

## ESTUDO EXPLORATÓRIO DE INFORMAÇÕES LEXICAIS RELEVANTES PARA A RESOLUÇÃO DE AMBIGUIDADES LEXICAL E ESTRUTURAL

Maria Paula Fiorim PIRUZELLI\*  
Bento Carlos DIAS-DA-SILVA\*\*

*ABSTRACT: Translation is an issue that stirs up discussions in the Linguistic domain. Nevertheless, with the fast technological development in computer sciences, the computer influence encompasses almost every area of human knowledge, which has given rise to many challenges to natural language processing and specifically to machine translation, which has become the target of a great number of research projects, the goals of which is to gather deep understanding of human languages to make machine translation viable and of good quality. In this context, this study discusses one of the most machine translation hard problems: to understand and solve the linguistic ambiguity resolution task. In particular, this task is mapped in the translation context from English into Portuguese and focuses on the lexical ambiguity resolution.*

*KEYWORDS: machine translation; ambiguity resolution; natural language processing.*

### 1. Introdução

Essa pesquisa norteia-se pelos seguintes objetivos, que são divididos em dois domínios complementares, com base na metodologia de estudo do Processamento Automático de Línguas Naturais (PLN), proposta por Dias-da-Silva (1996, 2006): o domínio linguístico e o linguístico computacional. No domínio linguístico, (i) estudam-se os aspectos léxico-gramaticais e semântico-conceituais de pares de frases automaticamente traduzidas do inglês para o português, comparando-as com traduções feitas por tradutores humanos extraídas do *corpus* paralelo português-inglês especificamente selecionado para o projeto, e (ii) catalogam-se, com base na literatura, as principais ambiguidades estudadas. No domínio linguístico-computacional, (iii) estudam-se as principais estratégias computacionais para a resolução dos diferentes tipos de ambiguidade catalogados em (ii).

Para isso, analisam-se ocorrências de pares de frases traduzidas de forma automática do inglês para o português pela Ferramenta de Idiomas do Google<sup>1</sup> (doravante FIG), comparando-as com as traduções feitas por tradutores humanos extraídas de um *corpus* paralelo português-inglês (descrito mais adiante). Essa análise visa detectar e catalogar os principais tipos de ambiguidade estudar as principais heurísticas para resolvê-las.

O *corpus* selecionado para o projeto, o COMPARA<sup>2</sup>, é um *corpus* paralelo bi-direcional, português e inglês, que reúne textos escritos originalmente em português e em inglês, contando com autores como Aluísio Azevedo, Chico Buarque, Edgar Allan Poe, José Saramago, Machado de Assis, Mary Shelley, Oscar Wild, entre outros. Esses textos são armazenados em uma base de dados e alinhados com as respectivas traduções nas duas

\* Aluna de mestrado; UNESP – Universidade Estadual Paulista ‘Júlio de Mesquita Filho’, Campus de Araraquara.

\* Professor Doutor; UNESP – Universidade Estadual Paulista ‘Júlio de Mesquita Filho’, Campus de Araraquara.

<sup>1</sup> Disponível em <http://translate.google.com.br/?hl=pt-BR&tab=wT#>. Acesso em: 12 jul. 2010

<sup>2</sup> Disponível em: <http://www.linguateca.pt/COMPARA/index.php>. Acesso em: 12 jul. 10

línguas e constitui o recurso a partir do qual estudam-se tanto a tradução humana quanto “a feita pela máquina (FRANKENBERG-GARCIA; SANTOS, 2002; 2003).<sup>3</sup>

A escolha da direção da tradução inglês-português foi feita com base nas técnicas de elaboração de dicionários bilíngues como ao Houaiss (2005) e Taylor (2003), que atestam que os dicionários inglês-português devem ser elaborados por falantes nativos do português, enquanto que aqueles, que têm o inglês como língua de chegada, devem ser produzidos por falantes nativos do inglês. Ressalta-se que o uso das versões em inglês dos textos originais em português como dados não compromete a análise, já que o objetivo da pesquisa não é julgar a autenticidade dos textos, mas sim determinar as ambiguidades presentes nas frases do inglês que podem ser causas de má-formação ou inadequação das frases do português que foram produzidas automaticamente pela FIG.

Adota-se a metodologia de pesquisa no âmbito do PLN proposta por Dias-da-Silva (1996; 2006), que defende a divisão do estudo em três domínios complementares de investigação: o linguístico, o linguístico-computacional e o computacional. Dentro do primeiro, explicitam-se os conhecimentos linguísticos necessários para descrever um determinado fenômeno da língua e que serão incorporados em algum tipo de sistema de PLN; no segundo, os conhecimentos descritos no domínio anterior são transformados em representações formais; por fim, no terceiro, as representações propostas no domínio linguístico-computacional são codificadas em uma linguagem de programação, domínio que não será abordado nesta investigação.

Assim, articulando-se nos dois primeiros domínios complementares - o linguístico e o linguístico-computacional-, as investigações catalogam ocorrências para ilustrar os principais tipos de ambiguidade sistematizados a partir do estudo da literatura. No domínio linguístico, descreve-se o conhecimento linguístico necessário para a resolução das ambiguidades; no domínio linguístico-computacional, fundamentando-se na descrição do conhecimento sistematizado no domínio anterior, as principais estratégias computacionais para a implementação de heurísticas de resolução desses tipos de ambiguidade são catalogadas.

### Questões de tradução

Segundo Vilela (1994, p. 13) “traduzir é transpor textos ou enunciados duma língua (= língua de partida) para outra língua (= língua de chegada)”. Por envolver a comparação entre uma ou mais línguas, a tradução sempre foi um tema intrigante para os pesquisadores das línguas naturais e da Linguística. No domínio das Letras, contudo, os estudos da tradução se separaram e sofreram uma especificação. A comparação entre línguas passou a se ocupar da reconstrução dos diferentes estágios diacrônicos das línguas (antigas ou não) e a tradução, por sua vez, concentrou-se na tradução literária, tradução-interpretação ou na tradução assistida por computador.

É possível afirmar que o tema “tradução” pode ser abordado de, pelo menos, dois pontos de vista: do ponto de vista do **tradutor humano** e do ponto de vista da **tradução automática** por sistemas de TA, mas as questões linguísticas são pertinentes a ambos (HATIM, 1990). As questões linguísticas, no âmbito da TA, por sua vez, dividem-se entre aspectos eminentemente linguísticos e aspectos linguístico-computacionais.

A tradução feita ou auxiliada por computadores tem sido discutida há tempos pelos pesquisadores de PLN, porque, envolvendo a comparação entre línguas, implica a modelagem

---

<sup>3</sup> É possível acessar o COMPARA *online* gratuitamente e, atualmente, o *corpus* conta com aproximadamente três milhões de palavras provenientes de textos de ficção. Entretanto, outros gêneros deverão ser acrescentados.

do comportamento linguístico humano, que engloba aspectos cognitivos, históricos e socioculturais, além dos aspectos puramente linguísticos.

No final do século XX, a automatização da tradução se tornou uma realidade, embora não tão bem sucedida, à medida que sistemas de TA passaram a ser desenvolvidos e lançados na Internet. Entretanto, esses sistemas não são, na verdade, capazes de traduzir sozinho textos de uma língua natural para outra, isso porque a tradução de qualidade só é alcançada através da pós-edição de um tradutor humano.

Wilks (2009) argumenta que apenas com os estudos realizados até hoje sobre a TA ainda é cedo para poder afirmar muitos fatos, porém ele aponta as inconsistências no que já é conhecido até agora já que é possível afirmar que, se de um lado a TA funciona, fato comprovado pela existência de sistemas que traduzem de forma completamente automática, trazendo benefícios para muitos usuários que necessitam recorrer a esse recurso, por outro, a afirmação de que é evidente a falta avanços teóricos que possibilitem uma TA de alta qualidade também é válida.

Traduzir é uma tarefa complexa até para tradutores humanos porque é necessário compreender o texto que é naturalmente ambíguo sob vários pontos de vista, além de que os conteúdos veiculados por eles estabelecem relações com conhecimentos exteriores. Além disso, ainda é preciso considerar a necessidade de se ter bastante conhecimento sobre as duas línguas e não deixar de lado suas diferenças e semelhanças (SANTOS, p. 03, 1988).

Entretanto, desde a década de 40, quando os computadores foram apresentados ao mundo ocidental, seu desenvolvimento tem sido constante e sua contribuição para com todos os domínios do conhecimento é evidente. O potencial dessas máquinas para auxiliar a investigação linguística assim como em muitas outras áreas do conhecimento é enorme. Todo o desenvolvimento trazido pelos computadores proporcionou o nascimento de uma grande diversidade de desafios, sempre com o foco no problema de fazer com que a comunicação entre o usuário e a máquina se torne “mais amigável”.

Foram os desafios que surgiram em torno da questão do tratamento computacional das línguas naturais que fizeram com que grandes investimentos materiais e humanos fossem aplicados nesse empreendimento, criando, dessa forma, um domínio de estudos novo: o PLN (DIAS-DA-SILVA, 2006). A TA se encontra inserida nesse contexto, já que ela faz parte de um domínio de estudos multidisciplinar que investiga como desenvolver programas computacionais (os sistemas de TA), que têm, como objetivo, a compreensão da linguagem humana, implicando na construção de interfaces em língua natural que venham a auxiliar os usuários das línguas e dos computadores em diferentes pontos do globo.

### **Questões da TA no âmbito do PLN**

Percebe-se uma grande diversidade de objetivos dentro do PLN por abordar questões linguísticas e também computacionais e para que seus objetivos sejam alcançados, é preciso desenvolver um trabalho que una esses dois grandes campos de conhecimento representados, respectivamente, pela Linguística e pelas Ciências da Computação. Mas, apesar da necessidade desse trabalho conjunto ser um fato bem reconhecido, nos últimos anos ele ainda tem ocorrido muito timidamente.

Além disso, embora haja certo reconhecimento de que a construção de conhecimentos linguísticos e metalinguísticos seja tarefa essencial e indispensável para que uma realização qualitativamente significativa possa ser alcançada no âmbito dos estudos do PLN, os fenômenos linguísticos, por razões diversas, não têm sido descritos com a necessária precisão. Em particular, as pesquisas que se ocupam da TA são, frequentemente, alvos de críticas, que as acusam de não considerar os conhecimentos descobertos e construídos pela Linguística.

Isso tudo corrobora para que, na prática, ainda haja um abismo na comunicação entre o desenvolvimento da Linguística Teórica e do PLN.

Assim como ressalta Santos (1999), o computador, quando utilizado como uma ferramenta, tem a característica de possibilitar que novas formas de descrição e sistematização das nossas próprias capacidades sejam descobertas e também testadas, e isso não só no domínio da tradução, bem como em todas as áreas do conhecimento a que ele seja aplicado (SANTOS, p. 04, 1999).

Os dados que precisam ser analisados quando se pretende desenvolver um estudo sobre línguas naturais são muito numerosos e, por natureza, complexos. O uso de computadores para auxiliar nesse aspecto representa uma ferramenta para ajudar a controlar a quantidade dos dados e até para amenizar a complexidade desses. Entretanto, os métodos utilizados para fazer os computadores lidarem com dados linguísticos, até nos dias de hoje, ainda necessitam de maior desenvolvimento.

Yehoshua Bar-Hillel, ilustre pesquisador do Instituto de Tecnologia de Massachussetes (MIT), quando nomeado pelo instituto e após analisar o assunto da TA, escreveu um artigo mostrando as abordagens básicas para a TA que eram utilizadas no período. Nessa época já era conhecido o fato de que o auxílio humano seria necessário para pré-editar ou pós-editar os textos, porque uma TA completamente automatizada e de alta qualidade seria impossível. Bar-Hillel criticava a noção de que o objetivo das pesquisas sobre TA deveria ser criar sistemas completamente automatizados e que produzissem traduções iguais às produzidas por seres humanos (HUTCHINS, 2001).

Nirenburg (1996) enfatiza que Bar-Hillel acreditava que a modelagem do conhecimento de mundo de forma que ele pudesse ser acessado pelas máquinas era uma condição essencial para o sucesso da TA.

*It seems now quite certain ... that with all the progress made in hardware, programming techniques and linguistic insight, the quality of fully autonomous mechanical translation, even when restricted to scientific or technological material, will never approach that of qualified human translators and that therefore MT will only under very exceptional circumstances be able to compete with human translation. [...] Expert human translators use their background knowledge, mostly subconsciously, in order to resolve syntactical and semantical ambiguities which machines will have either to leave unresolved or resolve by some "mechanical" rule which will every so often result in a wrong translation (NIRENBURG, 1996).*

### **Questões de tipologia dos sistemas de TA**

No que diz respeito à metodologia empregada, em linhas gerais, Hutchins (1992) classifica os sistemas de TA em bilíngues, quando trabalham com apenas um par de línguas, ou multilíngues, quando se ocupam de mais de duas línguas. Caracterizam-se ainda em sistemas unidirecionais, quando realizam a tradução em uma direção apenas, inglês-português, por exemplo, ou bidirecionais, quando traduzem nas duas direções, inglês-português ou português-inglês. O grau de sofisticação dos sistemas é medido de acordo com um ou mais dos três tipos de metodologia empregados no processo de tradução. A partir disso, são classificados basicamente em três tipos: os sistemas diretos, os sistemas de interlíngua e os sistemas de transferência (HUTCHINS, 2003).

Santos (1988) atesta que uma das possíveis distinções a ser feita em relação aos sistemas de TA é entre os chamados sistemas diretos, que traduzem diretamente a partir da língua de origem, e aqueles que são indiretos, utilizando alguma forma intermediária para

representar os conhecimentos e a estrutura da língua de origem e só depois geram o texto de chegada. São incluídos nesse tipo os sistemas de transferência e os baseados em interlíngua.

Os sistemas diretos, que são os mais simples, são bilíngues e unidirecionais. De acordo com essa abordagem, o texto de partida é analisado minimamente para poder originar textos na outra língua. Esses sistemas realizam a tradução procurando os correspondentes diretos entre os itens lexicais das línguas fonte e alvo.

Nos sistemas de transferência, a tradução processa-se por meio de regras sintáticas, a partir da análise da estrutura sintática da frase da língua fonte, gera-se uma representação sintática para a língua alvo, e se dá em três estágios: durante o primeiro estágio, o texto de partida é transformado em representações intermediárias, elimina-se, assim, a ambiguidade; em seguida, durante o segundo estágio, essas representações são transformadas em representações equivalentes para a língua de chegada; finalmente, na última etapa do processo, um texto na língua de chegada é gerado.

Os sistemas de interlíngua são os mais sofisticados e funcionam de maneira diferente: neles, a tradução é feita a partir do texto de partida para uma interlíngua, uma representação abstrata do significado que se aplica a qualquer língua, e desta para a língua de alvo (DIAS-DAS-SILVA, 2006; HUTCHINS, 1992). A tradução é feita, portanto, baseada na possibilidade de transformar textos em conceitos que podem ser representados em qualquer língua.

Santos (1988) menciona ainda uma divisão que atualmente não é mais tão significativa, estabelecida entre sistemas predominantemente sintáticos e outros predominantemente semânticos. Como já é um fato amplamente reconhecido que é necessário compreender o texto para poder traduzir, essa distinção acaba por deslocar-se “para uma opção metodológica de compromisso entre eficiência e qualidade” (SANTOS, 1988, p. 08). A autora propõe que, nesse aspecto, seria mais interessante classificar os sistemas em relação ao grau de conhecimento sintático, semântico, pragmático, etc., exigido por eles, enquanto que Hutchins (1986) defende que a distinção inicial poderia ser

“expressa em termos de quem controla quem, ou seja, um sistema de índole sintática manipularia marcadores semânticos ajudando a identificação das estruturas, mas possuiria como unidade básica, por exemplo, a frase (conceito eminentemente sintático). Por outro lado, um sistema de índole semântica poderia executar simultaneamente com o varrimento sintático do texto uma análise semântica, ou ter como única representação interna, a partir do texto, uma representação de casos ou na forma da dependência conceitual.” (SANTOS, 1988, p. 08).

Uma outra distinção é ainda comentada por Santos (1988), aquela que se estabelece entre sistemas de TA inspirados pela Inteligência Artificial e aqueles que se baseiam em teorias linguísticas. Mas, neste caso, e como também já foi argumentado anteriormente, a autora defende que a união entre abordagens de ambos os domínios é o caminho mais vantajoso.

Na sequência, este estudo abordará, na seção 2, as ambiguidades linguísticas, nomeando os principais tipos apontados na literatura e dando destaque à parte de dois dos cinco tipos: a lexical e a estrutural, que serão exemplificados na seção 4. Na seção 3, resumem-se as principais estratégias de resolução de ambiguidade desses dois tipos, mostrando a importância da construção de léxicos computacionais contendo informações robustas, cobrindo os domínios morfológico, sintático e semântico-conceitual. A seção 5 apresenta as considerações finais

## 2. As ambiguidades linguísticas

Diz-se que um item/expressão lexical é ambíguo quando apresenta mais de um sentido possível diferente. Porém, o termo restringiu-se para a nomeação daqueles itens/expressões lexicais que têm mais de um sentido estabelecido, isso se deve ao fato de todos os itens lexicais poderem ser ambíguos dentro de um ou outro contexto (CRUSE, 2006). No uso real da língua, o contexto sempre determina qual das alternativas possíveis de uma leitura ambígua é a adequada. Por isso, a ambiguidade não causa grandes dificuldades de interpretação.

Ide & Véronis (1998) apontam que as consequências problemáticas da manifestação das ambiguidades linguísticas era o ponto central já no artigo escrito por Bar-Hillel em 1960. Tomando como exemplo o pequeno texto: *Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy* (NERENBURG, 1996, p. 302), Bar-Hillel argumentou a impossibilidade de se determinar o sentido adequado do item lexical *pen* na frase *The box is in the pen* automaticamente, reconhecendo que para os humanos, o conhecimento de mundo sobre *pens* (canetas) e *playpens* (cercado dentro do qual crianças pequenas brincam) é o que os faz compreender o sentido tão facilmente. Bar-Hillel ainda acrescenta que se fosse possível desenvolver um sistema que tivesse acesso a esse tipo conhecimento, seria o mesmo que afirmar que os sistemas de TA deveriam ser equipados com, além de um dicionário, uma enciclopédia de conhecimentos universais.

O fenômeno da ambiguidade pode se manifestar em diferentes níveis de análise linguística: lexical, sintático, semântico, contextual-pragmático (DIAS-DA-SILVA, 1996; HIRST, 1992; SPECIA, 2007). Em particular, no nível lexical, por exemplo, um item léxico da língua fonte pode codificar mais de um sentido quando traduzido para a língua alvo e como consequência disso, a língua alvo oferece mais de uma opção disponível para a tradução. A divergência entre as várias culturas existentes é citada como um dos fatores que pode explicar essas diferenças, porque a cultura influencia a forma como os conceitos do mundo se realizam nos diferentes itens léxicos. Já no nível sintático, é possível organizar os itens lexicais que compõem uma frase em diferentes sequências, originando trechos ou frases inteiras ambíguos.

### Tipologia das ambiguidades

Em linhas gerais, a literatura aponta quatro grandes tipos de ambiguidades linguísticas:

1. **Ambiguidade lexical**, que se subdivide nos tipos 1.1 Ambiguidade por polissemia/homonímia, 1.2 Ambiguidade categorial e 1.3 Ambiguidade de transferência;
2. **Ambiguidade estrutural**, que se subdivide nos tipos 2.1 Ambiguidade de fixação de constituinte, 2.2 Ambiguidade de localização e de preenchimento de lacunas, 2.3 Ambiguidade analítica, 2.4 Ambiguidade de escopo da quantificação;
3. **Ambiguidade anafórica/referencial**;
4. **Ambiguidade temática**.

As ambiguidades de fixação de constituinte, por sua vez, subdividem-se em sete subtipos:

- 2.1.1 Ligação de um sintagma preposicional a mais de um sintagma nominal ou verbal;
- 2.1.2 Ligação de uma oração relativa a mais de um sintagma nominal disponível;
- 2.1.3 Ligação de um sintagma preposicional a uma oração adjetiva;

- 2.1.4 Possibilidade de ligação de um sintagma preposicional ou advérbio a posições pertencentes à oração ou à sua sub-oração;
- 2.1.5 Ligação do advérbio como modificador do sintagma verbal ou da frase;
- 2.1.6 Ligação de participípios ao sujeito estrutural da frase ou à frase;
- 2.1.7 Possibilidade de ligação simultânea de um advérbio a verbos de duas frases distintas.

Também apresentam subtipos as ambiguidades analíticas, que se subdividem-se em onze subtipos:

- 2.3.1 Detecção de partículas;
- 2.3.2 Diferenciação entre um sintagma preposicional e um sintagma adjetivo resultante de uma operação de alçamento e apagamento do verbo *ser/estar* aplicada ao complemento do verbo;
- 2.3.3 Diferenciação entre participípio presente e adjetivo;
- 2.3.4 Diferenciação entre participípio presente e substantivo;
- 2.3.5 Delimitação da extensão do sintagma nominal;
- 2.3.6 Diferenciação entre oração relativa reduzida e sintagma verbal da oração principal;
- 2.3.7 Delimitação da estrutura de um sintagma nominal complexo;
- 2.3.8 Interpretação ambígua de participípios e de orações adjetivas posicionados no final de frase;
- 2.3.9 Diferenciação entre frases clivadas e frases do tipo sujeito-verbo-objeto;
- 2.3.10 Diferenciação entre participípio passado e um sintagma verbal incompleto, resultando na ambiguidade entre pergunta e ordem;
- 2.3.11 Delimitação dos diferentes tipos de estrutura formados com esta sequência de elementos: *NP be ADJ to V*.

Contudo, neste trabalho, por razões da extensão da discussão, restringe-se a discussão a dois tipos de Ambiguidade Lexical (CRUSE, 2006; HIRST, 1992; HUTCHINS, 1992; SOMERS, 2000; SPECIA, 2007), a 1.1 Ambiguidade por polissemia/homonímia e a 1.2 Ambiguidade categorial, e a um tipo de Ambiguidade Estrutural, a 2.1.1 Ligação de um sintagma preposicional a mais de um sintagma nominal ou verbal.

### As ambiguidades lexicais

As ambiguidades que se manifestam no nível lexical sempre exigem que uma escolha seja feita entre as possíveis leituras, porque a escolha não adequada do item léxico resulta em proposições diferentes. Essa situação ilustra-se com a frase *But it's conditioning, brain-washing: more like a trained seal*, em que o sentido de *seal* deve ser desambiguado de forma adequada entre *selo*, *escudo*, *lacre* ou *foca* (SOMERS, 2000, p. 333; SPECIA, 2007, p. 12).

Embora não seja possível traçar uma fronteira rígida entre polissemia e homonímia, aceita-se que os itens léxicos polissêmicos são aqueles que os seus possíveis sentidos demonstram relações entre si. Para que os sentidos possam ser considerados como pertencentes ao mesmo item lexical, os falantes da língua precisam senti-los como relacionados. Algumas das relações responsáveis pela polissemia são a metáfora, a metonímia e a hiponímia. Já os itens lexicais homônimos apresentam sentidos que não permitem estabelecer nenhum tipo de relação entre si. Cruse (2006) aponta que a maior parte dos dicionários tradicionais confere entradas distintas para os homônimos, diferentemente do que

ocorre com os itens lexicais polissêmicos, que são identificados por números dentro da mesma entrada.

Nota-se que a distinção entre a polissemia e a homonímia é subjetiva e, se, em alguns casos, a distinção é bastante definida, em outros é impossível de se estabelecer uma diferença.

Almeida (1990) argumenta que, apesar da polissemia e homonímia demonstrarem diferenças em suas origens, ambos os fenômenos contribuem da mesma forma para a ambiguidade estrutural. Segundo o autor, o que realmente interessa são os múltiplos sentidos relacionados com uma única forma. Assim sendo, para o tratamento computacional essa divisão não é relevante, é suficiente a existência de algum tipo de biunivocidade entre forma e sentido.

Os itens léxicos categorialmente ambíguos são aqueles que podem pertencer a categorias sintáticas diferentes, variando de acordo com o contexto como, por exemplo, *canto*, que, além de ser alvo da homonímia, é também alvo da ambiguidade categorial, porque pode ser um substantivo ou um verbo na primeira pessoa do presente do indicativo. Na maioria das vezes, esse tipo de ambiguidade é solucionado pelo *parser* (analisador gramatical), não representando entraves mais sérios à TA (HUTCHINS, 1992; SPECIA, 2007).

De acordo com Hutchins (1992), as ambiguidades por polissemia/homonímia e categorial são monolíngues porque causam problemas para a análise da língua fonte. As ambiguidades de transferência são, por sua vez, ambiguidades bilíngues e se manifestam quando um item lexical da língua fonte pode ser traduzido por vários itens/expressões da língua alvo. Dessa forma, o problema só se manifesta sob a perspectiva da língua alvo, porque, para um falante nativo da língua fonte, o item lexical não é percebido como ambíguo. O item lexical do inglês *wall* ilustra esse tipo, porque, ao ser traduzido para o português, exige a escolha entre os itens *parede*, que denota paredes internas a uma construção, e *muro*, que são “paredes ao ar livre”.

### As ambiguidades estruturais

No nível estrutural (sintático), as diferentes formas possíveis de se agruparem sequências de itens lexicais podem ser a causa de trechos ambíguos ou até mesmo de frases inteiras ambíguas. A combinação de ambiguidades lexicais dos itens léxicos que compõem a frase é, muito frequentemente, apontada como uma das causas das ambiguidades estruturais. Considere, por exemplo, a sequência *I saw the man in the house with a telescope*. Como mostra Allen (1995), é possível, para um leitor humano, encontrar, pelo menos, cinco interpretações diferentes, devido a diferentes possibilidades de se diferentes interpretar o sintagma preposicional *with a telescope*.

Porém, apesar de grande parte das frases permitirem diversas análises gramaticais, de acordo com Hirst (1992, p. 09), após considerar aspectos semânticos e contextuais, apenas uma interpretação possível permanece. Considerando, por exemplo, a frase *Nadia left the university on the wrong bus*, para compreendê-la adequadamente é necessário aplicar o conhecimento de mundo de que universidades não andam de ônibus, e esse conhecimento o autor chama de “viés semântico”. Além desse viés, as línguas também exibem certas preferências sintáticas que Hirst (1992) denomina de “viés sintático”. Na frase *The landlord painted all the walls with crack*, o sintagma preposicional *with crack* pode ser fixado ao sintagma verbal, podendo ser interpretado como “as paredes estavam sendo pintadas em um estilo ‘rachaduras’” ou “as rachaduras foram usadas como instrumento para pintar as paredes”, interpretações que são semanticamente anômalas, e também pode se ligar ao sintagma nominal objeto sendo que, nesse caso, a interpretação seria “as paredes que apresentavam rachaduras foram pintadas”.

Muitas pesquisas foram feitas sobre esse viés sintático e também sobre como os humanos decidem sobre qual o local adequado para fixar um novo constituinte durante uma análise sintática. Como resultado desses estudos, alguns princípios gerais puderam ser afirmados, são eles: *Minimal Attachment* e *Right Association* ou *Late Closure* (ALLEN, 1995; HIRST, 1992).

*Minimal Attachment* é o princípio mais geral que afirma a existência de uma preferência para a estruturação sintática que cria o menor número possível de nós na árvore sintática. A frase *The man kept the dog in the house* (ALLEN, 1995, p. 160) exemplifica o princípio. Normalmente, essa frase é interpretada com o sintagma preposicional *in the house* modificando o verbo *kept*, o que, conseqüentemente, produz uma árvore sintática com um número menor de nós.

O princípio *right association* formula que um novo constituinte deve ser interpretado como parte do constituinte que está sendo construído e não deve ser fixado em nenhum outro constituinte pertencente a um nível superior na hierarquia da árvore sintática. Esse princípio é ilustrado por Allen (1995) com a frase *George said that Henry left in his car*, que pode ter duas interpretações sintaticamente aceitáveis – George falou que Henry saiu utilizando seu próprio carro e George falou, dentro do carro, que Henry saiu – sendo que a interpretação preferida é primeira. Essa interpretação preferida tem o sintagma preposicional fixado ao sintagma verbal que lhe é imediatamente anterior. A outra interpretação, por sua vez, faz a fixação do sintagma preposicional *in the car* ao sintagma verbal mais alto na árvore sintática.

Entretanto, na frase *The man kept the dog in the house*, esses dois princípios são conflitantes, porque o princípio *right association* aparentemente sugere que o sintagma preposicional seja fixado ao sintagma nominal *the dog*, já o princípio *minimal attachment* favorece a fixação do sintagma preposicional junto ao sintagma verbal *kept*. Como consequência disso, Allen (1995) afirma que haverá situações em que as preferências lexicais serão desejáveis em detrimento das preferências baseadas nesses princípios.

Quando um verbo exibe uma subcategorização que exige um sintagma preposicional, como o verbo *put*, por exemplo, que subcategoriza um sintagma preposicional com *in*, *on* ou *by*, então, nas situações que em houver necessidade de escolher um local para a fixação de um sintagma preposicional, o sintagma verbal terá preferência. Em outros casos, é o próprio sintagma preposicional que demonstra preferência por ser fixado dentro do sintagma verbal. Se essas duas situações não se manifestarem, então os princípios gerais devem ser obedecidos.

Contudo, em oposição às ambiguidades lexicais que sempre exigem que uma escolha seja feita, em determinadas situações a ambiguidade estrutural não precisa ser resolvida na língua alvo, por não apresentar dificuldade de compreensão para os falantes nativos. Essa situação pode ser ilustrada com a frase *The man saw the girl with a telescope*, na qual não se sabe se quem tinha o telescópio era o homem ou a menina. Essa ambiguidade pode ser transferida para a língua de chegada sem ser resolvida, porque se trata de fenômeno análogo na língua de chegada, e que só é solucionado com o conhecimento do contexto de uso desse tipo de frase.

A partir dessas considerações gerais sobre os tipos de ambiguidades alvos deste estudo, fica claro que as informações lexicais ajudam na resolução tanto das ambiguidades que se manifestam no nível lexical, quanto daquelas aparentes no nível estrutural, daí a importância de se equipar um sistema de TA com um léxico que contenha informações relevantes necessárias para o processo de desambiguação como, por exemplo, a categoria sintática, a afinidade de um verbo com uma determinada preposição, relações de sentido (sinonímia, hiperonímia, etc.), restrições sintáticas e semânticas, entre outras.

### 3. Em busca de soluções...

Hirst (1992) reforça a necessidade de modelagem computacional tanto do contexto discursivo quanto do co-texto em que o item léxico problemático ocorre para realizar a desambiguação. Essa necessidade decorre do fato de os itens léxicos próximos ao item léxico ambíguo poderem fornecer pistas para o sistema, ou seja, fornecerem indícios fortes para a desambiguação se um dos sentidos possíveis de um item léxico estiver semanticamente relacionado ao sentido de um outro próximo a ele. Daí, Hirst (1992, p.80) propor que os mecanismos necessários para a desambiguação no nível lexical são:

- (i) Reconhecimento do contexto;
- (ii) Associações semânticas entre itens lexicais;
- (iii) Informação sobre a sintaxe;
- (iv) Informação sobre restrições seletivas dos itens léxicos ambíguos
- (v) Inferências.

De acordo com Wilks (2009), as fontes de conhecimento necessárias para o funcionamento de um sistema de TA dependem do método por ele utilizado. Mas, é possível afirmar que a maioria dos sistemas utiliza, comumente, algumas das seguintes fontes de conhecimento: informações morfológicas, regras gramaticais e informações provenientes de léxicos. No caso do inglês, por ser uma língua que não apresenta muita flexão, a morfologia não é tão necessária como seria para uma língua muito flexionada, para as quais a informação morfológica é muito importante.

Por causa do aumento do número de dados legíveis por máquina disponíveis nos últimos anos e das técnicas estatísticas que podem ser aplicadas para identificar e utilizar as informações retiradas desses dados, as tentativas de desambiguar sentidos lexicais de forma automática cresceram.

De acordo com Stevenson e Wilks (2003), a tarefa de desambiguação lexical de sentido (DLS) é tema de interesse dos pesquisadores desde o começo dos estudos sobre a TA e é sempre reconhecida como um dos problemas mais importantes que carecem de solução dentro do campo de pesquisa do PLN. A DLS é uma tarefa intermediária (STEVENSON; WILKS, 2003) porque ela é necessária, ou pelo menos traz benefícios, para o desempenho de muitas outras tarefas de PLN, a TA é uma delas e para citar outras, tem-se, por exemplo, a recuperação de informação, análise gramatical e processamento de fala. Aliás, como se vê afirmado em Ide & Véronis (1998), os primeiros trabalhos sobre DLS foram desenvolvidos dentro do contexto da TA. Os mesmos autores apontam que a tarefa de DLS é descrita como “*AI-complete*”, o que significa dizer que é um problema que poderá ser solucionado apenas quando todos os outros problemas da Inteligência Artificial também tiverem alcançado uma solução.

Grosso modo, a tarefa da DLS é associar uma determinada unidade lexical de um texto com uma definição (ou seja, o sentido) dentre várias que podem ser potencialmente atribuídas a ela (IDE & VÉRONIS, 1998; SPECIA, 2007). A tarefa requer duas etapas: a primeira é a determinação de todos os sentidos diferentes relevantes para cada unidade léxica do texto e a segunda é a escolha de um meio de atribuir um sentido apropriado a cada ocorrência da unidade léxica. Para realizar a primeira etapa, geralmente conta-se com acervos de sentidos já definidos, assim como os sentidos registrados em um dicionário ou as informações retiradas de um *thesaurus*. A segunda etapa é realizada com base em informações provenientes do contexto do item léxico ambíguo e de outras fontes de conhecimento, como recursos lexicais e

enciclopédicos, e também de fontes de conhecimento manualmente construídas (IDE & VÉRONIS, 1998, p. 03).

Kilgarriff (1997, p. 212) argumenta que as informações lexicais podem resolver até mesmo grande parte das ambiguidades estruturais, sem que os sentidos dos itens lexicais precisem ser desambiguados. Para exemplificar sua argumentação, o autor considera duas frases-exemplo:

(i) *I love baking cakes with friends.*

(ii) *I love baking cakes with butter icing.*

Para resolver a ambiguidade de fixação do sintagma preposicional (*with...*), basta considerar a semântica do substantivo núcleo do sintagma nominal final (*friends* ou *butter icing*). Em (i) o núcleo do sintagma nominal é humano e por isso o sintagma preposicional deve ser fixado ao verbo (*baking*); em (ii) o núcleo do sintagma nominal é um tipo de ingrediente do bolo (a cobertura) e, conseqüentemente, fixa-se ao substantivo (*cakes*). Nesse caso, nem *friends* nem *icing* é ambíguo entre humano e ingrediente de bolo. Por essa razão, a DLS não é necessária.

Os estudiosos das ambiguidades linguísticas, incluindo aí a DLS, sempre enfatizam a importância do contexto. Ide & Véronis (1998) o apontam, inclusive, como a única fonte capaz de identificar sozinha o sentido adequado de um item lexical ambíguo. É por essa razão que todos os trabalhos em desambiguação de sentido utilizam informações provenientes do contexto do item lexical alvo da ambiguidade, contexto esse que, de alguma forma precisa ser modelado.

#### 4. Desenvolvimentos empíricos...

Apresentam-se, nesta seção, frases selecionadas no *corpus* descrito no início deste estudo, que exemplificam as ambiguidades lexicais e os problemas que elas representam para o processo de TA. Destaca-se o item lexical ambíguo em negrito e, abaixo da frase original, apresentam-se duas traduções: a do tradutor humano e a do sistema de TA.

Exemplo 1: *Shall I **ring** Phyllis Cameron and ask her?*

TH: **Telefone** à Phyllis Cameron para lhe perguntar?

TA: Devo **anel** Phyllis Cameron e perguntar-lhe?

Nesse exemplo, está ilustrado um caso de ambiguidade categorial relacionado às leituras nominal e verbal do item lexical *ring*. Nota-se que, nesse exemplo, a presença do pronome *I* deveria ser evidência suficiente para indicar a necessidade de um verbo em seguida, fato que impossibilitaria o emprego do substantivo. Portanto, conclui-se que informações sobre a categoria gramatical e suas restrições sintáticas devem fazer parte do léxico do sistema para que o analisador gramatical possa trabalhar corretamente.

No Exemplo 2, a seguir, a ambiguidade categorial é ilustrada pelo item lexical *steps*, que pode ser um substantivo no plural, correspondendo, em português, ao item lexical *etapas* ou *degraus*, ou a terceira pessoa do singular do presente simples do verbo *to step*, que em português corresponderia a *pisa*.

Exemplo 2: *Leslie **steps** forward with a smile, introduces himself to the couple, and inspects their tickets and passports.*

TH: Leslie **avança** com um sorriso nos lábios, apresenta-se ao casal e verifica os respectivos bilhetes e passaportes.

TA: Leslie **passos** para a frente com um sorriso, se apresenta ao casal, e inspeciona os ingressos e passaportes.

Um sistema de TA, ao processar o item *steps*, precisará, portanto, selecionar uma única opção dentre as disponíveis. A seleção depende das seguintes informações: da escolha da categoria gramatical, dos traços semânticos, das relações item-contexto, das relações de sentido, entre outras. No exemplo, essa escolha pode ser assim resolvida: registrando-se, para *step*, *verbo* + *toward*, a sinonímia entre *stairway*, *stairs*, *steps* e a restrição semântica [+lugar].

O Exemplo 3 ilustra a situação de ambiguidade em que o item lexical *paper* é a causa da ambiguidade lexical, podendo ser traduzido para o português como *papel*, *jornal*, *artigo*, entre outros. No primeiro exemplo, o sentido da expressão *newspaper boy*, que foi traduzida adequadamente pelo sistema de TA, fornece evidências para a resolução da ambiguidade de *paper*, apontando para o sentido expresso em português por *jornal*.

Exemplo 3: *The newspaper boy is late, or perhaps there is no **paper** today because of a strike.*

TH: O rapaz dos jornais está atrasado, ou talvez hoje não haja **jornais** por causa de uma greve qualquer.

TA: O menino do jornal é tarde, ou talvez não há **papel** hoje por causa de uma greve.

Já neste Exemplo 4, tratar os itens lexicais *morning paper* como uma *collocation* e incluir essa informação no léxico, resolveria a ambiguidade.

Exemplo 4: *A stewardess offers him the morning **paper**.*

TH: A hospedeira oferece-lhe o **jornal** da manhã.

TA: A aeromoça oferece-lhe o **papel** de manhã.

Considere, por fim, o Exemplo 5, do subtipo 2.1.1 Ligação de um sintagma preposicional a mais de um sintagma nominal ou verbal.

Exemplo 5: *The lecture theatre resonates like a drum **with the chatter** of a hundred-odd students.*

TH: O anfiteatro ressoa como um tambor **com o tagarelar** de uma centena de alunos.

TA: A palestra teatro ressoa como um tambor **com a vibração** de um cem alunos estranho.

Para solucionar esse tipo de ambiguidade, é preciso estar registrada, no léxico do sistema, a informação sobre a afinidade do verbo com a preposição em questão. Se o verbo apresentar essa característica, o sintagma preposicional deve ser fixado ao sintagma verbal; caso contrário, é ao sintagma nominal disponível que o sintagma preposicional deve ser fixado. No Exemplo 5, o verbo *resonate* admite a preposição *with*, informando ao sistema que o sintagma preposicional deve ser nele fixado.

## 5. Considerações finais

As breves discussões feitas neste estudo mostram que é indiscutível que as ambiguidades linguísticas representem um desafio para os sistemas de TA. Desenvolver um estudo sistemático dos tipos de ambiguidade e das suas manifestações no processo de

tradução automática do inglês para o português é o objetivo central que está sendo alvo dos estudos do mestrado em desenvolvimento. Merece destaque a importância de se representarem, nos léxicos dos sistemas de TA, informações sobre categorias, de subcategorização, sobre restrições seletivas, temáticas, sobre relações de sentido, colocações, traços e restrições semânticas, posto que, recordando Kilgarriff (1997), que argumenta que um léxico rico em informações lexicogramaticais e semântico-conceituais são essenciais para resolver grande parte das ambiguidades estruturais, sem que os sentidos dos itens lexicais precisem ser desambiguados.

## Referências

- ALLEN, J. *Natural Language Understanding*. Redwood City, CA: Benjamin/Cummings, 1995.
- ALMEIDA, J. Ambiguidade lexical. *Revista Alfa*, São Paulo. Vol 34, p. 187-193, 1990.
- CRUSE, A. *A Glossary of Semantics and Pragmatics*. Edinburgh University Press, 2006.
- DIAS-DA-SILVA, B.C. O estudo linguístico-computacional da linguagem. *Letras de hoje*, Porto Alegre, v. 41, p. 103-138, 2006.
- \_\_\_\_\_. *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais*. Araraquara, 1996. 272 f. Tese (Doutorado em Letras) - Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara, 1996.
- FRANKENBERG-GARCIA, A.; SANTOS, D. Introducing COMPARA, the Portuguese-English parallel translation corpus. In: ZANETTIN, F.; BERNARDINI S.; STEWART, D. (Eds.). *Corpora in Translation Education*. Manchester: St. Jerome Publishing, 2003. p. 71-87.
- \_\_\_\_\_. COMPARA, um corpus paralelo de português e de inglês na Web. *Cadernos de Tradução IX*, Santa Catarina, p. 61-79, 2002.
- HATIM, B., MASON, I. *Discourse and the translator*. New York: Longman Inc., 1990.
- HIRST, G. *Semantic interpretation and the resolution of ambiguity*. Cambridge: Cambridge University Press, 1992.
- HOUAISS, A. *Webster's: dicionário inglês-português*. 15. ed. Rio de Janeiro: Record, 2005.
- HUTCHINS, W.J. Machine translation: general overview. In: MITKOV, R. (Ed.). *The Oxford handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003. p. 501-511.
- \_\_\_\_\_. Machine Translation over fifty years. *Histoire, Epistemologie, Langage*, v. 22, n. 1, p. 7-31. 2001. Disponível em: <<http://www.hutchinsweb.me.uk/history.htm>>. Acesso em: 28 jul. 2010.
- \_\_\_\_\_. SOMERS, H. L. *An introduction to machine translation*. London: Academic Press, 1992.
- \_\_\_\_\_. *Machine translation: past, presence, future*. Ellis Horwood/Wiley, Chichester/New York, 1986.
- IDE, N.; VÉRONIS, J. Introduction to the Special issue on word sense disambiguation: The State of the Art. *Computational Linguistics*. Cambridge, v. 24, p. 2 – 40, Mar. 1998.
- Disponível em:  
<<http://portal.acm.org/citation.cfm?id=972749.972751&coll=GUIDE&dl=EF%BF%BD%C3%9C&idx=J25&part=affil&WantType=Affils&title=Computational%20Linguistics&CFID=96305886&CFTOKEN=64368211>>. Acesso em: 28 jul. 2010.
- KILGARRIFF, A. What is word sense disambiguation good for?. In: *Natural Language Processing in the Pacific Rim, 1997, Phuket, Thailand. Proceedings...*, Phuket, Thailand, 1997. p. 209-214. Disponível em: <<http://www.kilgarriff.co.uk/publications.htm>>. Acesso em: 23 jul. 2010.

- NIRENBURG, S. Bar Hillel and machine translation: then and now. In: BISFAI'95 The Fourth Bar-Ilan Symposium on Foundations of Artificial Intelligence, 4<sup>th</sup>., 1995, Jerusalem, Israel. *Proceedings...*, Jerusalem, Israel: AAAI Press, 1996. p.300-305. Disponível em: <<http://www.aaai.org/Papers/BISFAI/1995/BISFAI95-027.pdf>>. Acesso em: 16 jul. 2010.
- SANTOS, D. O computador e a tradução. In: II Seminário de Tradução Científica e Técnica em Língua Portuguesa, 2., 1999, Lisboa. *Actas do II Seminário de Tradução Científica e Técnica em Língua Portuguesa*. Lisboa, 1999. Disponível em: <<http://www.linguateca.pt/Diana/download/SantosSeminTradTecnica99.pdf>>. Acesso em: 15 jul. 2010.
- \_\_\_\_\_. *A fase de transferência de um sistema de tradução automática do inglês para o português*, 1988. 252 f. Dissertação (Mestrado em Engenharia Eletrotécnica e de Computadores) - Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, 1988. Disponível em: <<http://www.linguateca.pt/Diana/public.html>>. Acesso em: 15 jul. 2010.
- SOMERS, H. Machine translation. In: DALE, R.; MOISL, H.; SOMERS, H. *Handbook of natural language processing*. New York: Marcel Dekker, 2000. p. 329-346.
- SPECIA, L. *Uma abordagem híbrida relacional para a desambiguação lexical de sentido na tradução automática*. São Carlos, 2007. 245 f. Tese (Doutorado em Ciências) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2007.
- STEVENSON, M.; WILKS, Y. Word-sense Disambiguation. In: MITKOV, R. (Ed.). *The handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003. p. 249-265.
- TAYLOR, J. L. *Webster's: Portuguese-English dictionary*. 16. ed. Rio de Janeiro: Record, 2003.
- VILELA, M. *Tradução e Análise Contrastiva: Teoria e Aplicação*. Lisboa: Caminho, 1994.
- WILKS, Y. *Machine Translation: Its Scope and Limits*. Springer, New York, 2009.