imperativas no aspeto progressivo. Tal acontece porque, tradicionalmente, o imperativo tem só o tempo presente e denota uma ação pontual.

Na tradução automática feita através do sistema SYSTRANet, no caso da construção progressiva combinada com a frase imperativa negativa, foram encontrados os mesmos problemas de tradução detetados no controlo das frases imperativas. Neste caso, o controlo foi mais complicado porque foi preciso resolver dois tipos diferentes de problemas, relacionados entre si: o controlo da frase imperativa e o da frase imperativa combinada com a construção progressiva. Veja-se o exemplo:

(1a) Não estejas a perder tempo.

(1b) Non stai perdendo tempo.

LC: (1c) Não perder[65] tempo.

(1d) Non perdere tempo.

Para o controlo do *input* foi aplicada a *regra 18* do controlo das frases imperativas negativas (ou seja, a regra do controlo da segunda pessoa do singular em frases imperativas negativas), eliminando a construção progressiva. Nos outros casos podem ser utilizadas as mesmas regras conforme o sujeito da frase em que ocorre a frase imperativa. A frase (1b) é gramatical, dado que o sistema de tradução automática traduz o verbo na segunda pessoa do presente do indicativo, *stai*

[65] Para o controlo, veja-se a *regra 18* do Anexo, p. 132.

(equivalente de "*estás*"). O que está em causa, neste caso, é a interpretação do verbo, ou seja, é um problema de ambiguidade. Por esta razão, para evitar problemas devidos à ambiguidade, aplicou-se a *regra 18* do controlo das frases imperativas negativas, como se pode observar em (1c), pelo que a frase (1d) é gramatical. O mesmo controlo foi aplicado também ao exemplo seguinte:

(2a) Não estejas a comer antes do almoço.
(2b) Non stai mangiando prima del pranzo.
LC: (2c) Não comer[66] antes do almoço.
(2d) Non mangiare prima del pranzo.

No exemplo (2a) ocorrem os problemas acima mencionados, nomeadamente a frase imperativa negativa e a construção progressiva, com consequente ambiguidade na tradução - (2b). Em (2c) foi aplicado o mesmo controlo, em que a frase imperativa negativa com construção progressiva é substituída pela frase imperativa negativa, aplicando as regras do controlo das frases imperativas. O resultado de tradução (2d) não envolve questões de tipo gramatical.

***IR* + GERÚNDIO**. A língua portuguesa permite o uso da expressão verbal *ir* + *gerúndio* em contextos télicos, iterativos e incoativos, como no caso dos exemplos a seguir:

(1a) Vai pensando na minha proposta!
(1b) *Ne penserà nella mia proposta!
LC: (1c) Começa a pensar à[67] minha proposta!
(1d) Inizia a pensare alla mia proposta!

No exemplo acima é possível observar que - (1b) , resultado da tradução automática de - (1a) - é agramatical, dada a coocorrência da partícula multifuncional *ne* com o verbo *pensare* (equivalente de "*pensar*"), conjugado na terceira pessoa do singular

[66] *Cf.* nota 65.

[67] A preposição *em* da construção *pensar em* foi substituída pela preposição *a* + *determinante*, selecionada na língua italiana.

do futuro di indicativo, *penserà* ("*pensará*"). Em (1a), *vai pensando,* tem significado incoativo, porque é marcado o ponto em que começa o evento. Em italiano não é possível utilizar este tipo de expressão verbal e pode dizer-se "*comincia a pensare alla mia proposta*" ("*começa a pensar na minha proposta*"), dado que "*começa a pensar*" é equivalente de "*vai pensando*". Por esta razão, em (2c), *ir* + *gerúndio* foi substituído por *começar a* + *infinitivo*.

***ANDAR A* + INFINITIVO**. Citando Squartini (1998:282), a expressão verbal *andar a* + *infinitivo*, ocorre com eventos e estados em contextos não iterativos e é agramatical em italiano. Por estas razões, foi preciso controlar a língua portuguesa, como se pode observar no exemplo:

(1a) O que andas a fazer?
(1b) *Ciò che tu marce da fare?
LC: (1c) Que estás a fazer?
(1d) Cosa stai facendo?

Em - (1a) - ocorre a expressão *andas a fazer*, traduzida em (1b) por *ciò che tu marce da fare*, estrutura mal formada. No controlo (1c), a expressão verbal *andar a* + *infinitivo* é substituída pela forma progressiva *estar a* + *infinitivo*, com o auxiliar *estar* conjugado no tempo correspondente ao do verbo de (1a) . Há uma diferença semântica entre - (1a) - e - (1d) , sendo - (1a) - interpretável quer como equivalente de "*o que tens feito nos últimos tempos?*" quer como equivalente de "*o que estás a fazer*?". Contudo, o controlo aplicado em (1c), que produz o resultado de tradução - (1d) - é eficiente apenas para a segunda interpretação, ou seja, "*o que estás a fazer*?". O mesmo controlo é aplicado à frase seguinte, em que a expressão verbal *andar a* + *infinitivo* é equivalente da forma progressiva:

(2a) O Jorge há dois anos vivia muito ocupado, andava a escrever um livro sobre a aviação.

(2b) *Jorge due anni vivevano ha molto occupato, andava a scrivere un libro sull'aviazione.

LC (2c) O Jorge dois anos ele faz[68] vivia muito ocupado, estava a escrever um livro sobre a aviação.

(2d) Jorge due anni fa viveva molto occupato, stava scrivendo un libro sull'aviazione.

Em (2b), resultado de tradução automática de (2a), a expressão *andava a escrever* é traduzida por *andava a scrivere*, agramatical em italiano. Em (2c) foi aplicado o controlo acima referido e a expressão verbal andar *a + infinitivo* foi substituída pela forma progressiva, com consequente tradução gramatical em (2d). No caso de (2c) o verbo *estar* utilizado na formação da construção progressiva é conjugado no imperfeito do indicativo, que pode ser utilizado no aspeto progressivo porque denota um evento não concluído, que neste caso ocorre num momento anterior ao momento da fala.

[68] Em italiano a expressão correta é *due anni fa. Fa* é a terceira pessoa do singular do presente do indicativo do verbo *fare* (equivalente de "*fazer*"). No controlo, a expressão *há dois anos* é substituída por *dois anos ele faz*, em que a realização do sujeito permite uma tradução gramatical. Neste caso, a degradação do *input* produz um *output* aceitável.

5. QUESTÕES LEXICAIS

Na fase de teste de tradução ocorreram problemas de tipo lexical, que não cabem nos objetivos deste trabalho. Contudo, optou-se por dedicar um capítulo específico à descrição dos fenómenos mais recorrentes e decidiu-se utilizar a ferramenta *My Dictionary*, disponível no próprio tradutor automático, para evitar a criação de regras específicas de linguagem controlada para o léxico, utilizando um tipo de controlo "alternativo" à linguagem controlada.

Na *secção 5.1* apresentam-se algumas considerações gerais sobre o conceito de ambiguidade, em particular, sobre os conceitos de ambiguidade lexical e de polissemia.

Na *secção 5.2* apresentam-se as problemáticas encontradas no *corpus* no que diz respeito à desambiguação lexical, propondo um controlo "alternativo" através da ferramenta *My Dictionary*, integrada no sistema de tradução automática SYSTRANet.

5.1 AMBIGUIDADE

Na tradução automática das frases do *corpus,* foram encontrados vários casos de ambiguidade lexical de difícil resolução através do uso do controlo da linguagem. Seguem-se, antes de mais, algumas considerações sobre a ambiguidade, um dos problemas de mais difícil resolução nos estudos de processamento da linguagem (Pustejovsky e Boguraev, 1996:2). Citando Marrafa,

> "Um dos problemas maiores que se põem à modelização formal e computacional das línguas naturais respeita à representação e à resolução das ambiguidades, sejam de natureza lexical, sejam de natureza sintáctica" (Marrafa, 2004:3).

O que se tem como objetivo neste capítulo é a análise da ambiguidade lexical encontrada nas frases do *corpus*, com a finalidade de a controlar para que o sistema de tradução automática consiga produzir uma tradução gramatical e semanticamente adequada. Por esta razão, fala-se da ambiguidade lexical que, em termos gerais, ocorre quando a uma palavra estão associados dois ou mais sentidos, sendo assim apropriado falar de polissemia, termo amplamente utilizado na literatura da especialidade:

> "One of the most pervasive phenomena in natural language is that of systematic ambiguity or polysemy" (Ravin e Leacock, 2000:4).

Tradicionalmente, o conceito de ambiguidade está associado ao conceito de homonímia, no sentido em que as palavras homónimas são palavras sem correlação etimológica, mas que são fonologicamente idênticas e que, na grafia, são representadas, sem motivação, pela mesma sequência de carácteres[69].

Pustejovsky (1995) retoma as considerações feitas por Weinreich (1964) e introduz os conceitos de ambiguidade contrastiva e ambiguidade complementar. Na ambiguidade contrastiva, conhecida tradicionalmente com o nome de homonímia, já referida acima, um item lexical está associado a pelo menos dois significados distintos. Pustejovsky (1995:27), para exemplificar o conceito, dá alguns exemplos:

(1a) Mary walked along the bank river.
(1b) HarborBank is the richest bank in the city.

Os dois exemplos acima mostram um caso de ambiguidade contrastiva dado que palavra *bank* em (1a) significa "*margem de um rio*" e em (1b) "*banco*" (instituição financeira). No que diz respeito à desambiguação dos sentidos, o que está em causa é o enquadramento no contexto e o conjunto de conhecimentos relativos à palavra a desambiguar, dado que ambos fornecem informações úteis à desambiguação.

O outro tipo de ambiguidade ao qual Pustejovsky se refere é o da ambiguidade complementar, na qual os sentidos da palavra apresentam polissemia complementar, ou seja, em que as leituras alternativas dos sentidos lexicais são manifestações da mesma palavra que ocorre em contextos diferentes. Vejam-se os exemplos:

[69] Para uma leitura mais aprofundada sobre a matéria veja-se Ravin e Leacock, (2000), entre outros.

(2a) The bank raised its interest rates yesterday (Pustejovksy e Boguraev,1996:3).
(2b) The store is next to the new bank (*Ibidem).*

(3a) If the store is open, check the price of coffee (Pustejovsky, 1995:28).
(3b) Zac tried to open his mouth for the dentist (*Ibidem*).

Nos dois exemplos acima está em causa a polissemia complementar: em (2a) *bank* representa a instituição, enquanto em (2b) representa o edifício que acolhe essa instituição[70] e em (3a) *open* é adjetivo ("*aberto*") e em (3b) é o infinitivo do verbo *to open* ("*abrir*"). É possível observar que no exemplo (2) não há variação de categoria (em ambos os casos *bank* é substantivo), mas em (3) a variação de categoria ocorre, dado que *open* é adjetivo em (3a) e verbo em (3b). Isto, porque na polissemia complementar há dois tipos de complementaridade de sentidos, num, no outro essa relação é transcategorial. Pustejovsky (1995) designa o primeiro caso como polissemia lógica. Citando Pustejovsky,

> "[...] complementary polysemy is a slightly broader term than logical polysemy, since the former also describes how cross-categorial senses are related, for example with the use of hammer as both a noun and a verb" (Pustejovsky, 1995:28)

À luz destas considerações sobre a ambiguidade lexical, na secção que se segue são tratados casos de ambiguidade lexical encontrados nas frases do *corpus*, desambiguado através de uma ferramenta que o próprio sistema de tradução disponibiliza, propondo assim um controlo "alternativo" ao apresentado neste trabalho.

[70] A ligação entre os sentidos de *bank* enquanto edifício e instituição é motivada por uma representaçãosemântica para nomes e adjetivos, que tem o nome de Estrutura Qualia. Para uma leitura mais aprofundada sobre a matéria ver Pustejovsky (1995).

5.2 A FERRAMENTA *MY DICTIONARY*

Nesta secção são tratados alguns casos de ambiguidade lexical que não foram abordados na secção deste trabalho relativa ao controlo do português na tradução automática para italiano, por, como se referiu, estarem fora do escopo deste trabalho. Nestas circunstâncias, optou-se por utilizar a ferramenta *My Dictionary* não só para a desambiguação, mas também para impor determinadas traduções ao sistema de tradução, para resolver alguns problemas que não foi possível resolver através das regras de linguagem controlada propostas. Isto é, basicamente, para "treinar" o sistema de tradução no sentido de encontrar uma solução para os problemas que ocorreram ao longo da fase de teste e de controlo.

A ferramenta *My Dictionary*, em primeiro lugar, foi utilizada para a resolução das ambiguidades lexicais. No quadro apresentam-se os casos encontrados nas frases do *corpus*, com as correspondentes traduções para italiano:

Português	**Italiano**
encontrar	incontrare, trovare
achar	pensare, trovare

Quadro 16. Exemplos de ambiguidade lexical, extraídos do *corpus*.

Como é possível observar no Quadro 16., os dois verbos portugueses (*encontrar* e *achar*) têm mais do que um equivalente em italiano. Como um dos objetivos da linguagem controlada é a eliminação das ambiguidades, também neste tipo "particular" de controlo os verbos foram desambiguados com o auxílio de dicionários e da Wordnet.PT[71], disponível online. Recorrendo à Wordnet.PT, foi possível encontrar a definição das unidades lexicais das traduções para inglês (úteis também para a tradução para italiano) e as relações de hiponímia e hiperonímia. A seguir, foi utilizada uma ferramenta que o próprio sistema de tradução automática disponibiliza, a função *My Dictionary*. O SYSTRANet permite a criação de um dicionário pessoal depois de efetuar o *log in* no site, gratuito para todos os usuários, que podem também beneficiar de:

[71] Rede léxico-conceptual do português, desenvolvida no Centro de Linguística da Universidade de Lisboa pelo CLG – Grupo de Computação do Conhecimento Léxico-Gramatical, disponível no site http://www.clul.ul.pt/wn/.

- Significados alternativos nos resultados de tradução automática;
- Acesso aos dicionários SYSTRANet e Larousse;
- Tradução de textos com até 3.000 palavras;
- Traduções efetuadas por tradutores profissionais, disponíveis por e-mail;
- Tradução de ficheiros com a formatação do ficheiro original;
- Tradução de RSS;
- Criação de um dicionário personalizado.

No *My Dictionary* é possível escolher a função DNT (*do not translate*) para que o sistema utilize o dicionário integrado (o dicionário personalizado será sempre a primeira escolha do sistema) e a categoria gramatical da palavra (que pode ser detetada automaticamente ou escolhida pelo utilizador). É importante que seja o próprio utilizador a escolher a categoria gramatical, porque podem ocorrer erros na deteção automática com consequentes erros de tradução. Nas traduções das frases do *corpus*, para a desambiguação do significado, foram inseridas no *My Dictionary* as seguintes entradas:

Português	**Italiano**
1. achar	pensare
2. encontrar	incontrare
3. Ana	Ana
4. boa	buona

Quadro 17. Entradas inseridas no *My Dictionary*.

O Quadro 17. mostra as entradas que foram inseridas para a desambiguação lexical dos exemplos abaixo reportados. As entradas (1) e (2) são utilizadas como ferramentas para a desambiguação de casos de ambiguidade lexical encontrados no *corpus*, a entrada (3) para impor o controlo do sistema, dado o mau desempenho constatado nas secções anteriores deste trabalho, e a entrada (4) foi inserida posteriormente e só para o "controlo" feito no *My Dictionary*. Para a desambiguação, vejam-se os exemplos:

(1a) Acho que é uma boa ideia.

(1b) *Trovo che è un idea buonista.

My Dic. (1c) *Penso che è una buona idea.*

Como se observa, a ambiguidade lexical de (1a) foi eliminada (*achar* > *trovare*/*pensare*), mas permanecem outros problemas de gramaticalidade, já apresentados neste trabalho e que serão retomados nesta secção. Na frase (1a), o adjetivo *boa* é traduzido para italiano pelo adjetivo *buonista* em (1b), o que não aconteceu na fase de teste de tradução sem ter efetuado o *log in* e, além disso, dado que *idea* é um substantivo feminino, no artigo indefinido feminino há a elisão da vogal (*a*) quando o substantivo também começa por vogal (**un idea, un'idea*). Como referido anteriormente, a ambiguidade lexical foi eliminada, mas permanecem outros problemas de gramaticalidade e, por isso, é preciso aplicar as regras de linguagem controlada criadas nas secções anteriores. Veja-se o resultado:

(1a) Acho que é uma boa ideia

(1d) Acho que seja uma boa ideia[72].

(1e) Penso che sia una buona idea.

O exemplo mostra que, também nesta versão de SYSTRANet, os resultados de tradução automática são agramaticais e para obter uma tradução gramatical é possível aplicar as regras de linguagem controlada apresentadas neste trabalho. Isto quer dizer que, embora entre a versão com conta e sem conta haja diferenças em termos de resultado de tradução, as regras de linguagem controlada aqui propostas podem considerar-se universais, o que pode ser visto também nos exemplos apresentados nesta secção. Veja-se mais um exemplo:

(2a) Quando ia à universidade, encontrei a Joana

(2b) Quando andavo all'università, trovai Joana.

My Dic. (2c) *Quando andavo all'università, incontrai Joana*.

[72] Para o controlo da linguagem veja-se a *regra 5* do Anexo, p.127.

No exemplo acima observa-se mais um caso de ambiguidade lexical, dado que o verbo *encontrar* é traduzido pelo verbo *trovare* - (2b) - e não por *incontrare*. Depois de ter inserido no *My Dictionary* o verbo *incontrare* como tradução de *encontrar*, na frase (2c) o verbo é desambiguado e a frase resulta correta. É interessante notar que o verbo *encontrar*, conjugado no pretérito perfeito simples - (2a) -, é traduzido em (2b) e em (2c) no *passato remoto*, questão que é retomada mais à frente nesta secção.

No exemplo que se segue, a ferramenta *My Dictionary* foi utilizada para resolver um problema de tradução devido ao mau desempenho do sistema na tradução dos nomes próprios:

(3a) A Ana vai morar em Paris quando conclui o curso[73].
(3b) *Anne va abitare a Parigi quando conclude il corso.
My Dic. (3c) *Ana va∅ abitare a Parigi quando conclude il corso.

Como se observa no exemplo acima, o nome *Ana* - (3a) - é traduzido para o francês *Anne* - (3b). Por esta razão, foi imposto o controlo *Ana* > *Ana* ao sistema de tradução, para que a tradução do nome resultasse correta - (3c). Além disso, - (3c) - continua a ser agramatical, dado a omissão da preposição *a* na expressão *andare ad abitare*. O controlo foi feito através de uma função do *My Dictionary* que possibilita a tradução de aquilo a que o sistema chama de *sequences*. Esta função foi utilizada para o controlo da expressão *vai morar* da frase (3c), em que permanece um problema de tradução. Por isso, primeiro foi escolhida a função *sequences* e depois como entrada *ir morar*, controlado com o equivalente italiano *andare ad abitare*. Veja-se o resultado:

(4a) A Ana vai morar em Paris quando conclui o curso.
My Dic. (4b) Ana andrà ad abitare a Parigi quando conclude il corso.

No *My Dictionary* foi inserido o verbo *ir*, da expressão *vai morar*, conjugado no modo infinitivo e o mesmo foi feito com o verbo *andare*, equivalente de "*ir*". No

[73] A frase utilizada é escrita em linguagem controlada (*regra 6*, p. 127). O exemplo foi retomando por causa da tradução para francês *Ana* > *Anne*.

exemplo é possível observar que este verbo é conjugado na terceira pessoa do singular do presente do indicativo - (4a). O sistema de tradução reconheceu o sujeito de terceira pessoa singular, traduzindo-o corretamente em (4b). Para verificar que esta ferramenta pudesse ser útil para expressões que não sejam uma espécie de "*collocations*"/combinatórias preferenciais[74], escolheu-se utilizar a categoria *sequences* para a tradução da expressão verbal *ter + de,* que tinha causado problemas de tradução na fase de criação das regras de linguagem controlada. Por outras palavras, o uso desta função pode ser uma alternativa à criação de regras específicas para a tradução destes fenómenos. No *My Dictionary,* primeiro foi escolhida a função *sequences*, e como equivalente de *ter + de* foi introduzido o verbo *dovere*. Mais uma vez, o verbo *ter* e o verbo *dovere* foram inseridos no infinitivo. Veja-se o exemplo:

(5a) Ela teve de sair mais cedo para não perder o avião.
(5b) *Lei ebbe d'uscire per non perdere l'aereo più presto.
My Dic. (5c) Ela teve de sair mais cedo para não perder o avião.
(5d) Lei *dovette* uscire prima per non perdere l'aereo.

No exemplo acima observa-se que a frase (5c) é traduzida corretamente e que o sistema de tradução automática reconheceu o verbo conjugado na terceira pessoa do singular, traduzindo-o corretamente. Por esta razão, a função *sequences* pode ser útil para a tradução de outro tipo de expressões que não as envolvem algum tipo de "fixidez", dado que o sistema consegue reconhecer o sujeito e, consequentemente, conjugar o verbo corretamente. É interessante observar que em (5d) o verbo *dovere* é conjugado no *passato remoto*, o que não acontece na versão do sistema de tradução sem a criação de uma conta, em que o verbo é conjugado no pretérito perfeito composto. Como referido anteriormente nesta secção, o mesmo fenómeno é observado nas frases (2b) e (2c). Por esta razão, para verificar a utilidade desta ferramenta na tradução do pretérito perfeito simples para *passato remoto*, foram criados e traduzidos novos exemplos:

[74] Estas expressões são aqui usadas informalmente.

(6a) O meu pai visitou a África pela primeira vez nos anos 80.
(6b) Mio padre visitò l'Africa per la prima volta negli anni 80.

(7a) Quando viajou pela primeira vez, era muito pequeno.
(7b) Quando viaggiò per la prima volta, era molto piccolo.

Nos exemplos acima observa-se que os verbos no pretérito perfeito simples, - (6a) e (7a) -, são traduzidos corretamente por *passato remoto*, - (7a) e (7b). Como último teste de tradução, foram retomados os exemplos da *secção 4.2.3*:

(8a) Carducci nasceu em 1835.
(8b) *Carducci nascere in 1835.

(9a) Durante a guerra, os inimigos destruíram os antigos castelos da cidade.
(9b) Durante la guerra, i nemici distrussero i castelli antichi della città.

Nos dois blocos de exemplos acima, é possível observar que (8b) é agramatical, dado que o verbo *nascere* é traduzido no infinitivo. Por outro lado, em (9b) o verbo no pretérito perfeito simples de (9a) é traduzido corretamente, ou seja, pelo *passato remoto* em (9b). Depois destas considerações, à luz dos resultados dos testes, pode dizer-se que esta ferramenta é útil para a tradução do pretérito perfeito simples para *passato remoto*, de impossível tradução, como demonstrado neste trabalho, na versão sem conta do sistema de tradução automática SYSTRANet.

Para concluir, é interessante observar também que as traduções feitas utilizando as duas versões do sistema (com conta e sem conta) não são equivalentes, como mostra o quadro a seguir:

Tradução SYSTRANet sem *log in*	**Tradução SYSTRANet com *log in***
1. Trovo che è una buona idea	1. *Trovo che è un idea buonista*
2. Quando andavo all'università, ho trovato Joana	2. *Quando andavo all'università, trovai Joana*
3. Anne vivrà a Parigi quando conclude il corso	3. *Anne va abitare Parigi quando conclude il corso*
4. Ha dovuto uscireprima per non perdere l'aereo	4. *Lei ebbe d'uscire per non perdere l'aereo più presto*

Quadro 18. Diferença nas traduções das duas versões do SYSTRANet.

Provavelmente, isso é devido às especifidades próprias das duas versões do sistema, motivo pelo qual se aconselha a utilização ou da versão com conta, em que é possível usufruir das funções do *My Dictionary,* ou da versão sem conta aplicando as regras de linguagem controlada propostas neste trabalho. Nesta escolha, portanto, importa avaliar os problemas de tradução que podem ocorrer, para que o processo de tradução automática não se torne demasiado complexo.

Depois de ter utilizado a função *My Dictionary*, concluiu-se que é uma ferramenta útil para o controlo do léxico, sobretudo nos casos de controlo de sequências, acrónimos e nomes próprios, traduzidos muitas vezes para francês[75], mas também na tradução de frases que apresentam problemas de ambiguidade lexical. As traduções efetuadas com o *log in* são diferentes das feitas sem *log in*, o que não quer dizer que os resultados alcançados com *log in* sejam melhores, porque em alguns casos foi preciso aplicar as regras previamente criadas e em outros foi preciso criar regras novas por surgirem novos fenómenos. Pode ser uma ferramenta útil também para a tradução do pretérito perfeito simples para o *passato remoto* porque, como se demostrou nas secções anteriores, é impossível controlá-lo com regras sem ter efetuado o *log in*. A fase de controlo e de tradução resulta, por outro lado, mais complicada porque é preciso criar um dicionário personalizado e aplicar as regras de linguagem controlada. Por esta razão, em todos os outros casos, é preciso seguir primeiro as regras e as restrições gerais, assim como as sugestões de redação de textos presentes na secção *Help Center* do site do SYSTRANet e, a

[75] Nos exemplos (1b) e (1d), *Ana* é traduzido para francês, *Anne*.

seguir, aplicar as regras de linguagem controlada criadas sem ter efetuado o *log in* no site. Como referido anteriormente, é oportuno avaliar os problemas de tradução do texto a traduzir porque, como demostrado nesta secção, há casos em que o controlo imposto ao sistema de tradução através do *My Dictionary* resulta mais eficiente. Ou seja, é possível utilizar a ferramenta *My Dictionary* como "alternativa" à linguagem controlada só em casos em que não haja a possibilidade de obter uma tradução gramatical através da aplicação das regras de linguagem controlada que foram propostas neste trabalho.

6. CONCLUSÕES

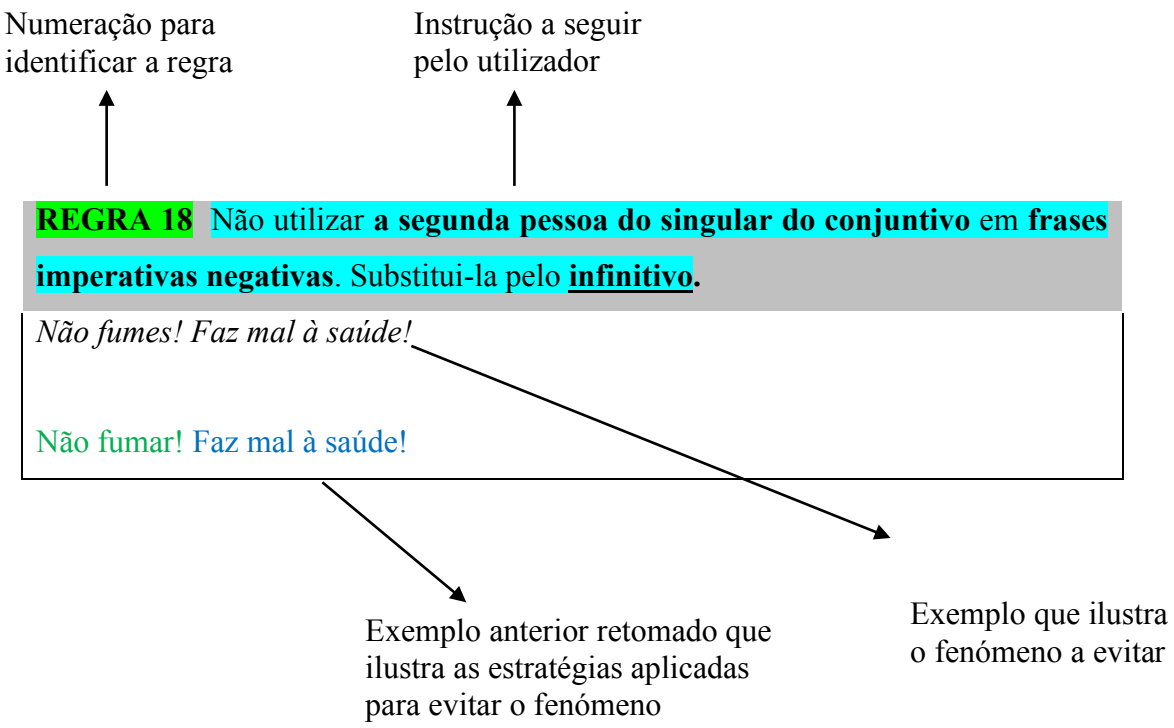
O objetivo deste trabalho é demostrar a possibilidade de criar um conjunto de regras para o controlo do português aplicáveis à tradução automática para italiano. Estas regras foram criadas a partir de um estudo das especificidades das duas línguas, que permitiu a criação de um *corpus* de exemplos que foi traduzido e testado no sistema de tradução automática SYTRANet. A fase de criação dos exemplos permitiu identificar as estruturas que causam problemas de tradução, verificados na primeira fase de teste. Na segunda fase, ou seja, a da criação e aplicação das regras, foi testado o controlo e as respetivas traduções. Esta etapa foi a mais importante porque permitiu a identificação de outras estruturas problemáticas, nomeadamente os casos de ambiguidade lexical. Além disso, foi testado também o desempenho do sistema de tradução e foi possível identificar fenómenos impossíveis de controlar, como a tradução de nomes próprios para francês (Ana >*Anne*), mas também problemas na tradução de verbos conjugados no imperativo e no conjuntivo. No caso da tradução de verbos no modo conjuntivo, foram tentadas várias formas de controlar a língua mas os resultados não foram satisfatórios. Houve casos de ambiguidade lexical e foi preciso pesquisar os termos equivalentes no dicionário integrado no próprio sistema de tradução, que foram utilizados no controlo. É interessante notar que, neste caso, o dicionário português > italiano apresenta um número inferior de entradas lexicais do que o dicionário italiano > português e por isso foi mais complicado encontrar alternativas que pudessem resultar na fase de teste do exemplo escrito em linguagem controlada. Como referido anteriormente, o português controlado apresentado neste trabalho analisa só determinados fenómenos relativos a modo, modalidade, tempo e aspeto. Apesar de o *corpus* ser bastante restrito, foi possível observar que na maioria dos casos o desempenho do sistema foi satisfatório. Por outro lado, houve casos em que foi necessário testar o mesmo exemplo ao longo do tempo. Esta fase de teste gerou problemas no que diz respeito à criação de regras para o controlo, porque foi preciso encontrar uma "solução" que resultasse estável. Em geral, depois de várias fases de teste, o controlo aplicado à língua portuguesa produziu resultados interessantes e satisfatórios, o que permitiu formalizar as regras e criar o fragmento restrito de português controlado aqui apresentado. O sistema de tradução SYTRANet permite também aos utilizadores a criação de uma conta gratuita online, que disponibiliza

funções extra. Escolheu-se testar, com a criação de uma conta, a ferramenta *My Dictionary* para a criação de um dicionário pessoal do utilizador, que permite a tradução de verbos, palavras, siglas, acrónimos e sequências inseridas diretamente dentro do sistema. Neste último tipo de controlo, foram testados todos os exemplos do *corpus* e foram escolhidos os que apresentavam problemas de ambiguidade lexical, resolvidos através da criação do dicionário personalizado *My Dictionary*. Isto permitiu também verificar que as duas versões do sistema (com conta e sem conta) não traduzem as frases da mesma maneira e foi interessante notar que um dos problemas verificados na primeira fase de teste, ou seja, o da tradução do pretérito perfeito simples para *passato remoto*, foi resolvido com o uso do sistema com a criação de uma conta. No que diz respeito ao desempenho do sistema, as regras de controlo foram aplicadas também nesta última fase, mas, como referido anteriormente, foi preciso criar e aplicar mais uma regra de controlo. Em termos gerais, pode-se concluir que a ferramenta *My Dictionary* pode ser utilizada nos casos em que não é possível aplicar as regras do português controlado, mas a tarefa de tradução torna-se mais longa e complexa.

Conclui-se que, dado que os resultados de tradução foram satisfatórios e, na maioria dos casos, estáveis ao longo do tempo, é possível continuar a aprofundar e a ampliar as regras para o controlo do português na tradução automática para italiano. É um desafio interessante e uma área ainda pouco explorada. As regras apresentadas neste trabalho podem servir como base para a criação de uma versão de português controlado aplicável a domínios específicos, como pode ser o caso da escrita técnica para a tradução automática ou da utilização por empresas para traduzir, por exemplo, documentação técnica para italiano. Contudo, dado que os resultados foram satisfatórios, o fragmento de linguagem controlada aqui proposto pode ser retomado e ampliado, na perspetiva futura de criar um sistema automático para a redação de textos em linguagem controlada, com o objetivo de tornar o *output* de tradução automática cada vez melhor.

ANEXO

A. ESTRUTURA DAS REGRAS DE LINGUAGEM CONTROLADA[76]



[76] *Cf.* Marrafa *et al.*, (2011).

B. REGRAS GERAIS

REGRA 1 Utilizar frases curtas e com estrutura simples quando possível.

Estão todos entre as dez espécies desta lista elaborada por um grupo de especialistas internacionais e que ontem foi divulgada pelo International Institute of Species Explorations do ESF, College of Environmental Science and Forestry, de Nova Iorque, para celebrar o dia de nascimento, a 23 de maio, de Carolus Linnaeus, que no século XVIII criou a moderna taxonomia - a classificação das espécies.

Estão todos entre as dez espécies desta lista elaborada por um grupo de especialistas internacionais. A lista foi divulgada ontem pelo International Institute of Species Explorations do ESF de Nova Iorque, para celebrar o dia de nascimento de Carolus Linnaeus, no dia 23 de maio. Linnaeus, no curso do século XVIII, criou a taxonomia moderna, ou seja a classificação das espécies.

REGRA 2 Escrever as frases utilizando sempre a ortografia correta.

O Miguel é um rapaz muito perguiçoso.

O Miguel é um rapaz muito preguiçoso.

O Bruno repara sempre nos promenores.

O Bruno nota sempre os pormenores.

O rui deve estar a escrever o relatório.

O Rui deve estar a escrever o relatório.

REGRA 3 Escrever as frases incluindo sempre os determinantes.

Ø Televisões, Ø imprensa escrita e Ø debates na rádio são palco de reflexões de especialistas.

As televisões, a imprensa e os debates na rádio são palco de reflexões de especialistas.

REGRA 4 Não usar expressões com sentido figurado. Usar sempre expressões com sentido literal.
Por este andar, o Rui deve ser ministro antes dos trinta. Se continuar assim, é provável que o Rui seja ministro antes dos trinta. *O João está em maus lençóis.* O João está numa situação complicada.

C. REGRAS ESPECÍFICAS

REGRA 5 Não utilizar ***achar* + indicativo** na **completiva finita**. Substituí-lo por ***pensar* + conjuntivo** na **completiva finita,** incluindo a **realização do sujeito**.

Acho que é uma boa ideia.

Eu penso que seja uma boa ideia.

Acho que não é uma coisa justa.

Eu penso que não seja uma coisa justa.

Eles acham que é melhor estudar na biblioteca.

Eles pensam que seja melhor estudar em biblioteca.

REGRA 6 Não utilizar **frases temporais finitas** com ***quando* + futuro do conjuntivo.** Substituir ***quando* + futuro do conjuntivo** por ***quando* + verbo no presente do indicativo**.

A Ana vai morar em Paris quando concluir o curso.

A Ana vai morar em Paris quando conclui o curso.

Vamos ao cinema quando eles saírem do trabalho.

Vamos ao cinema quando eles saem do trabalho.

REGRA 7 Não utilizar na **frase principal** de uma **construção condicional o verbo no imperfeito do indicativo.** Substituí-lo pelo **verbo no condicional simples**.

Se chovesse, ia de carro.

Se chovesse, iria em carro.

REGRA 8 Não utilizar na **frase principal** de uma **construção condicional** o **verbo no pretérito mais-que-perfeito do indicativo.** Substituí-lo pelo **verbo no condicional composto**.

Se não tivesse cuidado de mim, hoje tinha estado sem casa.

Se eu não me fosse tomado cuidado de mim, hoje eu teria estado sem casa.

Se ela tivesse chegado a tempo, ela tinha visto o filme.

Se ela tivesse chegado em tempo, ela teria visto o filme.

REGRA 9 Não utilizar **frases completivas sujeito** com o **verbo no infinitivo flexionado.** Substituir a **frase** o **verbo no infinitivo flexionado** pela **forma finita** com a estrutura **que + *conjuntivo***.

É importante estudares na biblioteca.

É importante que estudes em biblioteca.

É injusto eles serem castigados.

É injusto que eles sejam castigados.

REGRA 10 Não utilizar **frases restritivas não finitas** introduzidas por ***a***. Substituir a **frase restritiva não finita** com a estrutura ***a* + infinitivo flexionado** pela **forma finita** com a estrutura ***que* + indicativo**.

SCIgen foi criado em 2005 por investigadores a trabalharem no Instituto de Tecnologia de Massachusetts (MIT, sigla em inglês), nos Estados Unidos.

O SCIgen foi criado em 2005 pelos investigadores que trabalhavam no Instituto de Tecnologia do Massachusetts (MIT, sigla em inglês), nos Estados Unidos.

REGRA 11 Não utilizar **frases concessivas não finitas**.

REGRA 11.1 Não utilizar **frases concessivas não finitas** introduzidas por ***apesar de.*** Substituir a **frase concessiva não finita** com a estrutura ***apesar de* + infinitivo flexionado** pela **forma finita** com a estrutura ***embora* + conjuntivo**.

Apesar de estar triste, ela continua a sorrir.

Embora ela esteja triste, ela continua a sorrir.

Apesar de ter chorado, sorriu a todos os convidados.

Embora ele tenha chorado, ele sorriu a todos os convidados.

REGRA 11.2 Não utilizar **frases concessivas não finitas** introduzidas por ***não obstante.*** Substituir a **frase concessiva não finita** com a estrutura ***não obstante* + infinitivo flexionado forma finita** com a estrutura ***embora* + conjuntivo**.

Não obstante ser ainda jovem, conquistou posições invejáveis.

Embora ele ainda seja jovem, conquistou posições invejáveis.

REGRA 12 Não utilizar **frases temporais não finitas**.

REGRA 12.1 Não utilizar **frases temporais não finitas** introduzidas por ***ao.*** Substituir a **frase temporal não finita** com a estrutura ***ao* + infinitivo flexionado** pela **forma finita** com a estrutura ***quando* + indicativo**.

Ao ver a estátua, senti uma das maiores emoções da minha vida.

Quando vi a estátua, eu senti uma das maiores emoções da minha vida.

Ao rever o amigo, deu-lhe um longo beijo.

Quando reviu o amigo, deu-lhe um longo beijo.

Ao ir à universidade, encontrei a Joana.

Quando eu ia à universidade, encontrei a Joana.

REGRA 12.2 Não utilizar **frases temporais não finitas** introduzidas por ***até.*** Substituir a **frase temporal não finita** com a estrutura ***até* + infinitivo flexionado** pela **forma finita** com a estrutura ***até que* + conjuntivo**.

A Maria vai esperar até eu chegar.

A Maria vai esperar até que eu chegue.

Não vais sair até concluíres o trabalho.

Não vais sair até que concluas o trabalho.

REGRA 12.3 Não utilizar **frases temporais não finitas** introduzidas por ***depois de.*** Substituir a **frase temporal não finita** com a estrutura ***depois de* + infinitivo flexionado** pela **finita** com a estrutura ***depois de que* + indicativo**.

Depois de o António ter estacionado o carro, os amigos vieram ter com ele.

Depois de que o António tem estacionado o carro, os amigos andaram desde ele.

Ambos tiveram morte imediata depois de o condutor ter perdido o controlo do carro.

Ambos tiveram morte imediata depois de que o condutor perdeu o controlo do carro.

REGRA 13 Não utilizar **frases causais não finitas** com o verbo no **infinitivo.**

REGRA 13.1 Não utilizar **frases causais não finitas** introduzidas por ***por*** quando o **sujeito da frase principal e da frase causal são co-referentes**. Substituir a **frase causal não finita** com a estrutura ***por* + infinitivo flexionado** pela **forma finita** com a estrutura ***porque* + indicativo**.

O Rui não obteve bons resultados por não ter estudado.

O Rui não obteve bons resultados porque não estudou.

A Maria ficou em casa por estar doente.

A Maria ficou em casa porque estava doente.

REGRA 13.2 Não utilizar **frases causais não finitas** introduzidas por ***por*** quando o **sujeito da frase principal e da frase causal não são co-referentes**. Substituir a **frase causal não finita** com a estrutura ***por* + infinitivo flexionado** pela **forma finita** com a estrutura ***dado que* + indicativo**.

Eu gosto do meu pai por ser carinhoso e inteligente.

Eu gosto do meu pai, dado que é carinhoso e inteligente.

REGRA 14 Não utilizar **a segunda pessoa singular do imperativo** em **frases imperativas.** Substituí-la pela **segunda pessoa do singular do presente do indicativo.**

Faz o trabalho!

Fazes o trabalho!

REGRA 15 Não utilizar **a terceira pessoa singular do presente conjuntivo** em **frases imperativas.** Substitui-la por ***que* + segunda pessoa do singular do presente do conjuntivo.**

Durma bem!

Que durma bem!

REGRA 16 Não utilizar **a terceira pessoa plural do conjuntivo** em **frases imperativas.** Substitui-la pela **segunda pessoa do plural do presente do conjuntivo.**

Façam o trabalho rapidamente!

Façais o trabalho rapidamente!

REGRA 17 Não utilizar **a terceira pessoa do plural do presente do conjuntivo** em **frases imperativas**. **Fazê-la preceder** pelo complementador ***que***.

Façam o trabalho rapidamente!

Que façam o trabalho rapidamente!

REGRA 18 Não utilizar **a segunda pessoa do singular do conjuntivo** em **frases imperativas negativas.** Substituí-la pelo **infinitivo**.

Não bebas café!

Não beber café!

Não fumes! Faz mal à saúde!

Não fumar! Faz mal à saúde!

REGRA 19 Não utilizar a terceira pessoa do plural em frases imperativas negativas.

REGRA 19.1 Não utilizar a **terceira pessoa do plural do conjuntivo** em **frases imperativas negativas**. Substitui-la pela **segunda pessoa do plural do indicativo**.

Não fumem, faz mal à saúde.

Não fumais, faz mal à saúde.

Não mintam!

Não mintais!

Não chorem!

Não chorais!

REGRA 19.2 Não utilizar a **terceira pessoa do plural do conjuntivo** em **frases imperativas negativas**. **Fazê-la preceder** de ***ordeno que/peço que***.

Não *gritem*!

Ordeno que não gritem!

Não *gritem*!

Peço que não gritem!

REGRA 20 **Não utilizar *dever* com valor epistémico.**

REGRA 20.1 Não utilizar ***dever* + infinitivo** em **contextos epistémicos.** Substitui-lo por ***talvez* + indicativo**.

Atualmente, deve ser o pintor mais admirado.

Atualmente, talvez é o pintor mais admirado.

O Rui deve estar a escrever o relatório.

Talvez o Rui está a escrever o relatório.

O João está atrasado, deve ter perdido o comboio.

O João está em atraso, talvez tem perdido o comboio.

Ela estuda bem, deve passar o ano.

Ela estuda bem, talvez passa o ano escolar.

REGRA 20.2 Não utilizar ***dever* + infinitivo** em **contextos epistémicos.** Substitui-lo por ***é provável que* + conjuntivo**.

Por este andar, o Rui deve ser ministro antes dos trinta.

Por este andar, é provável que o Rui seja ministro antes dos trinta.

Ela passou o ano, deve estudar bem.

Ela superou o ano, é provável que tenha estudado bem.

REGRA 21 Não utilizar ***ter + de*** com **valor deôntico**. Substitui-la pelo verbo ***dever***, conjugado no modo e no tempo adequado.

Ela teve de sair mais cedo para não perder o avião.

Ela é devida sair mais cedo para não perder o avião.

REGRA 22 Não utilizar ***ter + de*** para a expressão da **modalidade de capacidade interna**. Substitui-lo por ***próprio***.

A: *Deves ir visitar o Presidente.*

B: *Não devo, tenho de ir. Eu prometi.*

A: Deves ir a visitar o presidente.

B: Não devo, vou próprio. Eu prometi.

REGRA 23 Não utilizar o **pretérito perfeito simples** como forma marcada no **aspeto perfetivo**. Substitui-lo pelo **pretérito perfeito composto.**

Hoje de manhã fui ao supermercado.

Hoje de manhã tenho andado ao supermercado.

REGRA 24 Não utilizar o verbo ***estar*** no **pretérito perfeito simples** na **construção progressiva *estar a* + *infinitivo*** para a expressão do **aspeto progressivo.**

REGRA 24.1 Não utilizar o verbo ***estar*** no **pretérito perfeito simples** na **construção progressiva *estar a* + infinitivo** para a expressão do **aspeto progressivo**. Substituir ***estar a* + infinitivo** pelo **verbo** no **pretérito perfeito composto**.

Estive a ler um romance muito interessante.

Tenho lido uma novela muito interessante.

REGRA 24.2 Não utilizar o verbo ***estar*** no **pretérito perfeito simples** como **auxiliar da forma progressiva** combinada com ***ontem***. Utilizar o **pretérito perfeito composto do verbo principal.**

Ontem estive a trabalhar todo o dia.

Ontem tenho trabalhado todo o dia.

REGRA 24.3 Não utilizar o **a forma progressiva** combinada com ***sempre***. Utilizar o **presente do indicativo do verbo principal**.

Aquele menino está sempre a discutir com os outros.

Aquele menino discute sempre com os outros.

Eles estão sempre a dizer a mesma coisa.

Eles dizem sempre a mesma coisa.

REGRA 25 Não utilizar a **frase imperativa negativa** com a **construção progressiva *estar a* + *infinitivo*** no **aspeto progressivo**. **Eliminar** a **construção progressiva e aplicar as regras para o controlo das frases imperativas**.

Não estejas a perder tempo.

Não perder tempo.

Não estejas a comer antes do almoço.

Não comer antes do almoço.

REGRA 26 Não utilizar ***ir* + gerúndio** em **contextos incoativos**. Substituir por ***começar a* + infinitivo**.

Vai pensando na minha proposta!

Começa a pensar à minha proposta!

REGRA 27 Não utilizar a expressão verbal ***andar* + infinitivo** em **contextos não-iterativos.** Substitui-la pela **construção progressiva *estar a* + infinitivo**, com o verbo ***estar*** no **indicativo,** conjugado no tempo adequado.

O que andas a fazer?

O que estás a fazer?

O Jorge há dois anos vivia muito ocupado, andava a escrever um livro sobre a aviação.

O Jorge dois anos ele faz vivia muito ocupado, estava a escrever um livro sobre a aviação.

REFERÊNCIAS BIBLIOGRÁFICAS

ALPAC (1966). *Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee*, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington D.C., National Academy of Sciences, National Research Council.

ARNOLD D., BALKAN L., MEIJER S., HUMPHREYS R.L., SADLER L. (1994). *Machine Translation: an Introductory Guide*. London, NCC Blackwell Ltd.

AZEVEDO M.A. (2004). *Portuguese. A Linguistic Introduction*. Cambridge, University Press.

BANJAR S.H. (2004). *Controlled Language and Machine Translation*. Assiut University, Bulletin of the Faculty of Arts, vol. 17, July 2004, pp. 34-91. Disponível em: http://www.academia.edu/1043565/Controlled_Language_and_Machine_Translation.

BECKER M.G., REMBERGER E.M. (2010). *Mood and Modality in Romance: Mood Interpretation, Mood Selection, and Mood Alternation*. Berlin, De Gruyter.

BERTINETTO P.M., EBERT K.H., DE GROOT C. (1995). The Progressive in Europe. In: Dahl O. (ed.), *Tense and Aspect in the Languages of Europe*. Berlin, De Gruyter, pp. 517-558.

BHATTACHARYYA P. (2012). Natural Language Processing: A Perspective from Computation in Presence of Ambiguity, Resource Constraint and Multilinguality. In: *CSI, Journal of Computing*, Vol.1, No. 2, 2012. Disponível em: https://www.cse.iitb.ac.in/~pb/papers/csi-nlp-pb-8aug12.pdf.

BHATTACHARYYA P. (2014). *Machine Learning for Machine Translation*. CSE Dept., IIT Bombay, ISI Kolkata, 6 January 2014. Disponível em: http://www.isical.ac.in/~acmsc/TMW2014/P_bhattacharyya.pdf.

BHATTACHARYYA P. (2015). *Machine Translation*. Boca Raton, CRC Press.

CLARK P., MURRAY W.R., HARRISON P., THOMPSON J. (2009). Naturalness vs. Predictability: a Key Debate in Controlled Languages. In: Workshop on Controlled Natural Language CNL 2009, Marettimo Island, Italy, June 8-10, *Controlled Natural Language,* Fuchs N.E. (ed.). Berlin, Springer, pp. 65-81. Disponível em: http://www.cs.utexas.edu/users/pclark/papers/cnl09.pdf.

COSTA CAMPOS, M.H. (1998). *Dever e Poder: um subsistema modal do português.* Lisboa, Fundação Calouste Gulbenkian.

COSTA-JUSSÀ M.R., FONOLLOSA J.A.R. (2015). Last trends in hybrid machine translation and its applications. In: Moore R.K. (ed.) *Computer Speech and Language*, vol. 32, Iss. 1, July 2015. Amsterdam, Elsevier, pp. 3-10. Disponível em: http://www.sciencedirect.com/science/article/pii/S0885230814001077.

CRABBE S. (2010). Controlled Languages for Technical Writing and Translation. In: Ninth Annual Portsmouth Translation Conference, Portsmouth, November 2009, *The Changing Face of Translation: Proceedings of the Ninth Annual Portsmouth Translation Conference*. Kemble I. (ed.). Portsmouth, University of Portsmouth, pp. 48-62. Disponível em: http://www.port.ac.uk/media/contacts-and-departments/slas/events/tr09-crabbe.pdf.

CUNHA C., CINTRA L. (1998). *Breve Gramática do Português Contemporâneo*. Lisboa, Edições Sá da Costa.

DARDANO M., TRIFONE P. (1995). *Grammatica Italiana con Nozioni di Linguistica*, 3ª edição. Bologna, Zanichelli.

DEANE P. (1988). Polysemy and Cognition. In: *Lingua, an International Review of General Linguistics*, vol. 75, July 1988, pp. 325-361.

DORR B.J., JORDAN P.W., BENOIT.W. (1999). A Survey of Current Paradigms in Machine Translation, In: Zelkowitz M.V. (ed.). *Advances in Computers*, Vol. 49. Amsterdam, Elsevier, pp. 1-68.

DOWTY D.R. (1979). *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Dodrecht, Holland, D. Reidel Publishing Company.

DUGAST L., SENELLART J., KOEHN P. (2008). Can we relearn an RBMT system? In: ACL 2008 Workshop on Statistical Machine Translation (WMT-08), June 2008, Columbus Ohio. *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 175-178. Diponível em: https://aclweb.org/anthology/W/W08/W08-0327.pdf.

ESPAÑA-BONET C., COSTA-JUSSÀ M.R. (2016). Hybrid Machine Translation Overview. In: Costa-jussà M.R., Rapp R., Lambert P., Eberle K., Banchs R.E., Babych B. (eds.) *Hybrid Approaches to Machine Translation*. London, Springer, pp. 1-26.

GASPARINI-BASTOS S.D (2014). *Distinções entre modalidade deôntica objetiva e subjetiva no português falado: o caso do verbo Dever*. São Jorge do Rio Preto, Universidade Estadual Paulista. Disponível em: http://llp.bibliopolis.info/confluencia/rc/index.php/rc/article/view/19/22.

GOMES DE OLIVEIRA R., ANASTASIOU D. (2011). Comparison of SYSTRAN and Google Translate for English-Portuguese. *Revista Tradumàtica: tecnologies de la traducció*, Traducció i software lliure, Número 09, Desembre 2011, pp. 118-136. Disponível em: http://www.raco.cat/index.php/Tradumatica/article/viewFile/248906/333152.

HARTLEY A., TATSUMI M., ISAHARA H., KAGEURA K., MIYATA R. (2012). Readability and Translatability Judgments for 'Controlled Japanese'. In: Proceedings of the 16th EAMT Conference, 28-30 May, Trento, pp. 237-244. Disponível em: http://hltshare.fbk.eu/EAMT2012/html/Papers/56.pdf.

HENDRICKX I., MENDES A., MENCARELLI S. (2012). Modality in a Text: a Proposal for Corpus Annotation. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC '12), Instabul. Disponível em: http://www.clul.ul.pt/files/amalia_mendes/modal_lrec2012_b.pdf.

HOGEWEG L., DE HOOP H., MALCHUKOV A. (2009). *Cross-Linguistic Semantics of Tense, Aspect and Modality*. Amsterdam and Philadelphia, John Benjamins Publishing Company.

HUTCHINS J.W. (2000). Machine Translation. In: Ralstion A., Reilly E.D., Hemmendinger D. (eds.), *Encyclopedia of Computer Science, 4th Edition*. New York, Grove's Dictionaries, pp. 1059-1066.

HUTCHINS J.W. (2002). *The state of machine translation in Europe and future prospects*. HLT Central, January 2002. Disponível em: http://hutchinsweb.me.uk/HLT-2002.pdf.

HUTCHINS J.W. (2005). *The history of machine translation in a nutshell*. Disponível em: http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.

HUTCHINS J.W. (2010). Machine Translation: A Concise History. In: *Journal of Translation Studies 13*, vol. 1-2, Special issue: The teaching of computer-aided translation, Chan Sin Wai (ed.). Chinese University of Hong Kong, pp.29-70.

HUTCHINS W.J., SOMERS H.L. (1992). *An Introduction to Machine Translation*. London, Academic Press.

IBRAHIMO N. (2010). *Para uma Tradução Automática baseada em Conhecimento: especificação da modificação e da predicação adjetival*, Dissertação de Mestrado. Lisboa, Universidade de Lisboa

JURAFSKY D., J.H. MARTIN (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognitions*. New Jersey, Prentice Hall.

KAJI H. (1999). Controlled Languages for Machine Translation: State of the Art. In: *Proceedings of MT Summit VII: MT in the Great Translation Era*, September 1999, Singapore, pp. 37-39. Disponível em: http://www.mt-archive.info/MTS-1999-Kaji.pdf.

KIT C., PAN H., WEBSTER J.J. (2002). Example-based Machine Translation: A New Paradigm. In: Sin-wai C. (ed.), *Translation and Information Technology*. Hong Kong, The Chinese University Press, pp. 57- 78.

KITTREDGE R. (2003). Sublanguages and Controlled Languages. In: Mitkow R. (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford, Oxford University Press, pp. 430-447.

KLIMOVA E. (2006). *Note sulla modalità del verbo dovere*. Études romanes de Brno, prací Filozofické fakulty brněnské univerzity. Řada L, romanistická, vol. 55, pp. 51-60. Disponível em: https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/113496/1_EtudesRomanesDeBrno_36-2006-1_6.pdf?sequence=1.

KUHN T. (2014). A Survey and Classification of Controlled Natural Language. In: *Computational Linguistics 40*, pp. 121-170. Disponível em: http://www.aclweb.org/anthology/J14-1005.

LOCKE W.N., BOOTH A.D (eds.) (1955). Machine Translation of languages: fourteen essays. Cambridge, Massachusetts, MIT Press.

LOPEZ A. (2008). *Statistical Machine Translation*. ACM Computing Surveys (CSUR), Vol. 40, Iss. 3, Article 8, August 2008. New York, ACM. Disponível em: http://dl.acm.org/citation.cfm?id=1380586.

MARQUES R. (1995). *Sobre o valor dos modos conjuntivo e indicativo em português*, Dissertação de Mestrado. Lisboa, Universidade de Lisboa.

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

MARRAFA P. (2004). Computação de ambiguidades sintácticas. Evidências em favor dos modelos baseados em conhecimento linguístico. In: *In Cognito*, Vol. 2.1, 2004, pp.1-10.

MARRAFA P., AMARO R., FREIRE N., MENDES S. (2012). Portuguese Controlled Language: Coping with Ambiguity. In: Third International Workshop, CNL 2012, Zurich, Switzerland, August 2012, Proceedings. *Controlled Natural Language*, Kuhn T., Fuchs N.E. (eds.). Berlin, Springer, pp. 152-166.

MARRAFA P., AMARO R., MENDES S., IBRAHIMO N. (2011). *CLG – Português controlado para tradução automática e para ensino/aprendizagem do Português*. Lisboa, CLUL/Instituto Camões.

MATEUS M.H.M., BRITO A.M., DUARTE I., FARIA I.H. (2003). *Gramática da Língua Portuguesa*, 5ª edição. Lisboa, Editorial Caminho.

MITAMURA T. (1999). Controlled Language for Multilingual Machine Translation. In: *Proceedings of MT Summit VII: MT in the Great Translation Era*, September 1999, Singapore, pp. 46-52. Disponível em: http://www.lti.cs.cmu.edu/Research/Kant/PDF/MTSummit99.pdf.

MITAMURA T., NYBERG E. (1995). Controlled English for Knowledge-Based MT: Experience with the KANT System. In: *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation* - TMI-95, July 1995, Leuven, Belgium, pp. 158-172. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.1257&rep=rep1&type=pdf.

NAGAO M.A. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In: A. Elithorn and R. Banerji (eds.), *Artificial and Human Intelligence*. Amsterdam, Elsevier, pp. 173–180.

NIRENBURG S., SOMERS H., WILKS Y. (2003). *Readings in Machine Translation*. Massachusetts, The MIT Press.

NUYTS J., VAN DER AUWERA J., (2016). *The Oxford Handbook of Modality and Mood*. Oxford, Oxford University Press.

NYBERG E., MITAMURA T., HUIJSEN W.O. (2003). Controlled Language for Authoring and Transaltion. In: Somers H. (ed.), *Computers and Translation*. Amsterdam and Philadelphia, John Benjamins Publishing Company, pp. 245-281.

PALMER F.R. (1986). *Mood and Modality.* Cambridge, University Press.

PING K. (2009). Machine Translation. In: Baker M., Saldanha G. (eds.), *Routledge Encyclopedia of Translation Studies*. London and New York, Routledge, pp- 162-169.

PUSTEJOVSKY J, BOGURAEV B. (1996). Introduction: Lexical Semantics in Context. In: Pustejovsky J., Boguraev B. (eds.) *Lexical Semantics. The Problem of Polysemy*. Oxford, Clarendon Paperbacks, pp. 1-14.

PUSTEJOVSKY J. (1995). *The Generative Lexicon*. Cambridge, Massachusetts, London, England, The MIT Press.

PUSTEJOVSKY J. (2005). *Introduction to Generative Lexicon*. Disponível em: https://www.cse.iitb.ac.in/~pb/papers/csi-nlp-pb-8aug12.pdf.

QUAH C. (2006). *Translation and Technology*. London, Palgrave Macmillian.

QUARESMA P., MENDES A., HENDRICKX I., GONÇALVES T. (2014). Tagging and Labelling Portuguese Modal Verbs. In: *Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014*, São Carlos/SP, Brazil, October 6-8 2014, Proceedings, Baptista J., Mamede N., Candeias S., Paraboni I., Pardo T.A.S (eds.). Springer, London, pp. 70-81.

RAVIN Y., LEACOCK C. (2000). Polysemy: an Overview. In: Ravin Y., Leacock C. (eds.) *Polysemy: Theoretical and Computational Approaches*. Oxford, Oxford University Press, pp. 1-29.

REUTHER U. (2003). *Two in one - can it work? Readability and Translatability by means of Controlled Language*. EAMT-CLAW03, Dublin City University, 15-17 May 2003, pp.124-132. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.515.8031&rep=rep1&type=pdf.

ROCCI A. (2005). *On the nature of epistemic readings od the Italian modal verbs: the relationship between propositionality and inferential discourse relation*. Disponível em: https://ssl.lu.usi.ch/entityws/Allegati/pdf_pub1569.pdf.

ROTHSTEIN B., THIEROFF R. (2010). *Mood in the Languages of Europe*. Wien, John Benjamins Publishing Company.

SENELLART J., DIENES P., VÁRADI T. (2001). *New Generation Systran Translation System*. In: MT Summit VIII, 18-22 September 2001, Santiago de Compostela. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.568&rep=rep1&type=pdf.

SERIANNI L. (2010). *Grammatica Italiana. Italiano comune e lingua letteraria*. Novara, Utet Università.

SLOCUM J. (1985). A Survey of Machine Translation: its History, Current Status, and Future Prospects. In: Slocum J. (ed.), *Machine Translation Systems*. Cambridge, University Press, pp. 1-48.

SOBRERO A.A., MIGLIETTA A. (2006). *Introduzione alla Linguistica Italiana*. Roma, Edizioni Laterza.

SOMERS H.L. (2000). Machine Translation. In: Dale R., Moisl H., Somers H.L. *Handbook of Natural Language Processing*. New York, Marcel Dekker, Inc., pp. 329-346.

SQUARTINI M. (1998). *Verbal Periphrases in Romance. Aspect, Actionality and Grammaticalization*. Berlin, De Gruyter.

SQUARTINI M., BERTINETTO P.M. (1995). The Simple and Compound Past in Romance Languages. In: Dahl O. (ed.), *Tense and Aspect in the Languages of Europe*. Berlin, De Gruyter, pp. 403-440.

TEIXEIRA NOGUEIRA M., LOPES VASCONCELOS M.F. (2011). *Modo e Modalidade. Gramática, Discurso e Interação*. Fortaleza, Edições UFC.

TRIPATHI S., SARKHEL J.K. (2010). Approaches to machine translation. In: *Annals of Library and Information Studie*s, December 2010, pp. 388-393.

TUCCI I. (2005). L'espressione della modalità nel parlato: i verbi modali nei corpora italiano e spagnolo C-Oral-Rom. In: *Atti del VIII Convegno Internazionale della SILFI "Lingua, Cultura e Intercultura"*, Korzen. I (ed.). Copenhagen, Samsfundslitteratur Press, pp. 295- 308

VERSTRAETE J.C. (2001). Subjective and objective modality: Interpersonal and ideational functions in the English modal auxiliary system. In: *Journal of Pragmatics*, Vol. 22, Iss. 10, October 2001, pp. 1505-1528.

ZAMAGNI A. (2014). *Italiano Controlado para a Tradução Automática (italiano-português). Linguagem especializada: informática*, Dissertação de Mestrado. Lisboa, Universidade de Lisboa.

SITES *CORPUS*

CONJUGA-ME: http://www.conjuga-me.net.

DIZIONARIO LAREPUBBLICA.IT: http://dizionari.repubblica.it/italiano.php.

ENCICLOPEDIA TRECCANI: http://www.treccani.it/vocabolario/.

INFOPEDIA: https://www.infopedia.pt/dicionarios/portugues-italiano/.

PRIBERAM: http://www.priberam.pt.

SYSTRAN: http://www.systransoft.com.

SYSTRANET: http://www.systranet.com/translate.

WORDNET.PT: http://www.clul.ul.pt/wn/.