

**Centro Universitário de Araraquara – UNIARA**

Departamento de Ciência da Administração e Tecnologia  
Ciência da Computação – com ênfase em Análise de Sistema

***Abordagem Automática para Criação de  
Cópus Etiquetados com Sentidos para  
Desambiguação Lexical de Sentido na  
Tradução Inglês – Português***

**SYLLAS FREITAS DE OLIVEIRA NETO**

Araraquara  
**2004**

# **Centro Universitário de Araraquara – UNIARA**

Departamento de Ciência da Administração e Tecnologia  
Ciência da Computação – com ênfase em Análise de Sistema

## ***Abordagem Automática para Criação de Cópus Etiquetados com Sentidos para Desambiguação Lexical de Sentido na Tradução Inglês – Português***

**SYLLAS FREITAS DE OLIVEIRA NETO**

ORIENTADORA: PROF<sup>a</sup>. LUCIA SPECIA

Monografia apresentada ao Departamento de Ciência da  
Administração e Tecnologia do Centro Universitário de  
Araraquara, como parte dos requisitos para obtenção do título de  
bacharel em Ciência da Computação com ênfase em Análise de  
Sistema.

Araraquara  
2004

# Agradecimentos

A Deus pela capacitação em concluir esse trabalho.

À minha esposa, Andreza, que em todos os momentos foi paciente em suportar minha ausência para que eu pudesse me empenhar na decorrência desse trabalho e me auxiliando nos momentos difíceis que nos acometeram.

À professora Lucia, por todo auxílio técnico, empenho e motivação.

A todos que de alguma forma me auxiliaram, motivaram e torceram para que esse trabalho chegasse ao fim obtendo bons resultados.

# Resumo

A necessidade em se obter traduções de textos de um idioma para outro de maneira rápida e satisfatória vem impulsionando, há algumas décadas, várias pesquisas e desenvolvimentos nessa área. Como resultado, dispõem-se hoje de muitos sistemas de tradução automática. Contudo, esses sistemas apresentam uma deficiência grave e facilmente perceptível em suas traduções, que é a falta de tratamento ao problema da ambigüidade lexical de sentido das palavras. O objetivo deste trabalho é fornecer subsídios, em termos de pré-processamento, para o desenvolvimento de módulos de desambiguação lexical de sentido para sistemas de tradução automática inglês-português. Para tanto, foram investigadas, propostas e implementadas estratégias para automatizar a criação de córpus de exemplos a serem utilizados em abordagens empíricas para o desenvolvimento de módulos dessa natureza. Além da tradução das palavras, as estratégias foram desenvolvidas de modo que os córpus resultantes apresentem outras informações úteis para a tarefa de desambiguação. Tais estratégias foram implementadas em dois sistemas e então avaliadas. Os resultados se mostraram promissores em termos tanto de abrangência quanto de corretude. Os sistemas implementados podem ser facilmente estendidos e utilizados de modo a facilitar e viabilizar as etapas de pré-processamento para a geração de modelos de desambiguação lexical de sentido.

# Sumário

<b>Capítulo 1 - Introdução .....</b>	<b>1</b>
1.1 Contextualização e motivação .....	1
1.2 Objetivos .....	5
1.3 Metodologia.....	5
1.4 Resultados .....	6
1.5 Organização .....	6
<b>Capítulo 2 - Ambigüidade Lexical de Sentido na TA.....</b>	<b>7</b>
2.1 Método baseado em conhecimento lingüístico .....	7
2.2 Método baseado em córpus.....	13
2.3 Método híbrido.....	16
2.4 Considerações finais .....	17
<b>Capítulo 3 - Abordagens para a criação de córpus de exemplos.....</b>	<b>19</b>
3.1 Córpus etiquetados manualmente .....	19
3.2 Córpus etiquetados automaticamente.....	21
3.3 Considerações finais .....	23
<b>Capítulo 4 - As estratégias de pré-processamento desenvolvidas .....</b>	<b>24</b>
4.1 Conjunto de palavras .....	24
4.2 Córpus originais .....	25
4.3 Seleção das sentenças .....	26
4.4 Identificação automática das traduções.....	27
4.4.1 Pré-processamento.....	27
4.4.2 Pré-supostos .....	28
4.4.3 Dicionários.....	29

4.4.4	<i>Heurísticas</i>	30
4.4.5	<i>Avaliação 1</i>	32
4.4.6	<i>Avaliação 2</i>	34
<b>4.5</b>	<b>Extração das características</b>	<b>40</b>
4.5.1	<i>Possíveis características</i>	40
4.5.2	<i>Interface com o usuário</i>	42
4.5.3	<i>Exemplos de combinações testadas</i>	44
<b>4.6</b>	<b>Considerações finais</b>	<b>46</b>
	<b>Conclusão</b>	<b>48</b>
	<b>Referências</b>	<b>49</b>

## Lista de Figuras

Figura 1. Exemplo de sentença paralela.....	28
Figura 2. Sentença paralela pré-processada.....	28
Figura 3. Alguns resultados do sistema.....	31
Figura 4. Cobertura do sistema.....	35
Figura 5. Precisão do sistema.....	36
Figura 6. Precisão do sistema x baseline.....	37
Figura 7. Relação entre o nº de possíveis sentidos de cada verbo e a precisão do sistema para tal verbo.....	38
Figura 8. Relação entre o nº de sentidos usados de cada verbo e a precisão do sistema para tal verbo.....	39
Figura 9. Interface do sistema extrator de características.....	43
Figura 10. Exemplos parciais de córpis gerados pelo sistema.....	46

## Lista de Tabelas

Tabela 1. Exemplos de sentenças do BNC com verbos problemáticos.....	25
Tabela 2. Quantidades de sentenças e palavras.....	27
Tabela 3. Quantidades de traduções possíveis para os verbos.....	29
Tabela 4. Precisão do processo de etiquetagem de sentido.....	33
Tabela 5. Cobertura do sistema.....	35
Tabela 6. Precisão do sistema.....	36
Tabela 7. Causas dos erros do sistema.....	37
Tabela 8. Diferentes traduções utilizadas nas 200 sentenças.....	39

# Capítulo 1 - Introdução

## 1.1 Contextualização e motivação

No mundo globalizado, existe a necessidade de uma comunicação mais ativa, dinâmica e sem fronteiras, para que seja possível o fluxo contínuo e eficaz de informações. Independentemente de sua origem ou destino, língua ou cultura, a informação deve permanecer original, sem distorções ou dúvidas de seu significado para que se mantenha seu valor verdadeiro, sem acréscimos ou diminuições e para que não sejam geradas interpretações distorcidas.

Nos dias de hoje, tradutores de texto entre diversas línguas são extremamente difundidos e inseridos no cotidiano de várias atividades profissionais, principalmente das atividades que utilizam a comunicação como matéria-prima e também como produto final a ser comercializado. Pode-se observar que notícias correm o mundo em poucos instantes, estando em países e continentes diferentes (nos quais a língua pode variar grandemente) no formato já traduzido e, muitas vezes, adaptado à cultura e costumes subjacentes às línguas específicas. Isto ocorre devido à velocidade com que se pode traduzir um texto para qualquer outra língua, por meio do auxílio de programas de computador, ou seja, de sistemas de Tradução Automática (TA).

Existem diversos sistemas capazes de traduzir textos para diversas línguas, mas a qualidade das traduções ainda não é satisfatória, principalmente em se tratando de traduções para o português. Uma das maiores dificuldades desses sistemas está relacionada às diferenças lexicais entre as línguas, ou seja, aos diferentes modos pelos quais as línguas caracterizam ou classificam o mundo, quais conceitos elas escolhem para expressar via palavras, e quais elas decidem não utilizar. Em função dessas diferentes classificações do mundo, uma palavra na língua fonte pode ter um uso e uma realização muito diferentes que sua equivalente na língua alvo. Como consequência, a uma única palavra da língua fonte podem corresponder várias palavras da língua alvo, com diferentes significados, dando origem ao problema denominado Ambigüidade Lexical. Esse problema é ainda mais grave nos casos em que as diversas palavras ambíguas, na língua alvo, são da mesma categoria gramatical (por exemplo, verbo, substantivo, etc.). Nestes casos, o problema é denominado

Ambigüidade Lexical de Sentido, já que ocorre variação apenas de sentido (ou significado). Alguns exemplos desse problema, considerando-se traduções do inglês para o português, foco deste trabalho, são (a) *know*, que pode ser traduzido por “saber” e “conhecer”; (b) *bank*, que pode significar “instituição financeira”, “assento” ou, ainda, “banco de areia”; e (c) *light*, que pode ser traduzida como “leve” ou “luz”.

A ambigüidade lexical de sentido não é um problema exclusivo da TA: até mesmo para os tradutores humanos, que são dotados de inteligência e são capazes de armazenar e processar grandes quantidades de informação sobre a língua em uso e sobre o mundo, existe dificuldade em definir o significado de certas palavras em uma tradução. Dotar sistemas computacionais com tal capacitação é, portanto, uma tarefa bastante complexa. A área que se ocupa do tratamento da ambigüidade lexical de sentido, não apenas na TA, mas também em aplicações monolíngües, é denominada Desambiguação Lexical de Sentido (DLS), do inglês *Word Sense Disambiguation*.

Segundo Oliveira et al. (2000), a presença da ambigüidade lexical na TA entre o inglês e o português é bastante freqüente, justificando a necessidade de estratégias de desambiguação nas ferramentas de tradução. Os autores afirmam que a qualidade das escolhas lexicais afeta o processo de tradução em vários graus, principalmente se a escolha incorreta ocorrer em itens lexicais em posições de núcleo, como verbos em um predicado verbal ou substantivos em um sujeito. Nesses casos, a ambigüidade lexical pode prejudicar a coerência local e global da sentença, freqüentemente tornando-a incompreensível.

Também é comprovada a necessidade de tratamento para a DLS na TA por meio do estudo de Fossey et al. (2004), no qual fica evidente que a ambigüidade lexical compromete profundamente a qualidade das traduções produzidas automaticamente e que a solução das questões envolvendo esse problema se mostra um dos caminhos necessários para a obtenção de resultados mais satisfatórios nas produções das ferramentas de TA.

Outro estudo, diretamente relacionado a este trabalho, foi desenvolvido por Specia & Nunes (2004) e é descrito na Seção 4.1. O estudo, realizado com um conjunto de verbos e três sistemas de TA, evidencia que a porcentagem de sentenças nas quais ocorre ambigüidade nos verbos selecionados é bastante grande. Com base em um critério que considera sentenças problemáticas somente aquelas cuja acepção do verbo em foco não era corretamente identificada por pelo menos dois sistemas, 62,6% das sentenças foram

consideradas problemáticas. Se forem consideradas problemáticas as sentenças nas quais a aceção correta do verbo não havia sido identificada por pelo menos um sistema, esse número aumenta para 74,4%. Esse número alto mostra que os sistemas estudados não dispõem de mecanismos de DLS. Normalmente, escolhem uma das possíveis aceções de um verbo, provavelmente a mais comum, e essa aceção é utilizada na tradução da maioria das suas ocorrências, excetuando-se alguns casos do uso do verbo em *phrasal verbs* ou em expressões comuns. O tratamento dispensado a *phrasal verbs* é também bastante simplificado: muitas vezes, um verbo seguido de uma preposição (dois elementos que poderiam compor um *phrasal verb*) é diretamente traduzido como o *phrasal verb* correspondente, mesmo que não seja usado com tal função na sentença.

Em termos gerais, a tarefa de DLS consiste em associar a uma dada palavra de uma sentença um sentido que é distinguível dos outros sentidos potencialmente atribuídos a tal palavra. Vários trabalhos têm sido propostos para a DLS, especialmente voltados para aplicações monolíngües, que apresentam diferenças significativas com relação à TA. Esses trabalhos desenvolvem soluções baseadas em diferentes métodos de Processamento da Língua Natural (PLN) para resolver o problema: método profundo, baseado em conhecimento lingüístico manualmente ou semi-automaticamente especificado; método empírico, baseado em córpus, ou seja, em conhecimento adquirido de córpus de exemplos de tradução, a partir de técnicas de aprendizado de máquina; e método híbrido, baseado em conhecimento lingüístico e em córpus. Trabalhos sob o método baseado em conhecimento são, em geral, mais precisos. Contudo, a necessidade de codificação manual (ou semi-automática) de grandes quantidades de conhecimento acaba limitando sua aplicação a domínios e cenários muito restritos. Trabalhos sob o método baseado em córpus, por outro lado, são menos dependentes de especialistas e mais robustos, mas seus resultados geralmente são menos precisos. Além disso, esses trabalhos são fortemente dependentes de córpus significativos e suficientemente abrangentes das línguas em questão. Já os trabalhos sob o método híbrido permitem combinar as características dos outros dois métodos, em teoria, unindo suas vantagens e minimizando suas deficiências.

Para aplicações multilíngües, em particular, há poucas propostas de DLS. Geralmente, elas seguem métodos profundos, delimitando o problema a um recorte bastante pequeno das línguas e, portanto, não têm aplicação efetiva em sistemas para a TA irrestrita. O contexto deste trabalho são as propostas que visam uma aplicação mais

abrangente, ou seja, as proposta que seguem métodos baseados em *córpus* ou híbridos. Um dos motivos para a pequena quantidade de trabalhos baseados nesses métodos, certamente, é a inexistência de *córpus* adequados, ou seja, suficientemente abrangentes e contendo as informações necessárias sobre a tradução entre as línguas em questão. A criação de *córpus* dessa natureza requer um trabalho de pré-processamento, de preparação dos dados para a sua utilização no processo de aprendizado automático.

No caso da desambiguação para a tradução, o pré-processamento consiste na criação de *córpus* de exemplos de tradução anotados, ou seja, de conjuntos de sentenças na língua fonte, com as traduções das palavras sob consideração indicadas. Além das traduções, esses exemplos podem possuir outros tipos de informação, como a categoria gramatical das palavras. Quanto maior a quantidade de informações relevantes, potencialmente, melhores resultados serão obtidos no aprendizado. Por outro lado, maior será o esforço para a criação do *córpus*.

Essa etapa de pré-processamento é, portanto, indispensável para os trabalhos baseados em *córpus* e híbridos e deve ser realizada de tal maneira a garantir a qualidade dos dados resultantes, bem como a fornecer o máximo de informações relevantes. Uma estratégia para a realização dessa etapa é a obtenção e representação dos exemplos e das informações sobre eles manualmente. Em um extremo, sentenças na língua fonte poderiam ser traduzidas manualmente e tais traduções, juntamente com as demais informações sobre as palavras, também manualmente identificadas e anotadas. Essa estratégia garante a qualidade dos exemplos, mas demanda muito tempo e esforço, tornando-se inviável quando se pretende representar diferentes tipos de informação e/ou uma grande quantidade de exemplos. Estratégias mais viáveis consideram a aquisição e/ou representação semi-automática dos exemplos e informações. Por exemplo, podem ser utilizados *córpus* paralelos entre as duas línguas, ou seja, *córpus* com os textos originais e suas traduções, e a identificação e anotação manual das traduções das palavras em foco, evitando-se com isso a necessidade de gerar traduções. Podem também ser utilizados processos de PLN para a geração automática das informações, como um etiquetador gramatical e um analisador sintático. No outro extremo, ideal, estão as estratégias para a aquisição e representação dos exemplos e informações de maneira completamente automatizada. Essas estratégias, embora exijam também tempo e esforço para o seu desenvolvimento, uma vez criadas minimizam enormemente os esforços com o trabalho de pré-processamento para métodos

de aprendizado de máquina, principalmente quando se visa a grandes quantidades de exemplos e informações.

## 1.2 Objetivos

O objetivo deste trabalho foi investigar, propor e implementar estratégias para automatizar a criação de *cópus* de exemplos a serem utilizados em abordagens empíricas para o desenvolvimento de módulos de DLS para sistemas de TA inglês-português. Além da tradução, as estratégias foram planejadas e desenvolvidas de modo que os *cópus* possam apresentar outras informações úteis para a tarefa de desambiguação.

## 1.3 Metodologia

As estratégias desenvolvidas neste trabalho podem ser divididas em dois grupos: (1) identificação e anotação das traduções; e (2) extração de características. Como escopo inicial do trabalho para verificar a viabilidade das estratégias, foram considerados sete verbos altamente ambíguos e problemáticos na TA: *to come*, *to get*, *to give*, *to go*, *to look*, *to make* e *to take*.

As estratégias do primeiro grupo são baseadas principalmente em *cópus* paralelos inglês-português alinhados em nível de sentenças, ou seja, com a indicação das correspondências entre as sentenças nas duas línguas. Além disso, utilizam informações fornecidas por alguns processos de PLN, basicamente, etiquetadores gramaticais para as duas línguas e um lematizador parcial para o português. Considerando-se as informações providas por esses recursos e processos, são definidas heurísticas para identificar, no *cópus* paralelo, a tradução de cada um dos verbos em questão. Essa tradução é então anotada na correspondente sentença em inglês, juntamente com as demais informações fornecidas pelos processos de PLN, gerando um *cópus* intermediário.

As estratégias do segundo grupo consistem em, a partir do *cópus* em inglês anotado com a tradução e demais informações, extrair as informações (ou características) relevantes, de acordo com um conjunto pré-definido de características, e representá-las em

um formato adequado para a sua utilização por algoritmos de aprendizado de máquina, neste caso, os algoritmos do ambiente Weka<sup>1</sup> (Witten & Frank, 2000). Essa extração é realizada de maneira parametrizável, permitindo diversas variações e combinações de características, bem como a inclusão de informações provenientes de outros recursos ou processos, ainda não representadas no cópús intermediário, por exemplo, as relações provenientes de um analisador sintático.

## 1.4 Resultados

Com a implementação das estratégias e propostas, foram produzidos dois sistemas que geram dados para serem diretamente utilizados como entrada para processos de aprendizado de máquina, com a finalidade de geração de um modelo de DLS na TA. O primeiro sistema identificado gera um cópús intermediário anotado com a tradução do verbo e com atributos necessários. O segundo sistema extrai desse cópús características de maneira parametrizável, por meio de uma interface com o usuário, gerando um novo cópús que será utilizado no processo de aprendizado automático.

## 1.5 Organização

Este trabalho está organizado da seguinte maneira: No Capítulo 2 são descritos os principais trabalhos de DLS voltados especificamente para a TA, seguindo os diferentes métodos de PLN. No Capítulo 3 são descritos alguns trabalhos seguindo as duas abordagens para a criação de cópús de exemplos a serem utilizados em sistemas baseados em cópús ou híbridos de DLS: criação manual e criação automática. No Capítulo 4 são apresentadas as estratégias desenvolvidas neste trabalho para a automatização da criação e configuração dos cópús de exemplos de DLS. Por fim, são apresentadas algumas conclusões e trabalhos futuros.

---

<sup>1</sup> <http://www.cs.waikato.ac.nz/~ml/>

## Capítulo 2 - Ambigüidade Lexical de Sentido na TA

Os trabalhos de DLS podem seguir os diferentes métodos de PLN: 1) método baseado em conhecimento lingüístico manualmente especificado; 2) método baseado em corpus de exemplos e em algoritmos de aprendizado de máquina para adquirir conhecimento automaticamente a partir dos exemplos; ou 3) método híbrido, que combina características de ambos os métodos. A seguir, são descritos alguns exemplos de trabalhos de DLS voltados para a TA que seguem esses três métodos, incluindo alguns trabalhos efetivamente inseridos no contexto de algum sistema de TA. Vale notar que a maioria desses trabalhos considera a TA entre outras línguas, não envolvendo o português.

### 2.1 Método baseado em conhecimento lingüístico

Goodman & Nirenburg (1991) descrevem a criação de um sistema de TA por interlíngua para a tradução de manuais técnicos (sobre computadores) entre o inglês e o japonês. Esse sistema, também baseado em conhecimento lingüístico profundo, é denominado KBMT (*Knowledge-Based Machine Translation*). A sua interlíngua consiste de uma hierarquia conceitual que foi manualmente construída, especificamente para esse sistema. Os itens lexicais são representados em dicionários monolíngües e mapeados nos conceitos dessa ontologia, que são independentes de língua e, em princípio, não ambíguos.

Nesse sistema, não há um módulo específico de DLS, mas as ambigüidades na língua-fonte são resolvidas durante o processo de mapeamento dos itens lexicais da língua-fonte em conceitos não ambíguos da interlíngua, por meio de restrições de seleção. Isso é possível porque a ontologia é delimitada a um único domínio. Para sistemas independentes de domínio, abordagens de DLS fundamentadas principalmente em uma ontologia seriam pouco viáveis, dada a complexidade para a construção de uma ontologia dessa natureza e a quantidade limitada de conhecimento que ela poderá prover.

Outros sistemas de TA, como o EUROTRA (Copeland, 1991) e METAL (Gajek, 1991), empregam procedimentos mais simples de DLS. Eles procuram tratar a ambigüidade lexical por meio da definição de estruturas argumentais e de restrições ou

preferências de seleção sobre essas estruturas. No sistema EUROTRA, em particular, uma hierarquia simples de tipos semânticos (entidade, humano, não-humano, etc.) é utilizada para tratar os casos de desambiguação mais refinada, com base em preferências de seleção. O sistema aplica a noção de distância semântica entre os nós dessa hierarquia. Para a desambiguação de um substantivo que complementa o verbo em uma sentença, por exemplo, a hipótese é de que quanto menor a distância entre o nó que representa o sentido de um substantivo e os nós que representam as restrições impostas na estrutura argumental do verbo em questão, maior a indicação de que esse é o sentido do substantivo. Contudo, de modo geral, essas restrições são simples e limitadas, de modo que resolvem apenas alguns casos mais simples de ambigüidade.

Sistemas comerciais de TA em uso atualmente que oferecem algum tratamento à DLS empregam métodos ainda mais simples, em função da necessidade de abrangência a qualquer gênero e domínio de textos. O Systran®<sup>2</sup>, considerado por muitos como o melhor sistema de TA disponível atualmente, adota uma visão bastante prática do processo de DLS: procura identificar o domínio do texto sendo traduzido para acessar dicionários específicos de cada domínio. Isso é feito com base na análise de traços sintático-semânticos (objeto concreto, sujeito humano, etc.) e das categorias semânticas (dispositivo, propriedade, etc.) das palavras do contexto, armazenadas nos dicionários do sistema. Contudo, nem todas as entradas possuem essas informações e o seu uso não é efetivo, na maior parte dos casos. Além disso, dependendo do tamanho do texto a ser traduzido e da sua natureza, a identificação do domínio não é possível. Para os casos mais simples, o Systran também possui entradas específicas para algumas expressões idiomáticas, locuções comuns e termos técnicos de diversas áreas.

O sistema UNITRAN de tradução automática por interlíngua entre o inglês, o espanhol e o francês (Dorr, 1993) é um exemplo representativo das abordagens mais refinadas empregadas para o tratamento da ambigüidade lexical nos sistemas de TA. Contudo, não se tem como objetivo, neste caso, obter um sistema comercial.

O UNITRAN não dispõe de um mecanismo específico para esse problema. Seu tratamento é embutido em outros módulos do sistema. O sistema utiliza estruturas conceituais lexicais, tanto para a representação dos itens lexicais quanto das estruturas

conceituais compostas por vários itens. A interlíngua do sistema corresponde à composição de várias dessas estruturas para a representação de sentenças específicas, de acordo com as palavras da sentença.

No UNITRAN, todos os problemas (ou “divergências”) de tradução, em diversos níveis, são tratados de acordo com uma estratégia similar. A ambigüidade lexical, em especial, é considerada uma das variações do problema de divergência lexical. Ela é tratada como um problema de seleção lexical, na realização das estruturas conceituais compostas para a língua-alvo. A necessidade de escolha ocorre quando uma parte da estrutura conceitual composta (que representa um conceito) pode combinar com mais de uma estrutura conceitual lexical da língua-alvo, ou seja, quando um conceito pode ser realizado por mais de uma palavra. Essa escolha é feita por meio da verificação das estruturas lexicais que satisfazem as restrições de seleção sintáticas e semânticas presentes na estrutura conceitual composta, por meio de um processo similar ao de unificação. Várias estruturas podem combinar em todos esses itens, de modo que, em alguns casos, o sistema retorna mais de uma realização lexical. Contudo, segundo a autora, a idéia não é, de fato, encontrar a melhor combinação, mas simplesmente encontrar combinações satisfatórias.

Dorr afirma que essas restrições não capturam distinções que não sejam caracterizadas por propriedades puramente sintáticas, por exemplo, por distinções que dependem de conhecimento do discurso, de domínio ou de mundo, conhecimento de usos idiomáticos, etc. Essa limitação nos tipos de conhecimento considerados no sistema, bem como o fato de o sistema poder retornar várias realizações lexicais, implica, certamente, que muitos casos de ambigüidade não são resolvidos.

Egedi et al. (1994) apresentam um sistema de TA por transferência entre coreano e o inglês, que possui um módulo de DLS para tratar da ambigüidade de alguns verbos. Ele se baseia na unificação de restrições de seleção semânticas definidas na estrutura argumental desses verbos com os traços semânticos definidos para os substantivos que podem ser utilizados como seus argumentos. As regras de transferência, incluindo as restrições de seleção e os traços semânticos, são manualmente especificadas.

---

<sup>2</sup> <http://www.systransoft.com>

A DLS ocorre no processo de transferência lexical, com base nas possíveis traduções especificadas em um dicionário bilíngüe e nas restrições de seleção e traços semânticos especificados na língua-alvo. Os autores justificam a especificação desse conhecimento na língua-alvo porque, segundo eles, a seleção lexical normalmente depende da existência de traços semânticos nos elementos da língua-alvo que são completamente irrelevantes para a língua-fonte. Eles citam, como exemplo, a tradução do verbo *wear*, do inglês para o coreano. No coreano, a tradução depende do complemento do verbo: “*wear clothes*” e “*wear socks*” são traduzidos por verbos completamente diferentes. No entanto, no inglês, não há distinção.

Pedersen (1997) descreve uma abordagem baseada em teorias da semântica lexical para a desambiguação de um subconjunto de verbos de movimento polissêmicos na TA do dinamarquês para o inglês. A autora considera apenas o fenômeno da polissemia sistemática desse subconjunto. O seu objetivo é identificar padrões para o tratamento de polissemia sistemática, ou seja, que possam ser aplicados a diversos verbos com significado relacionado, dentre os verbos de movimento, formalizar e implementar esses padrões na forma de regras lexicais que possam ser usadas para a DLS na TA.

Para tanto, primeiramente, é realizada uma análise das ocorrências de 100 verbos de movimento em diferentes corpúscos do dinamarquês para verificar propriedades estatísticas (frequência, co-ocorrências, etc.) e outras características do uso desses verbos, bem como os tipos de conhecimento que são necessários para diferenciar os seus sentidos. Para essa análise, foram selecionados de 100 a 300 exemplos de ocorrência de cada um dos verbos. A partir da análise, os exemplos foram manualmente categorizados de acordo com suas propriedades sintáticas e semânticas. Nessa etapa, foram estabelecidas várias delimitações, por exemplo, foram descartados exemplos do uso do verbo em expressões idiomáticas e metafóricas. No processo de categorização foram agrupados os exemplos de acordo com os padrões de valência do verbo de movimento, separados os exemplos com verbos que possuíam elementos modificadores de direção dos que não possuíam, etc. Como resultado dessa etapa, foram formados grupos de exemplos com propriedades similares, por exemplo, exemplos de verbos de movimento que têm uma direção específica, cujo agente é animado e que implica o movimento de partes do corpo ou de uma máquina.

A partir dessa análise, os verbos foram classificados em uma taxonomia para os verbos de movimento, de acordo com propriedades sintáticas e semânticas e, principalmente, com as regularidades nos desvios do significado básico para os demais sentidos. Nesses verbos, segundo a autora, a polissemia deve se manifestar de maneira sistemática, de modo que todos os verbos do grupo podem receber o mesmo tratamento na DLS.

Para representar os verbos dos grupos, a autora definiu um modelo lexical. Também foram definidas uma hierarquia conceitual parcial e restrições de seleção para substantivos distribuídos nessa hierarquia. Os verbos são então especificados de acordo com o modelo definido, utilizando uma grande quantidade de informações lingüísticas na língua-fonte, em diversos níveis, que indicam os desvios de significado e, portanto, podem auxiliar na desambiguação. Os esquemas especificados foram implementados na forma de regras lexicais e incorporados a um sistema de interpretação do dinamarquês. A autora realiza um teste com 42 sentenças com os verbos ambíguos. Desses verbos, 39 foram corretamente desambiguados.

Dorr & Katsova (1998) definem um mecanismo de seleção lexical para verbos e substantivos derivados de verbos na TA (entre o inglês e o espanhol) que se baseia na estrutura argumental desses elementos, representada por meio de estruturas conceituais lexicais (LCSs), e nos sentidos da WordNet (Miller et al., 1990). A hipótese é de que a tradução de um elemento da língua-fonte pode ser desambiguada se forem escolhidos, na língua-alvo, elementos que tenham a mesma LCS e que pertençam ao mesmo grupo de sinônimos (*synset*) da WordNet, ou seja, que sejam sinônimos do elemento na língua-fonte.

Para testar sua hipótese, as autoras implementam um algoritmo de seleção lexical que utiliza um sistema já existente para codificar sentenças em suas representações LCSs. Esse sistema também possui um léxico do inglês e outro do espanhol, cujas entradas estão codificadas como LCSs, com um código correspondente ao *synset* da WordNet ao qual pertencem (anotado manualmente). Com base na estrutura gerada pelo sistema para uma sentença, o algoritmo extrai a estrutura LCS genérica do verbo a ser desambiguado, sem as constantes que representam as palavras da sentença, e recupera do léxico do espanhol todas as entradas correspondentes a verbos que têm a LCS com as mesmas propriedades estruturais. Por exemplo, para o verbo *sap*, são recuperados 358 verbos do espanhol com a

mesma estrutura de LCS. Desse conjunto de verbos, o algoritmo seleciona apenas aqueles que apresentam o mesmo código do *synset* que o verbo sendo desambiguado. Se o verbo puder pertencer a vários *synsets*, são selecionados todos os verbos em todos os seus *synsets*. Para o verbo *sap*, apenas um verbo (*escurir*) pertence ao mesmo *synset*. Esse verbo é então escolhido como a tradução mais adequada para o verbo do inglês.

Caso haja mais de um verbo com a mesma estrutura e o mesmo código de *synset*, o algoritmo retorna todos eles. Por outro lado, caso não seja encontrado nenhum verbo com a LCS equivalente no mesmo *synset*, o algoritmo estende a busca aos *synsets* hiperônimos em um nível (mais genéricos) de todos os *synsets* aos quais o verbo pertence.

O algoritmo pode operar também na DLS monolíngüe. Nesse caso, as buscas por LCSs equivalentes são feitas no léxico da própria língua. As autoras realizam experimentos para a DLS de três verbos do inglês, monolíngüe e multilíngüe. Na DLS monolíngüe, dois dos verbos possuem exatamente um equivalente em estrutura e *synset*, enquanto para o outro verbo só é encontrado um equivalente quando são analisados os *synsets* hiperônimos. Na DLS multilíngüe, um verbo possui exatamente uma tradução, enquanto os outros dois possuem duas e quatro traduções. Nenhum outro tipo de conhecimento é empregado para filtrar essas possíveis traduções.

As autoras afirmam que o seu método é mais efetivo para a DLS monolíngüe. Mencionam também que se a DLS monolíngüe for realizada como um pré-processamento para a multilíngüe, ela pode reduzir o número de ambigüidades, melhorando a precisão na tradução.

Não são realizados experimentos de avaliação mais abrangentes, mostrando se a abordagem é realmente viável. Um problema dessa abordagem é que como são recuperadas todas as LCSs estruturalmente equivalentes, podem ser recuperados verbos que, apesar de estarem no mesmo *synset*, não são válidos como tradução do verbo na língua-fonte. Um possível filtro, bastante simples, seria buscar apenas as estruturas dos verbos que podem ser traduções do verbo na língua-fonte, a partir da consulta a um dicionário bilíngüe. Com relação à abrangência, o sistema é limitado às LCSs que já estão codificadas no léxico, às quais já foi atribuído um código de *synset*. Além disso, o fato de serem retornadas todas as traduções possíveis indica que o sistema não elimina todas as ambigüidades.

O único trabalho multilíngüe voltado explicitamente para a DLS e envolvendo o português é o de Leffa (1998), que focaliza a importância do uso do contexto local da palavra ambígua, isto é, das palavras vizinhas a ela na sentença, na forma de colocações (*collocations*), para a desambiguação na TA. Ele afirma que colocações são mais efetivas para a DLS que outras características mais profundas, como conhecimento de mundo, devido à dificuldade em se representar e utilizar esse conhecimento e à natureza dinâmica do uso das palavras.

Leffa também defende a análise do uso das palavras em córpus para definir o seu conjunto de possíveis sentidos. Segundo ele, em um contexto multilíngüe, é possível estabelecer uma metodologia bastante objetiva para a definição desse conjunto de sentidos, a partir de exemplos de tradução. Para investigar sua hipótese, o autor realiza um experimento para desambiguar 20 substantivos ambíguos do inglês para o português, contextualizados em exemplos de tradução extraídos de um córpus de 20.000.000 de palavras.

Para cada palavra, foram aleatoriamente selecionados 200 exemplos, sendo que cada exemplo consiste de um segmento com 20 palavras, em média. O autor não menciona como as regras de desambiguação são construídas, apenas que são incorporadas às regras de um sistema de TA inglês-português em fase inicial de construção. Ao que tudo indica, as regras são manualmente codificadas, com base em um conjunto de colocações pré-definidas, que também não são explicitadas no trabalho. Na sua avaliação, o autor relata uma acurácia média de 94% para as 20 palavras. No entanto, como esse modelo se baseia apenas nas palavras da sentença, na forma de colocações, sua abrangência deve ser bastante limitada. Tanto o módulo de DLS quanto o sistema de TA mencionados não foram concluídos.

## **2.2 Método baseado em córpus**

Entre os trabalhos baseados em córpus para a DLS, alguns seguem o modo não-supervisionado de aprendizado, ou seja, utilizam córpus de exemplos apenas com informações da língua fonte, não anotados com as respectivas traduções das palavras. Em se tratando de trabalhos para a TA, no entanto, a maioria dos trabalhos segue o modo

supervisionado, ou seja, a partir de *córpus* de exemplos anotados com as traduções das palavras em questão, ou utiliza *córpus* bilíngües paralelos como fonte de informação para identificar tais traduções. O modo supervisionado (ou o uso de *córpus* paralelos) é mais indicado para a DLS na TA, uma vez que o conjunto de possíveis traduções precisa ser previamente definido, diferentemente do que ocorre na desambiguação monolíngüe.

Brown et al. (1991) usam um modelo estatístico, baseado em informação mútua, para a seleção lexical de itens ambíguos na TA do francês para o inglês. Para tanto, são extraídas de um *córpus* paralelo entre as duas línguas as possíveis traduções das palavras do francês para o inglês que apresentam um alinhamento direto (um para um). Em seguida, é definido um conjunto de possíveis características úteis para distinguir, com base no contexto local das sentenças na língua-fonte (francês) ou na língua-alvo (inglês, considerando a sentença parcialmente traduzida por um sistema de TA), qual é a tradução adequada para uma palavra na língua-alvo. As características incluem diferentes palavras do contexto da palavra ambígua na língua-fonte, por exemplo, o primeiro substantivo à direita, o primeiro verbo à direita, a primeira palavra à esquerda, etc.

O modelo estatístico empregado tenta encontrar, para cada palavra ambígua, uma única característica (dentre as pré-definidas) que indica, com um alto nível de confiabilidade, qual a sua tradução. O algoritmo considera uma desambiguação binária, ou seja, a escolha entre apenas duas possíveis traduções de uma palavra ambígua. O processo é iterativo e, a cada interação, o algoritmo procura aumentar a informação mútua obtida com o emprego da característica para a desambiguação da palavra em questão. O critério de parada é indicado pela estabilização da informação mútua, ou seja, na iteração em que essa medida não pode mais ser aumentada.

Definida a característica que melhor divide o conjunto de treinamento, os exemplos (em francês) podem ser divididos em dois grupos, para as duas traduções possíveis, de acordo com o valor que apresentam para essa característica. Na verdade, cada um dos grupos pode ter várias traduções, mas elas são ranqueadas de acordo com uma estimativa da probabilidade de cada uma das traduções no *córpus* do inglês. A tradução com a maior probabilidade de ocorrência em cada grupo é então escolhida para etiquetar a palavra ambígua em todos os exemplos do francês selecionados.

Para a avaliação do módulo de DLS, o modelo foi treinado para a desambiguação das 500 palavras mais comuns do inglês e as 200 mais comuns do francês e o módulo resultante foi incorporado a um sistema de TA por transferência, também estatístico, na fase de análise. Na tradução de 100 sentenças aleatoriamente selecionadas com essas palavras, os autores relatam uma diminuição de 13% na taxa de erro das traduções resultantes do sistema com o uso do módulo.

Lee (2002) apresenta uma abordagem de DLS para um sistema de TA do inglês para o coreano que segue o método direto de tradução por palavras, empregando técnicas estatísticas para a seleção lexical e a re-ordenação das palavras na língua-alvo. A DLS é, portanto, embutida no módulo de seleção lexical.

A abordagem de TA utiliza cópulas paralelos entre as duas línguas e dicionários bilíngües. Para a DLS, com base nos documentos paralelos, é criado um dicionário de tradução para cada palavra da língua fonte, que consiste de todas as suas possíveis traduções na língua-alvo, extraídas do cópulas paralelo. A partir desses dicionários, o problema de DLS é estruturado como um problema de classificação. Para tanto, são usadas como características “co-ocorrências”, que correspondem a todas as combinações de palavras (tomadas de duas a duas) na sentença a ser traduzida. O algoritmo de aprendizado supervisionado empregado é o SNoW, que aprende, como modelo, uma rede de funções lineares com regras de atualização.

Como cópulas paralelo, é utilizado um conjunto de 689 documentos (17.846 sentenças) manualmente traduzidos do inglês para o coreano, manualmente alinhados por palavras. Assim, os dicionários de tradução são diretamente extraídos desses documentos. Para avaliar sua abordagem, são selecionados exemplos de 121 substantivos ambíguos que possuem mais de 50 exemplos no cópulas. Os resultados da classificação foram comparados à *baseline* da escolha pelo sentido mais freqüente e à classificação utilizando um algoritmo Naïve Bayes. O classificador gerado pelo algoritmo SNoW apresenta uma precisão média de 57.46%, superior à precisão da *baseline* (53.87%) e do classificador Naïve Bayes (47.49%).

Dihn et al. (2003) descrevem um sistema de TA do inglês para o vietnamita, desenvolvido de acordo com um método híbrido: parte do sistema é constituída de regras manualmente criadas e outra parte, de regras aprendidas a partir de cópulas, com base no

aprendizado baseado em transformações (Brill, 1995). Esse sistema possui módulos específicos para cada tipo de ambigüidade, incluindo um módulo para a DLS. As regras desse módulo são geradas por uma abordagem baseada em córpus.

O córpus de exemplos é criado a partir de textos paralelos entre as duas línguas, de diversos gêneros e domínios, por meio do alinhamento automático das palavras, revisado manualmente. As características para o aprendizado consistem de n-gramas (de uma a quatro palavras), etiquetas gramaticais e funções sintáticas. Além disso, o algoritmo considera as etiquetas já atribuídas às palavras vizinhas na sentença, ou seja, as palavras já traduzidas. Isso é possível porque no córpus de exemplos, todas as palavras estão etiquetadas com a tradução correspondente. Os autores não avaliam os módulos individuais do sistema, tampouco a influência desses módulos no desempenho geral do sistema de TA.

### **2.3 Método híbrido**

O único trabalho híbrido de DLS voltado para a TA de que se tem conhecimento é o de Zinovjeva (2000). A autora emprega o método de aprendizado por transformações (Brill, 1995) com o objetivo de aprender automaticamente regras (simbólicas) para traduzir corretamente palavras ambíguas do inglês para o sueco, em textos irrestritos, de qualquer gênero e domínio.

Um conjunto de exemplos de treinamento é criado a partir de sentenças manualmente etiquetadas com a tradução dos verbos e substantivos ambíguos. A partir desses exemplos, são realizados alguns experimentos de aprendizado, cada um considerando determinados tipos de conhecimento. Com esses experimentos, a autora pretende verificar quais conhecimentos são mais adequados para, assim, empregá-los na construção do seu modelo de DLS. Os experimentos consideram cada palavra ambígua, individualmente, e assumem que as palavras da sentença já possuem etiquetas gramaticais, corretamente atribuídas em uma etapa de pré-processamento.

No primeiro experimento, são consideradas apenas as palavras vizinhas à palavra ambígua na sentença. São criados modelos para três palavras, dois substantivos e um verbo. Os exemplos incluem 4.800 ocorrências de cada substantivo e 780 ocorrências do

verbo. Cerca de 10% desses exemplos são usados para teste e o restante, para o treinamento. A acurácia obtida foi de 92.1%, 95.2% e 73.1%.

O segundo experimento considera, em vez das palavras vizinhas, as suas categorias gramaticais. Considerando a mesma configuração que a do primeiro experimento, as acurácias obtidas mudaram para 93.6%, 85.4% e 80.8%.

O terceiro experimento considera as relações sintáticas das palavras do contexto da palavra ambígua, produzidas por um *parser*. Apenas o modelo para o verbo é gerado, considerado 78 das suas ocorrências. A acurácia obtida foi de 83.3%. Segundo a autora, essa acurácia relativamente baixa deve-se ao tamanho reduzido do conjunto de treinamento.

O quarto experimento considera a combinação das etiquetas gramaticais com as relações sintáticas. Novamente, apenas o modelo para o verbo, com 78 das suas ocorrências, é gerado. A acurácia obtida foi de 84.6%, pouco maior que a do experimento anterior.

A cada experimento, uma etapa subsequente de alteração manual das regras foi realizada, visando aperfeiçoar regras muito genéricas ou muito específicas. A avaliação dos modelos considerando essas alterações levou a uma acurácia superior, em todos os casos. Vale notar que essa alteração só foi possível porque as regras são simbólicas.

As regras que apresentaram a maior acurácia são incorporadas a um sistema de TA já existente, também baseado em transformações. Sem o módulo de DLS, o sistema necessita da interação com o usuário para que ele escolha entre todas as possíveis traduções de palavras ambíguas.

## **2.4 Considerações finais**

Independente do método de PLN utilizado nos trabalhos descritos, eles evidenciam que para um sistema de TA obter resultados satisfatórios é imprescindível a utilização de um mecanismo de DLS. Pode-se perceber, também, que os trabalhos atuais ainda apresentam uma série de limitações, principalmente com relação à sua abrangência, e que

nas pesquisas mais recentes o foco parece estar voltado para abordagens baseadas em *cópus*, visando justamente minimizar tal limitação.

Considerando-se principalmente a tradução inglês-português, observa-se que não há trabalhos relevantes. Portanto, o desenvolvimento de recursos que facilitem a geração de dados de entrada, ou seja, de *cópus* de exemplos para abordagens empíricas, pode contribuir enormemente para que trabalhos surjam nessa área, e conseqüentemente, para que módulos de DLS para a TA inglês-português efetivos possam ser propostos.

## Capítulo 3 - Abordagens para a criação de **córpus de exemplos**

O foco deste trabalho está no desenvolvimento de módulos de DLS seguindo o método baseado em córpus e o modo supervisionado de aprendizado, mais indicado para a TA, conforme mencionado. Para tanto, é necessária a criação de córpus de exemplos apropriados, ou seja, anotados.

Os córpus para a DLS supervisionada podem ser criados manual ou automaticamente. Neste capítulo, são brevemente apresentados alguns córpus criados manualmente comumente utilizados pelos trabalhos de DLS, principalmente monolíngües. Na seqüência, são apresentadas algumas abordagens de criação automática de córpus. Vale notar, todavia, que os córpus existentes são, em geral, monolíngües. Da mesma maneira, as abordagens existentes são voltadas para criação automática de córpus monolíngües, anotados com o sentido das palavras, em vez de suas traduções.

### **3.1 Córpus etiquetados manualmente**

Os principais exemplos de córpus disponíveis e que são comumente utilizados para o treinamento e avaliação de trabalhos de DLS são os córpus DSO (Ng & Lee, 1996) e SEMCOR (Miller et al., 1994). Ambos os córpus foram criados para a desambiguação monolíngüe do inglês, utilizando os sentidos da WordNet.

O maior e mais significativo desses córpus é o DSO. Ele consiste de 192.800 sentenças de exemplo contendo 192.874 ocorrências dos 121 substantivos e 70 verbos mais freqüentes da língua inglesa, extraídas do córpus Brown (Francis & Kucera, 1979) e de um córpus de artigos do *Wall Street Journal*. Em média, cada verbo considerado possui 12 sentidos, enquanto cada substantivo possui 7.8 sentidos. Para cada palavra, foram extraídos até 1.500 exemplos. O processo de etiquetagem manual do córpus estendeu-se por um ano.

O córpus SEMCOR também consiste de um subconjunto do córpus Brown, com cerca de 200.000 palavras, sendo que as palavras de conteúdo foram manualmente etiquetadas com os sentidos da WordNet.

Outros *córpus* menores são os criados em determinados trabalhos de DLS e disponibilizados para uso em outros trabalhos. Por exemplo, os *córpus* criados por Leacock et al. (1993) e Bruce & Wiebe (1994), cada um com pouco mais de 2.000 sentenças de exemplos com seis diferentes sentidos da palavra *line* e *interest*, respectivamente. Outros exemplos são os *córpus* usados nas três edições do exercício de avaliação conjunta de DLS SENSEVAL<sup>3</sup>. Com exceção da primeira edição, os demais *córpus* são baseados nos sentidos da WordNet.

Contudo, como afirma Ng (1997b), esses *córpus*, incluindo o DSO, são ainda muito pequenos para serem utilizados para a criação de abordagens irrestritas de DLS. Com base no DSO, o autor examina o efeito do tamanho do *córpus* de treinamento, em termos do número de exemplos para a DLS. Para tanto, ele define uma abordagem baseada em instâncias e realiza testes com vários subconjuntos do *córpus*, de modo a obter as curvas de aprendizado nesse *córpus*. Os resultados do experimento mostram que a precisão aumenta à medida que o número de exemplos do *córpus* cresce e que todos os exemplos do *córpus* são efetivamente utilizados pelo algoritmo empregado.

Como conclusão desses experimentos, o autor estima que um *córpus* de 3.200 palavras diferentes etiquetadas com seus sentidos é suficiente para construir um sistema de DLS de ampla cobertura e alta precisão, considerando-se qualquer palavra de conteúdo, em textos irrestritos da língua inglesa. Assumindo uma média de 1.000 ocorrências etiquetadas por sentido por palavra, isso significa um *córpus* de 3.2 milhões de palavras etiquetadas. Com base na sua experiência com a criação do DSO, segundo o autor, a produção manual desse *córpus* demandaria um tempo de 16 anos, considerando-se o esforço de um etiquetador humano.

Uma alternativa para o problema da etiquetagem manual que tem sido investigada ultimamente e é o foco deste trabalho, é a etiquetagem automática dos sentidos (ou traduções) dos exemplos.

---

<sup>3</sup> <http://www.senseval.org/>

## 3.2 Córpus etiquetados automaticamente

Segundo Agirre & Martínez (2004), a criação automática de córpis é uma das estratégias mais indicadas para minimizar o problema do gargalo da aquisição do conhecimento, contudo, é ainda muito pouco explorada. Para Dagan & Itai (1994), além de permitir a aquisição de córpis mais representativos, a etiquetagem automática permite capturar distinções diferentes das que seriam atribuídas por um anotador humano, por exemplo, distinções específicas de algum domínio ou pouco comuns.

Uma possibilidade para a criação automática de córpis é a exploração de textos paralelos. O uso dessa estratégia pode facilitar principalmente a criação de córpis para trabalhos multilíngües. Contudo, essa estratégia tem sido pouco investigada nesse sentido. Alguns exemplos do uso de córpis paralelos para a criação de córpis para a DLS monolíngüe são os trabalhos de Ide et al. (2002) e Diab & Resnik (2002).

Ide et al. (2002) utilizam textos paralelos em sete línguas para verificar em que nível as traduções para os diferentes significados de um item polissêmico do inglês são lexicalizadas por itens diferentes nessas línguas. Um algoritmo de *clustering* é utilizado para criar grupos de sentidos de acordo com as diferentes traduções de cada palavra do inglês, nas diferentes línguas. As distinções de sentido são, então, adquiridas a partir do córpis.

Diab & Resnik (2002), por sua vez, propõem uma abordagem para a criação de um córpis etiquetado com sentidos a partir de córpis paralelos bilíngües, produzidos por um sistema de TA e de um inventário de sentidos pré-definido da língua para a qual se pretende criar o córpis etiquetado (língua-alvo). Os textos paralelos são automaticamente alinhados por sentenças e por palavras. Esse alinhamento permite identificar, nos textos da língua-alvo, quais as traduções correspondentes a palavras da língua-fonte. As palavras que são traduções de uma mesma forma na língua-fonte são, então, agrupadas. Para cada um dos grupos gerados, são considerados todos os possíveis sentidos para cada palavra. A etiqueta de sentido adequada para cada palavra é atribuída de acordo com a sua similaridade semântica com as outras palavras no grupo. Apesar da facilidade na geração do córpis paralelo alinhado, é importante ressaltar que esse córpis pode apresentar

diversos erros decorrentes de traduções automáticas ou alinhamentos automáticos inadequados, os quais podem propagar-se pelo processo de criação do *corp*us.

Seguindo uma metodologia diferenciada, sem a utilização de *corp*us paralelos, Agirre & Martínez (2004) descrevem um processo de criação automática de *corp*us de exemplos etiquetados, também monolíngüe. O método empregado é o proposto por Leacock et al. (1998), que se baseia nos “parentes” não-polissêmicos dos itens ambíguos para obter exemplos etiquetados com sentidos para esses itens. Os parentes, nesse caso, são os sinônimos dos itens ambíguos. Para cada item polissêmico, são realizadas buscas na *web*, considerando sentenças de busca com os sinônimos não-polissêmicos para recuperar exemplos contendo esses sinônimos. A suposição do método é de que para um determinado sentido da palavra ambígua, se for possível encontrar um sinônimo não-ambíguo desse sentido, então os exemplos que contêm esse sinônimo devem ser muito similares ao sentido da palavra ambígua e podem, portanto, ser usados para gerar um modelo supervisionado para tal sentido da palavra.

Assim como Agirre & Martínez, Fernández et al. (2004) também apresentam uma estratégia para a criação automática de *corp*us baseada na formação de sentenças de busca a partir das definições e relações da WordNet e na busca de exemplos com essas sentenças em *corp*us ou na *web*. Cada *synset* a que pertence uma palavra na WordNet é caracterizado, por meio de suas relações com outros *synsets* ou palavras, como uma potencial sentença de busca. Contudo, os critérios para a construção das sentenças de busca são mais elaborados e flexíveis. Na abordagem de Agirre & Martínez, a estrutura das sentenças de busca é fixa, definida previamente. Por exemplo, ela é constituída sempre do contexto da palavra-alvo e de mais um sinônimo não ambíguo dessa palavra. Fernández et al., por outro lado, definem uma linguagem para especificação de padrões de sentenças de busca, de modo que várias estratégias de busca possam ser previamente definidas para formar diferentes sentenças para a busca nos *corp*us. Com isso, a abordagem se torna mais flexível e as buscas podem retornar um número muito maior de exemplos. Em um experimento com o *corp*us do SEMCOR, foram criadas seis estratégias de busca e essas estratégias foram aplicadas às 73 palavras ambíguas usadas no SENSEVAL-2. As sentenças de busca geradas foram então utilizadas para recuperar exemplos no SEMCOR. Como cada estratégia envolve um possível sentido da palavra ambígua e as sentenças de busca mantêm esse sentido, os exemplos recuperados já possuem, automaticamente, uma etiqueta de sentido. Para todas

as palavras, as sentenças de busca de todas as estratégias recuperaram, em conjunto, 48.980 exemplos (não necessariamente todos corretos de acordo com sentido buscado). Esse pode ser considerado um número alto, já que o SEMCOR é um córpus relativamente pequeno.

### **3.3 Considerações finais**

Como pode ser verificado pelos trabalhos descritos nessa seção, não se dispõem de córpus de exemplos anotados para a criação de abordagens baseadas em córpus supervisionadas de DLS para a TA. Tanto os córpus criados manualmente quanto os criados a partir de abordagens automáticas são voltados para aplicações monolíngües, muito embora se utilizem de córpus paralelos bilíngües. Novamente, considerando-se a TA para a língua portuguesa, não se tem conhecimento de nenhum trabalho prévio envolvendo a criação manual ou automática de córpus de exemplos.

## Capítulo 4 - As estratégias de pré-processamento desenvolvidas

Neste capítulo são descritas as estratégias desenvolvidas para a criação automática de um corpus de exemplos para a DLS na TA inglês-português. Elas são denominadas, aqui, “estratégias de pré-processamento”, considerando-se que a criação de corpus constitui uma etapa prévia (portanto, de pré-processamento) ao processo de aprendizado para a criação de um módulo de DLS, contexto deste trabalho. Antes das estratégias, propriamente ditas (Seção 4.4), são apresentados o conjunto de palavras consideradas e a motivação para a sua escolha (Seção 4.1); os corpus paralelos que serviram de base para as estratégias (Seção 4.2); e o conjunto de sentenças resultante do processo desse corpus (Seção 4.3).

### 4.1 Conjunto de palavras

Os problemas causados pela ambigüidade lexical de sentido na tradução envolvendo o português do Brasil foram recentemente analisados em alguns estudos experimentais. Um desses estudos consistiu da realização de um experimento com o corpus BNC (*British National Corpus*) (Burnard, 2000) com o objetivo de investigar as conseqüências da ambigüidade lexical de sentido em traduções automáticas de textos reais, a fim de delimitar a proposta de um modelo de DLS aos casos mais problemáticos de ambigüidade (Specia & Nunes, 2004). Nesse sentido, tal estudo serviu para delimitar o escopo deste trabalho, definindo um conjunto de palavras a ser manipulado.

Esta atividade foi desempenhada com base em três sistemas de TA inglês-português comumente utilizados, a saber, Systran, FreeTranslation e Globalink Power Translator Pro. Foram considerados para análise somente os verbos das sentenças, inicialmente, o subconjunto dos 15 verbos mais freqüentes do BNC. Essa categoria gramatical foi escolhida porque os verbos são altamente ambíguos e porque da sua desambiguação pode depender a desambiguação de outras palavras da sentença, principalmente dos seus argumentos.

Para a análise, 531 sentenças do BNC contendo os 15 verbos foram aleatoriamente selecionadas e submetidas aos tradutores. As traduções foram, então, manualmente analisadas para verificar a ocorrência da ambigüidade lexical de sentido, seus efeitos na tradução das sentenças e o comportamento dos sistemas diante desse fenômeno.

Nesse estudo foram definidos critérios específicos para identificação de um subconjunto de verbos mais problemáticos com relação à ocorrência de ambigüidade lexical de sentido e à ineficiência no tratamento dispensado a ela pelos sistemas de TA. Com base nesses critérios, foram selecionados sete verbos: *to go*, *to get*, *to make*, *to take*, *to come*, *to look* e *to give*. Alguns exemplos de casos de ambigüidade lexical de sentido encontrados no uso desses verbos e não manipulados adequadamente pelos tradutores avaliados são ilustrados na Tabela 1.

Sentença	Tradução correta	TA		
		Systran	Free-Translation	Power Translator
The war may well just <b>go</b> on and on.	continuar	ir	vai	ir
Stand in a French village when the Tour de France <b>goes</b> by and you are participating in an event which is unambiguously French.	passa (passar)	vai	vai	passa
It's best to be alone when the noises <b>get</b> this loud.	ficam (ficar)	recebem	começam	adquirem
A lot of international help will be needed to <b>get</b> things moving.	fazer	receber	começar	adquirir
They <b>take</b> more foreign holidays.	têm (ter)	tomam	fazem exame	levam
" <b>Take</b> that money out of your mouth!" said her mother.	tire (tirar)	toma ... fora	faça exame ... fora	objeto pegado ... fora
Now eat your supper, both o' ye, afore it <b>takes</b> cold.	fique (ficar)	toma	faz exame	leva
"This city has suddenly <b>come</b> alive," said her husband, an off-duty border guard.	renasceu (renascer)	veio vivo	vivo ... vindo	veio viva
"Yes, I'm <b>coming</b> , but I've one or two things to attend to first," she explained.	indo (ir)	venho	vindo	vindo
Mr Gonzalez has also <b>come</b> in for criticism from within his own party.	recebeu (receber)	entrou	entrou	entrou

Tabela 1. Exemplos de sentenças do BNC com verbos problemáticos

## 4.2 Córpus originais

Os córpus originais, ou seja, os textos paralelos alinhados em inglês e português, sem quaisquer outras marcações, contendo sentenças com os verbos em questão, foram

extraídos de duas origens: do *cópus Compara* (Frankenberg-Garcia & Santos, 2003) e do *cópus Europarl* (Koehn, 2002).

O *cópus Compara* compreende livros de ficção originalmente em português ou em inglês e suas traduções manualmente elaboradas, tanto do inglês para o português como vice-versa. As traduções foram realizadas ou digitalizadas de modo que a cada unidade (sentença ou grupo de sentenças) de uma língua correspondesse exatamente uma unidade na outra língua. Assim, os *cópus* já foram criados de maneira corretamente alinhada em nível de sentença. Embora o *cópus* contenha livros traduzidos do português do Brasil e também de Portugal, somente as traduções para o português do Brasil foram consideradas nessa pesquisa.

O *cópus Europarl* compreende sentenças em inglês e português de Portugal de textos extraídos do Parlamento Europeu. As sentenças em ambos idiomas foram automaticamente segmentadas e alinhadas, o que resultou em um *cópus* com várias sentenças incorretamente alinhadas. Para minimizar os efeitos de tais erros nas estratégias desenvolvidas neste trabalho, alguns casos identificados de alinhamentos incorretos foram corrigidos manualmente.

### **4.3 Seleção das sentenças**

A partir dos *cópus* originais, foram selecionadas para o desenvolvimento deste trabalho as sentenças que continham os sete verbos problemáticos escolhidos, citados anteriormente. Para a seleção utilizou-se dois concordanciadores: o primeiro foi o fornecido pelo projeto *Compara*<sup>4</sup> e o segundo foi desenvolvido neste trabalho especialmente para extrair as sentenças do *Europarl*.

Os números das sentenças para cada verbo em inglês (i) e português (p), bem como o número total de palavras é mostrado na Tabela 2. O número de sentenças em ambos idiomas é o mesmo, portanto, somente o número total (i e p) é exibido.

---

<sup>4</sup> <http://www.linguateca.pt/COMPARA/>

Verbo	Sentenças Compara (i e p)	Sentenças Europarl (i e p)	Total
go	2.000	46.848	48.848
get	1.662	15.542	17.204
make	1.590	94.426	96.016
take	1.530	84.480	86.010
come	1.688	28.748	30.436
look	1.474	15.734	17.208
give	1.108	49.946	51.054
Total	11.052	335.724	346.776
palavras i	133.712	6.228.239	6.361.951
palavras p	120.754	6.371.370	6.492.124

Tabela 2. Quantidades de sentenças e palavras

## 4.4 Identificação automática das traduções

### 4.4.1 Pré-processamento

Alguns procedimentos de pré-processamento foram realizados sobre as sentenças selecionadas do córpus original com o intuito de transformar o córpus em um formato apropriado e também de obter as informações necessárias para a identificação dos sentidos:

1. Tokenização das sentenças em ambos os idiomas;
2. Etiquetagem morfossintática das sentenças em ambos idiomas utilizando o MXPOST (Ratnaparkhi, 1996);
3. Lematização dos verbos e expressões incluindo verbos nas sentenças em português;
4. Anotação XML das sentenças em ambos idiomas utilizando o esquema XML de Hofland (Hofland, 1996).

As ferramentas utilizadas em todas as etapas já haviam sido desenvolvidas para outros propósitos e estavam disponíveis no repositório de ferramentas do NILC (Núcleo Interinstitucional de Lingüística Computacional)<sup>5</sup>. Cada verbo de cada um dos dois córpus foi tratado separadamente, ou seja, foram gerados 14 arquivos com os córpus pré-

<sup>5</sup> <http://www.nilc.icmc.usp.br/nilc/index.html>

processados referentes a cada verbo/cópus, anotados com as informações mencionadas. Um exemplo de par de sentenças resultantes das etapas de pré-processamento, para representar as sentenças da Figura 1, é ilustrado na Figura 2.

<p>“I’d rather do without whatever I came for.” “Prefiro sair sem o que for que tenha vindo comprar.”</p>
---

Figura 1. Exemplo de sentença paralela

<pre>&lt;s id=tagged_en/tagged_tokenized_compara_come_en..s17&gt;I_PRP would_MD rather_RB do_VBP without_IN whatever_WDT I_PRP came_VBD for_IN . &lt;/s&gt;  &lt;s id=lemmatized_pt/lemma_tagged_tokenized_compara_come_pt..s17&gt;Prefiro_NP sair VERB INF//main:sair sem PREP o ART que PRON for VERB FIN-FUT1/SUB/main:ser</pre>
---

Figura 2. Sentença paralela pré-processada

#### 4.4.2 Pré-supostos

Para a identificação automática das traduções a partir do cópus pré-processado, assume-se um cópus no formato apropriado, em que as sentenças estão corretamente segmentadas, em que suas palavras e símbolos (pontuações, por exemplo) estão corretamente marcadas com etiquetas gramaticais apropriadas. Assume-se, também, que os cópus inglês e português estão corretamente alinhados no nível de sentença e que esse alinhamento é corretamente indicado pela marcação XML.

Partindo-se desses princípios, algumas suposições foram consideradas para a identificação automática das traduções:

- Uma vez que para cada sentença (ou unidade) no cópus em inglês há uma única sentença (ou unidade) representando sua tradução em um cópus em português, a tradução correta para os verbos de cada sentença pode ser encontrada naquela sentença (ou unidade);
- Somente verbos em português são utilizados como tradução dos verbos em inglês, e não palavras de outras categorias gramaticais;

- Cada verbo do inglês possui um conjunto pré-definido de possíveis traduções, incluindo aquelas referentes a *phrasal verbs*, sendo que essas traduções podem ser definidas a partir das traduções fornecidas por dicionários bilíngües;
- *Phrasal verbs* possuem traduções específicas, que são usadas com mais frequência para traduzir o *phrasal verb* do que a tradução do verbo individualmente;
- Se há mais de uma possível tradução para um verbo do inglês na sentença em português, as traduções em posições mais similares que a posição do verbo original são provavelmente as mais corretas.

#### 4.4.3 Dicionários

Para definir o conjunto de possíveis traduções para cada verbo, foram consultados dois dicionários inglês-português, Houaiss (edição de 1982) e Collins Gem (edição de 2001), e um dicionário específico de *phrasal verbs* Michaelis (edição de 2003). Os números de possíveis traduções para cada verbo (incluindo o seu uso em *phrasal verbs*) é ilustrado na Tabela 3. A média de possíveis traduções para os verbos em inglês é de 157. Apenas as traduções com uma única palavra foram consideradas nesse estágio. Além disso, não foi considerado o uso do verbo em expressões complexas, como expressões idiomáticas. Ambos os casos devem ser tratados posteriormente, em um trabalho futuro.

Verbos	Quantidade de Traduções
take	271
go	140
get	165
make	197
come	162
give	111
look	54
<b>Média</b>	<b>157</b>

Tabela 3. Quantidades de traduções possíveis para os verbos

#### 4.4.4 Heurísticas

Com base no cópús pré-processado, nas suposições citadas e no dicionário criado para cada verbo, foram definidas algumas heurísticas para identificar as traduções de cada um dos sete verbos no cópús:

1. Identifica-se no cópús em inglês cada uma das ocorrências do verbo em questão na sentença, sua posição na sentença e se é um *phrasal verb* ou não. Cada uma das ocorrências do verbo é manipulada individualmente.
2. Para cada ocorrência identificada, busca-se por suas possíveis traduções na sentença em português correspondente, com base na sua lista de possíveis traduções. Essa busca é feita com base nas anotações existentes em cada palavra, neste caso, busca-se apenas pelos lemas dos verbos, de acordo com a marcação fornecida pelo etiquetador morfossintático e pelo lematizador.
3. Se o verbo identificado na sentença em inglês ocorrer em um *phrasal verb*, procura-se primeiramente pelas traduções específicas de tal *phrasal*. Caso nenhuma tradução seja encontrada, busca-se pelas traduções do verbo considerado individualmente.
4. Se o verbo identificado na sentença em inglês não ocorrer em um *phrasal verb*, procura-se apenas pelas traduções do verbo considerado individualmente.
5. Em ambos os casos (*phrasal* ou não *phrasal*), na busca pelas traduções, caso seja encontrada apenas uma tradução, ela é escolhida.
6. Em ambos os casos (*phrasal* ou não *phrasal*), na busca pelas traduções, caso seja encontrada mais de uma possível tradução, considera-se, como critério de escolha, a posição dos verbos em inglês e das suas possíveis traduções em português nas respectivas sentenças: é escolhida a tradução em uma posição mais similar à posição do verbo na sentença em inglês.
7. A tradução escolhida é então utilizada para anotar a sentença correspondente do cópús em inglês.

Por exemplo, para o par de sentenças do cópulo paralelo mostrado na Figura 1, considerando o verbo “come” (lema de “came”) na posição 8, o sistema identifica corretamente que a tradução para esse verbo é o verbo “vir” (lema de “vindo”), que está na posição 9. Vale observar que de acordo com o grupo de possíveis traduções para o verbo “come”, duas outras traduções poderiam ter sido selecionadas: “sair” (posição 2) e “ir” (lema do verbo “for”) (posição 6). No entanto, a heurística de similaridade de posições adotada evita erros dessa natureza na identificação. A saída do sistema para esse par de sentenças e para alguns outros pares relativos ao verbo “come” é ilustrada na Figura 3. As traduções identificadas (os seus lemas), bem como os verbos (ou *phrasal verbs*) sendo analisados, estão destacados.

```

s17#came_VBD#7#NRM#tenha=vindo_VERB_FIN-PRES/SUB/+ter/main:vir#I'd_NNP rather_RB
do_VBP without_IN whatever_WDT I_PRP came_VBD for_IN .

s18#come_VBN#33#NRM#tinha=sáido_VERB_FIN-PAST2/IND/+ter/main:sair#She_PRP said_VBD ,
«` It_PRP hasn't_VBZ any_DT paragraphs_NNS , why_WRB is_VBZ that_IN ? »_ and_CC I_PRP
explained_VBD that_IN I_PRP was_VBD out_RB of_IN practice_NN in_IN writing_VBG
paragraphs_NNS , I_PRP was_VBD used_VBN to_TO writing_VBG lines_NNS , speeches_NNS ,
so_IN my_PRP$ self-description_NN had_VBD come_VBN as_IN a_DT kind_NN of_IN
monologue_NN .

s19#comes_VBZ#7#PHR#me=vem_VERB_FIN-PRES/IND/main:vir#That's_IN the_DT sort_NN of_IN
thought_NN that_WDT comes_VBZ to_TO you_PRP in_IN the_DT middle_NN of_IN the_DT
night_NN .

s20#come_VBN#20#NRM#vindo_VERB_GER//main:vir#I_PRP was_VBD down_RB at_IN the_DT
Club_NNP the_DT other_JJ day_NN with_IN my_PRP$ physically-challenged_JJ peer_NN group_NN

```

Figura 3. Alguns resultados do sistema

Este sistema foi implementado em Java utilizando o pacote de desenvolvimento J2EE versão 1.4.2.6, as bibliotecas de entrada/saída para manipulação de arquivos e as bibliotecas padrão para manipulação de *strings*.

#### 4.4.5 Avaliação 1

Aplicado aos *corpus* paralelos, esse estudo foi capaz de determinar uma tradução para 87% dos verbos do *corpus* Compara e 70% dos verbos do Europarl. Estes valores podem ser considerados como uma medida de cobertura do sistema.

Como previsto, as traduções das ocorrências de alguns verbos não foram encontradas. Isto ocorreu, a princípio, por quatro motivos: (a) a lista de traduções possíveis está incompleta; (b) não se considera traduções de expressões e traduções realizadas com mais de uma palavra; (c) há alguns problemas com as ferramentas utilizadas no procedimento de pré-processamento; (d) há problemas nas sentenças nos *corpus* originais.

A lista de traduções possíveis pode estar incompleta devido aos dicionários utilizados (e dos dicionários, em geral). Em um estágio subsequente desse trabalho, procurou-se consultar mais dicionários, incluindo dicionários especializados de *phrasal verbs*, mas sabe-se que desenvolver uma lista completa de traduções é uma tarefa praticamente impossível, devido às constantes mudanças nas línguas naturais. Algumas traduções muito possivelmente ainda não serão cobertas por serem muito novas, raras ou gírias, ainda não incluídas nos dicionários.

Os problemas derivados das ferramentas de pré-processamento consistem da escolha incorreta de etiquetas por parte do etiquetador morfossintático ou da identificação incorreta dos lemas. De qualquer maneira, esses erros são raros.

As traduções constituídas por multi-palavras não foram consideradas por apresentarem uma complexidade muito grande, fora do escopo deste trabalho. Finalmente, sobre os problemas dos *corpus* paralelos, uma vez que não possuem traduções literais, há casos de omissão e adição de palavras, assim como de outras mudanças nas sentenças traduzidas. Além disso, no *corpus* Europarl existem diversos erros derivados do processo de segmentação e alinhamento automáticos que não foram manualmente corrigidos.

No que diz respeito à corretude do *corpus* etiquetado pelo sistema, para realizar uma avaliação apropriada, seria necessário analisar manualmente cada tradução identificada de acordo com a tradução da sentença em inglês na sentença paralela em português. Embora esse procedimento seja facilitado pela existência das traduções no *corpus* paralelo, ainda requer uma grande quantidade de tempo.

Por esta razão, efetuou-se primeiramente uma avaliação preliminar, considerando apenas uma amostra de córpus etiquetado. Com isso, pretendia-se ter uma idéia da acurácia do sistema: se a acurácia se mostrasse muito baixa, já indicaria que as heurísticas precisariam ser aperfeiçoadas, e muito tempo não teria sido despendido nesta avaliação. Caso contrário, se a acurácia se mostrasse satisfatória, uma avaliação mais pormenorizada seria realizada.

Para a avaliação preliminar, selecionou-se aleatoriamente 20 sentenças de cada um dos verbos em ambos os córpus (totalizando 280 sentenças). A medida de acurácia usada é a tradicionalmente utilizada em DLS (também chamada de “precisão”). Com ela, obtém-se a proporção das traduções corretamente identificados pelo sistema com relação ao total de traduções identificadas. Os resultados são mostrados na Tabela 4.

<b>Compara verbo</b>	<b>% correto</b>	<b>EPC verbo</b>	<b>% correto</b>
<i>go</i>	80	<i>go</i>	55
<i>get</i>	70	<i>get</i>	85
<i>make</i>	90	<i>make</i>	80
<i>take</i>	80	<i>take</i>	85
<i>come</i>	75	<i>come</i>	70
<i>look</i>	95	<i>look</i>	100
<i>give</i>	90	<i>give</i>	95
<b>Média</b>	82.9%	<b>Média</b>	81.4%

Tabela 4. Precisão do processo de etiquetação de sentido

Como se pode observar na Tabela 4, os resultados mostram que o conjunto de heurísticas implementados identifica, em média, a tradução correta de 82.9% dos verbos do córpus Compara e 81.4% dos verbos do Europarl. Os resultados para o Compara são levemente melhores, pois ocorrem menos problemas no paralelismo das sentenças.

Os erros de etiquetação em ambos os córpus são, em geral, conseqüências dos quatro problemas listados anteriormente. Se a lista de possíveis traduções estiver incompleta para um verbo e, com isso, não estiver incluída a tradução correta para uma dada ocorrência, mas houver uma das possíveis traduções para o verbo na sentença, contudo, referindo-se a um outro verbo, esta tradução é (erroneamente) adotada pelo sistema para o verbo em questão. O mesmo ocorre quando há uma expressão na sentença-origem (sem uma tradução correspondente). Alguns poucos erros são devidos a problemas com as ferramentas usadas no pré-processamento. Entretanto, a maioria dos erros na

identificação do sentido e etiquetagem dos verbos são devidos a características das sentenças paralelas originais. Primeiramente, elas não são traduções literais. Por exemplo, se um tradutor humano não traduz ou muda parte da sentença que contém o verbo, não será possível para o sistema identificar a tradução correta. Além disso, algumas sentenças de ambos os corpuses, principalmente do Europarl, apresentam um número muito grande de palavras (por exemplo, 179 palavras), dificultando ainda mais a identificação. Há também alguns problemas causados por erros (não corrigidos) no alinhamento das sentenças originais no Europarl, como mencionado.

É importante salientar que os verbos considerados nessa pesquisa são altamente ambíguos e de uso geral. Então, é inevitável que tenham possíveis traduções em comum com outros verbos. O alinhamento prévio sentencial certamente reduziu o número de possíveis traduções em cada ocorrência de um dos verbos. De qualquer maneira, após o alinhamento, cada sentença em português ainda apresentava, em média, 3 possíveis traduções para o verbo em inglês correspondente nas 280 sentenças analisadas (em alguns casos, até oito possíveis traduções foram encontradas na sentença).

A despeito dos problemas mencionados, a avaliação preliminar mostrou que os resultados são promissores, considerando que foram empregadas somente heurísticas simples. Como próximo passo, foram realizados alguns ajustes nos dicionários, de modo, por exemplo, a incluir algumas possíveis traduções identificadas como necessárias nessa avaliação, e partiu-se para uma avaliação mais abrangente, considerando-se mais sentenças.

#### **4.4.6 Avaliação 2**

Com o intuito de obter estatísticas mais significativas a respeito das heurísticas de identificação das traduções desenvolvidas, bem como para analisar mais profundamente as razões dos erros cometidos pelo sistema, realizou-se uma segunda avaliação, considerando agora 200 sentenças de cada verbo, todas do corpus Compara. Apenas esse corpus foi utilizado com o objetivo de minimizar a interferência dos problemas citados (como o alinhamento sentencial do Europarl) nos resultados da etiquetagem.

Nesta avaliação, as 1400 sentenças foram analisadas considerando-se os mesmos critérios que da primeira avaliação: abrangência (ou cobertura) e acurácia (ou precisão) do

sistema. Os números com relação à abrangência são ilustrados na Tabela 5 e Figura 4, enquanto os números com relação à precisão são ilustrados na Tabela 6 e Figura 5.

Verbo	% de traduções identificadas
<i>go</i>	89
<i>get</i>	95
<i>make</i>	84
<i>take</i>	86
<i>come</i>	85
<i>look</i>	51
<i>give</i>	67
<b>Média</b>	79.57%

Tabela 5. Cobertura do sistema

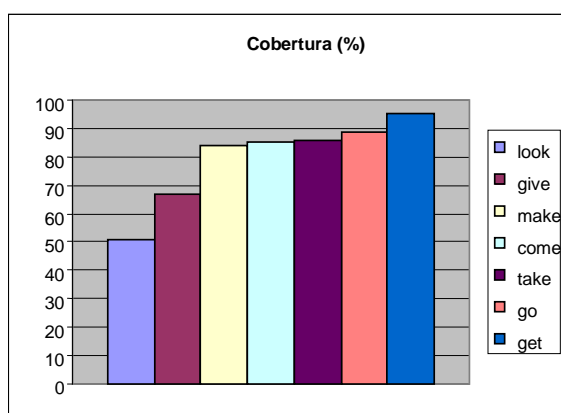


Figura 4. Cobertura do sistema

Como pode-se perceber, o sistema foi capaz de identificar alguma tradução (não necessariamente correta) para 79.57% das ocorrências dos sete verbos nas 1400 sentenças, em média. As razões para a não identificação de traduções em determinadas sentenças são as mesmas que as citadas na seção anterior. É importante ressaltar que apesar desse número não ser muito alto, o foco das heurísticas desenvolvidas está na acurácia do sistema, sendo a sua abrangência um fator secundário. Essa decisão foi tomada porque pretende-se utilizar, conforme mencionado, os resultados da identificação das traduções como fonte de informação para a criação automática de um *corpus* de exemplos para a

DLS baseada em *cópus*. Assim, é de extrema importância que os resultados estejam corretos. As traduções não identificadas podem, posteriormente, ser manualmente indicadas.

Verbo	% de traduções corretamente identificadas	% baseline (tradução mais freqüente no <i>cópus</i> )
go	87	17
get	75	14
make	68	56
take	84	19
come	78	43
look	89	47
give	91	78
Média	81.7%	39.1%

Tabela 6. Precisão do sistema

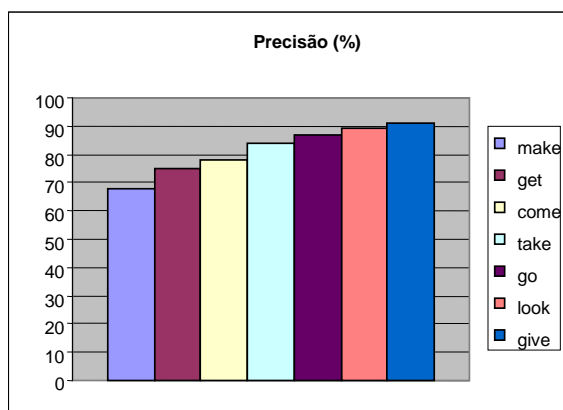


Figura 5. Precisão do sistema

Como indicado na segunda coluna da Tabela 6, a precisão do sistema na identificação das traduções é similar à verificada na primeira avaliação: média de 81.7% para todos os verbos. Na terceira coluna da Tabela 6 é indicada a precisão que seria obtida caso o sistema fosse capaz de identificar a tradução mais usada para cada verbo nas sentenças em português e simplesmente indicasse, como tradução para cada ocorrência de um verbo do inglês, a sua tradução mais freqüente. Essa precisão, chamada *baseline*, pode ser considerada uma precisão mínima esperada do sistema. É importante notar que, como ilustrado na Figura 6, o sistema ultrapassa a *baseline* para todos os verbos.

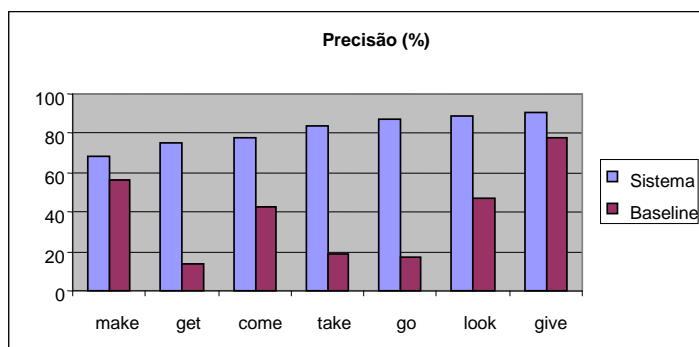


Figura 6. Precisão do sistema x baseline

Durante a verificação da precisão das traduções identificadas, foram sistematicamente analisadas as causas de todos os erros e estas foram então compiladas e organizadas em alguns grupos para cada verbo, visando uma futura correção das heurísticas. A Tabela 7 ilustra a lista das causas dos erros, bem como o número de ocorrências de cada tipo de erro decorrente dessa razão para cada verbo.

Causa / Verbo	Erros do concordanciador ou etiquetador	Dicionário incompleto	Ocorrência do verbo em expressões	Traduções modificadas	Deficiências das heurísticas
<i>go</i>	1	5	6	10	5
<i>get</i>	0	4	20	28	7
<i>make</i>	1	5	2	26	13
<i>take</i>	1	6	8	15	4
<i>come</i>	1	5	2	26	13
<i>look</i>	2	2	1	15	3
<i>give</i>	3	1	4	9	2

Tabela 7. Causas dos erros do sistema

Como pode-se perceber pela Tabela 7, as causas dos erros são similares às citadas na seção anterior, referentes à primeira avaliação. Incluiu-se, aqui, mais um tipo de causa, denominada “deficiências das heurísticas”, à qual se atribui todos os erros diretamente causados por decisões das heurísticas. Por exemplo, a escolha incorreta entre duas possíveis traduções derivada da preferência pela tradução com a posição mais próxima à do verbo na sentença em inglês. Pela tabela, fica evidente, novamente, que a maior parte

dos erros se deve a características do córpus originais, ou seja, ao uso do verbo em expressões sem traduções literais ou mesmo equivalentes para o português; e à modificação, pelo tradutor humano, das traduções, alterando, incluindo ou excluindo palavras.

Como uma análise adicional dos resultados, procurou-se observar a relação entre a precisão do sistema para cada verbo e o número de possíveis traduções de tal verbo. O gráfico de correlação é ilustrado na Figura 7. Segundo o gráfico, pode-se notar que não há relação direta com o número de possíveis traduções. Vale observar que há algumas diferenças com relação ao número de possíveis traduções ilustradas na Tabela 3, já que os dicionários foram alterados para essa segunda avaliação.

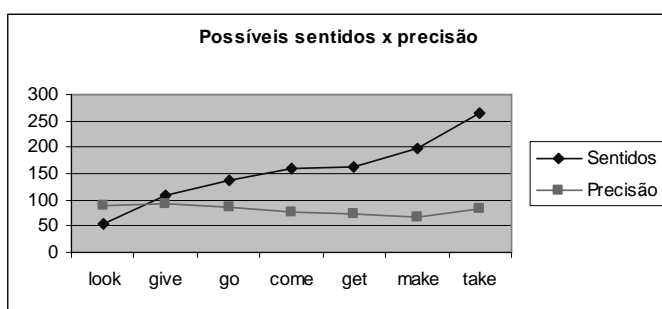


Figura 7. Relação entre o nº de possíveis sentidos de cada verbo e a precisão do sistema para tal verbo

Conforme mencionado, o córpus intermediário gerado pelo sistema, que corresponde à implementação da primeira das duas estratégias de pré-processamento propostas neste trabalho, deve ser usado como entrada para a segunda estratégia desenvolvida. Para tanto, durante a verificação da corretude das traduções, as traduções incorretamente identificadas foram também manualmente corrigidas. Assim, seria possível analisar mais facilmente os resultados da segunda estratégia.

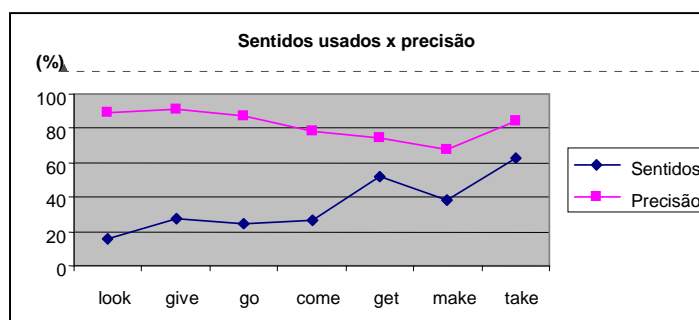
Com essa correção manual, outro resultado importante dessa avaliação é que ela permitiu verificar a grande variação do uso dos verbos em questão, mesmo no número relativamente pequeno de sentenças analisadas: média de 35.3. Os números para cada

verbo, considerando as traduções corrigidas, é ilustrado na Tabela 8 para cada um dos verbos.

Verbo	Número de traduções utilizadas
<i>go</i>	25
<i>get</i>	52
<i>make</i>	38
<i>take</i>	63
<i>come</i>	26
<i>look</i>	16
<i>give</i>	27
<b>Média</b>	35.3

Tabela 8. Diferentes traduções utilizadas nas 200 sentenças

Novamente, procurou-se analisar a relação entre a precisão do sistema e a variação no uso de cada verbo, ou seja, o seu número de diferentes traduções, mas agora considerando-se apenas as traduções efetivamente utilizadas no conjunto de 200 sentenças. O gráfico de correlação é ilustrado na Figura 8. Como na análise anterior, pode-se verificar que não há relação direta entre essas duas variáveis.



Formatted: Font: Bold

Figura 8. Relação entre o nº de sentidos usados de cada verbo e a precisão do sistema para tal verbo

Partindo-se do cópulus intermediário anotado com as traduções identificadas pelo sistema, o próximo passo foi a criação de um outro sistema para permitir a extração das informações necessárias desse cópulus, bem como representá-las de maneira adequada para os algoritmos de aprendizado de máquina.

## 4.5 Extração das características

Para a extração de características do *cópus* foi desenvolvido um sistema que, de forma automática e parametrizável, permite a escolha por uma característica ou a combinação de uma série de características, por meio de uma interface com o usuário, e gera um novo *cópus*, já no formato do ambiente de aprendizado de máquina Weka<sup>6</sup>. Esse ambiente possui vários algoritmos de aprendizado, sendo que todos utilizam o mesmo formato de entrada. Assim, podem ser realizados testes variados considerando os *cópus* gerados pelo sistema. Para a criação do sistema, primeiramente foi definido o conjunto de possíveis características a serem extraídas.

### 4.5.1 Possíveis características

A definição das possíveis características foi feita com base nas características usadas e sugeridas como relevantes por outros trabalhos de DLS, incluindo os citados no Capítulo 2. Essas características são:

- Contexto local (*narrow context*): cinco palavras de conteúdo à direita e à esquerda do verbo. Palavras de conteúdo são verbos, substantivos, adjetivos e advérbios. Para a sua seleção, foram considerados, portanto, somente as palavras (a partir da posição do verbo) com etiquetas gramaticais referentes a essas categorias de palavras, ou seja: JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ e WRB.
- Categoria gramatical do contexto local (*POS of the narrow context*): categorias gramaticais das 10 palavras consideradas como contexto local. Vale notar que, mesmo que as palavras não sejam escolhidas como características, as etiquetas podem ser utilizadas. As etiquetas foram diretamente extraídas das palavras constituindo o contexto local, já que estão anexadas a essas palavras no *cópus*.
- Contexto global (*broad context*): de 1 a 100 palavras (sem qualquer restrição) à esquerda e à direita do verbo. Para essa característica, palavras de todas as

---

<sup>6</sup> <http://www.cs.waikato.ac.nz/~ml/weka/arff.html>

classes gramaticais são consideradas. O número de palavra pode ser escolhido, mas o mesmo número será utilizado para os lados esquerdo e direito do verbo.

- Categoria gramatical do contexto global (*POS of the broad context*): categorias gramaticais das 1 a 100 palavras consideradas como contexto global. Vale notar que, mesmo que as palavras não sejam escolhidas como características, as etiquetas podem ser utilizadas. Novamente, o número de etiquetas será o mesmo para ambos os lados do verbo.
- 10 colocações definidas por Stevenson (2003): primeira e segunda palavras à esquerda e à direita do verbo, primeiro substantivo, primeiro adjetivo e primeiro à esquerda e à direita do verbo em questão. Colocações podem ser definidas como padrões de co-ocorrência entre as palavras no córpus com algumas características específicas (posição, categoria gramatical, etc.). Neste caso, são analisadas apenas as posições das palavras relativas ao verbo para a seleção das quatro primeiras colocações (primeira e segunda palavras à esquerda e à direita do verbo) e a posição e a etiqueta da palavra para as demais colocações, já que cada categoria possui um conjunto específico de etiquetas, a saber, substantivo: NN, NNP, NNPS ou NNS; adjetivo: JJ, JJR ou JJS; e verbo: VB, VBD, VBG, VBN, VBP ou VBZ.
- Fonte (*source corpus*): córpus de origem dos exemplos. Essa opção é definida pelo usuário durante a execução do sistema, e não extraída do córpus intermediário, produzido pelo primeiro sistema.
- Relações sintáticas (*syntactic relations*): palavras nas relações sintáticas de sujeito e objeto com o verbo. Diferentemente das demais informações, que são extraídas diretamente do córpus intermediário ou definidas pelo usuário, as relações sintáticas são geradas por um processo auxiliar, independente deste extrator de características, que utiliza a análise sintática previamente produzida por um *parser* e incorpora as informações referentes às relações sintáticas à saída do extrator de características.

Além dessas sete características, todas opcionais, outra característica é automaticamente incorporada ao córpis resultante: a tradução do verbo para a sentença em questão. Essa característica é obrigatória no aprendizado supervisionado, por isso não é apresentada como opção. Ela é extraída automaticamente a partir do córpis intermediário gerado.

#### **4.5.2 Interface com o usuário**

A interface do sistema, ilustrada na Figura 9, é composta basicamente por três partes:

**Entrada e saída de dados**, em que se define o arquivo de entrada (o córpis intermediário gerado pelo primeiro sistema), o arquivo de saída (arquivo no qual o córpis gerado pelo sistema será armazenado) e o verbo a ser analisado, para que seja criado o cabeçalho do córpis de saída com as informações referentes a tal verbo. As escolhas de arquivos de entrada e saída são realizadas por meio da interface padrão de janela de listagem de arquivos e diretórios.

**Definição de características e parâmetros**, na qual se define o nome do córpis que está sendo analisado (caso essa informação seja utilizada), as características que deverão ser extraídas do córpis e os parâmetros de algumas dessas características.

**Visualização da saída**, que exibe o mesmo conteúdo do arquivo gerado como saída, para uma conferência prévia dos resultados do sistema, durante a execução, sem a necessidade de abrir o arquivo gerado.

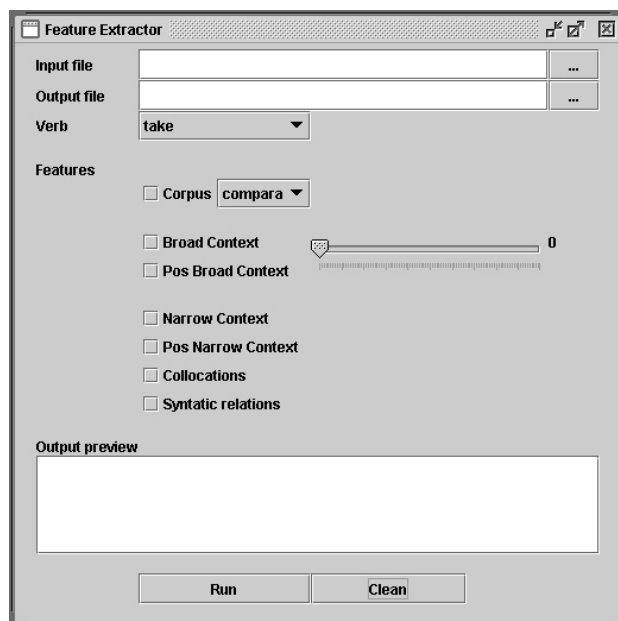


Figura 9. Interface do sistema extrator de características

A seção de definição de características e parâmetros permite a escolha entre os sete tipos de características citados, individualmente ou considerando quaisquer combinações entre eles. Além disso, permite a definição de alguns parâmetros: (1) caso seja marcada a opção para representar o nome do córpus, qual é o córpus sendo utilizado; e (2) qual o número de palavras a ser utilizado como contexto global (*broad context*) e, conseqüentemente, de etiquetas do contexto global, caso essas características sejam escolhidas.

O sistema foi implementado em Java utilizando o pacote de desenvolvimento J2EE versão 1.4.2.6, utilizando bibliotecas como a Swing para o desenvolvimento da interface, bibliotecas de entrada/saída para manipulação de arquivos e bibliotecas padrão para a manipulação de *strings*.

### 4.5.3 Exemplos de combinações testadas

Para realizar alguns testes com o extrator de características, foram escolhidas as seguintes combinações de características (além de características individuais) e parâmetros para serem extraídos pelo sistema:

- Narrow context
- POS of the narrow context
- Broad context: 5
- Broad context: 20
- Broad context: 50
- Broad context: 100
- POS of the broad context: 5
- POS of the broad context: 20
- POS of the broad context: 50
- POS of the broad context: 100
- Collocations
- Narrow context e POS of the narrow context
- Broad context: 5 e POS of the broad context: 5
- Broad context: 20 e POS of the broad context: 20
- Broad context: 50 e POS of the broad context: 50
- Broad context: 20 e POS of the broad context: 20 and Collocations
- Broad context: 50 e POS of the broad context: 50 and Collocations
- Narrow context e Syntactic relations
- POS of the narrow context e Syntactic relations
- Narrow context e POS of the narrow context e Syntactic relations
- Broad context: 5 e POS of the Broad context: 5 e Syntactic relations
- Broad context: 20 e POS of the Broad context: 20 e Syntactic relations
- Collocations e Syntactic relations
- Syntactic relations

Na Figura 10 são ilustradas as saídas parciais do sistema, incluindo o cabeçalho da base de dados (seções @relation e @attribute) e os valores dos atributos para os 10 primeiros exemplos (seção @data), para o verbo “look”, considerando-se duas características: *collocations* e *POS of the narrow context*.

Collocations
@relation look
@attribute 1_col string
@attribute 2_col string
@attribute 3_col string
@attribute 4_col string
@attribute 5_col string

<p>@attribute 6_col string  @attribute 7_col string  @attribute 8_col string  @attribute 9_col string  @attribute 10_col string  @attribute sense { admirar, aguardar, ansiar, antecipar, apressar, arregalar, assemelhar, assistir, atender, avaliar, buscar, confiar, considerar, consultar, contemplar, cuidar, demonstrar, desdenhar, desprezar, encarar, encarregar, esperar, estimar, examinar, examiner, fitar, folhear, hesitar, ignorar, indicar, inspecionar, investigar, lançar, lembrar, lidar, melhorar, menosprezar, mostrar, observar, olhar, parecer, pavonear, pensar, perdoar, pesquisar, prestar, procurar, progredir, proteger, recordar, recuperar, relembrar, repassar, respeitar, rever, ter, venerar, ver, visitar, voltar}  @data  at, to, us, over, caribbean, room, black, private, see, turn, olhar  like, hair, a, his, plump, hair, ?, blue, have, wearing, parecer  at, be, x-rays, nizar, x-rays, nizar, lighted, ?, hold, be, examiner  very, it, funny, think, ?, work, funny, ?, ?, think, parecer  through, like, the, image, porthole, image, powerful, circular, ?, was, olhar  like, what, a, be, silvery, ?, slim, ?, bite, be, parecer  like, what, a, with, brillo, chest, doormat-sized, ?, grow, cover, parecer  like, would, a, it, continuation, beard, ?, afraid, ?, be, parecer  like, without, a, ponytail, ponce, ponytail, real, enough, believe, wear, parecer  much, t, of, now, turn-on, university, erotic, ?, ?, call, parecer</p>
<b>POS of the narrow context</b>
<p>@RELATION look  @attribute sent string  @attribute 1_pawl {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 2_pawl {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 3_pawl {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 4_pawl {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 5_pawl {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 1_pawr {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 2_pawr {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 3_pawr {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 4_pawr {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute 5_pawr {JJ, JJR, JJS, NN, NNP, NNPS, NNS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ, WRB}  @attribute sense { admirar, aguardar, ansiar, antecipar, apressar, arregalar, assemelhar, assistir, atender, avaliar, buscar, confiar, considerar, consultar, contemplar, cuidar, demonstrar, desdenhar, desprezar, encarar, encarregar, esperar, estimar, examinar, examiner, fitar, folhear, hesitar, ignorar, indicar, inspecionar, investigar, lançar, lembrar, lidar, melhorar, menosprezar, mostrar, observar, olhar, parecer, pavonear, pensar, perdoar, pesquisar, prestar, procurar, progredir, proteger, recordar, recuperar, relembrar, repassar, respeitar, rever, ter, venerar, ver, visitar, voltar}  @DATA  s2, ?, ?, VBD, ?, WRB, ?, ?, ?, VBD, olhar  s4, NN, ?, ?, NN, ?, ?, ?, NN, NN, NN, parecer  s5, VBD, NNP, ?, ?, ?, NNS, ?, ?, NN, examinar  s6, ?, VBD, NN, ?, NN, RB, JJ, ?, ?, ?, parecer  s7, ?, NN, JJ, JJ, JJ, ?, ?, NN, ?, ?, olhar  s8, ?, VBD, VB, ?, ?, ?, JJ, NN, NN, parecer  s11, ?, ?, VBN, VBZ, NN, ?, ?, JJ, NNP, NN, parecer</p>

s13, ?, ?, JJ, VBD, ?, ?, ?, NN, ?, ?, parecer
s16, ?, NN, ?, ?, NN, ?, ?, NN, ?, NN, parecer
s17, ?, RB, ?, VB, ?, RB, ?, ?, JJ, NN, parecer

Figura 10.Exemplos parciais de córpus gerados pelo sistema

Como se pode verificar, além dos dados extraídos do córpus intermediário (ou definidos pelo usuário ou gerados pelo *parser*), o sistema gera também o cabeçalho para a configuração em questão, declarando a relação, os atributos (com nomes e tipos adequados) e demais informações necessárias para o Weka.

Como primeiro teste do desempenho do sistema, verificou-se se os arquivos gerados podiam ser devidamente utilizados como entrada para o ambiente Weka. Qualquer problema com o formato do arquivo ou inconsistências diversas, como diferença entre o número de atributos declarados no cabeçalho e o número de valores disponíveis para atributos, ou, ainda, incompatibilidades de tipos entre atributos declarados e seus valores, seriam prontamente acusados pelo ambiente. Para todas as características citadas, não houve problemas com a abertura no Weka dos arquivos gerados, indicando que não havia problemas com a formatação dos arquivos.

Além desse teste com o formato, os córpus de cada configuração foram manualmente analisados, a partir do córpus intermediário, para verificar a correteude dos dados gerados. Observou-se que o sistema teve êxito em 100% dos testes nos quais foi aplicado, ou seja, todas as configurações e parâmetros solicitados durante os testes, foram extraídos e representados de forma precisa.

## 4.6 Considerações finais

De modo geral, ambos os sistemas implementados, correspondentes às duas estratégias de pré-processamento investigadas e propostas, mostraram resultados bastante positivos. O primeiro sistema, de identificação das traduções, apesar da precisão estimada em cerca de 80%, é particularmente muito promissor, uma vez que tais resultados foram obtidos com heurísticas relativamente simples. Considera-se que algumas medidas de aperfeiçoamento nessas heurísticas poderiam melhorar significativamente a precisão do sistema.

O segundo sistema, apesar de realizar apenas a extração das informações já geradas pelo primeiro (ou por outros sistemas), mostrou-se bastante útil e eficaz para a geração de conjunto de atributos no formato padrão do Weka, amplamente aceito, facilitando assim a realização de experimentos diversos nesse ambiente.

Com ambos os sistemas, o processo de criação de cópús para utilização em métodos empíricos torna-se, de modo geral, muito mais viável, rápido e prático, permitindo manter o foco na geração do modelo de DLS, propriamente dita, e não nas etapas de pré-processamento relacionadas à criação de cópús de exemplos.

## Conclusão

Nesta pesquisa foram investigados trabalhos de DLS para a tradução automática e também de criação de córpus de exemplos para abordagens de DLS baseadas em córpus e híbridas. Considerando-se a inexistência de córpus de exemplos etiquetados com traduções para a TA inglês-português e a necessidade de córpus dessa natureza para o desenvolvimento de módulos de DLS abrangentes e precisos, foram propostas, implementadas e avaliadas algumas estratégias de pré-processamento para a criação desses córpus.

As palavras consideradas para essa pesquisa foram sete verbos frequentes, altamente ambíguos e problemáticos na tradução por parte dos sistemas disponíveis atualmente. Foram utilizados dois córpus paralelos inglês-português, contendo sentenças com esses sete verbos.

As estratégias foram implementadas na forma de dois sistemas de pré-processamento. O primeiro sistema é responsável por identificar, em um córpus paralelo alinhado por sentenças, a tradução correta de cada ocorrência do verbo em inglês e armazenar essa informação em um novo córpus, juntamente com outras informações relevantes para o processo de DLS (etiqueta gramatical, por exemplo). O segundo sistema gera um novo córpus, baseado no resultado do primeiro, extraindo e representando diversos tipos de informações escolhidas pelo usuário, de forma que possam ser utilizadas diretamente como entrada para o ambiente de aprendizado de máquina Weka.

Na avaliação dos sistemas e durante os vários testes no desenvolvimento desses sistemas, pode-se observar que os resultados obtidos são bons e, principalmente, bastante promissores, no sentido de que os sistemas podem ser facilmente estendidos e aperfeiçoados. Com isso, tais sistemas podem facilitar e viabilizar o desenvolvimento de modelos de DSL para a TA e, assim, contribuir para a criação de abordagens mais eficazes de TA que as disponíveis atualmente para a tradução do inglês para o português.

## Referências

- Agirre E., Martínez, D. (2004). Unsupervised WSD Based on Automatically Retrieved Examples: The Importance of Bias. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona.
- Brill, E. (1995). *Transformation Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging*. *Computational Linguistics*, 21(4), pp. 543-565.
- Brown, P.F.; Della Pietra, S.A.; Della Pietra, V.J.; Mercer, R.L. (1991). Word Sense Disambiguation Using Statistical Methods. In Proceedings of the 29th Annual Meeting of Association for Computational Linguistics, pp. 264-270. Berkley, CA.
- Bruce, R.; Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp.139-145. Las Cruces.
- Burnard, L. (2000). *Reference Guide for the British National Corpus (World Edition)*, Oxford University Press.
- Copeland, C.; Durand, J.; Krauwer, S.; Maegaard, B. (1991). The Eurotra Formal Specifications. *Studies in Machine Translation and Natural Language Processing*, 2. Commision of European Communities.
- Dagan, I.; Itai, A. (1994). Word Sense Disambiguation Using a Second Language Monolingual Corpus. *Computational Linguistics*, 20, pp. 563-596.
- Diab, M.; Resnik, P. (2002). An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, Philadelphia.
- Dihn, D.; Kiem, H.; Hovy, E. (2003). BTL: a Hybrid Model for English-Vietnamese Machine Translation. In *Proceedings of the MT Summit IX*, pp. 23-27. New Orleans.
- Dorr, J. B. (1993). *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge.
- Dorr, B.; Katsova, M. (1998). *Lexical Selection for Cross-Language Applications: Combining LCS with WordNet*. Maryland.
- Egedi, D. et al. (1994). *Korean to English Translation Using Synchronous TAGs*. Pennsylvania.
- Fernández, J.; Castilho, M.; Rigau, G.; Atserias, J.; Turmo, J. (2004). Automatic Acquisition of Sense Examples using ExRetriever. In Proceedings of the

- International Conference on Language Resources and Evaluation, pp. 25-28. Lisbon.
- Fossey, M. F.; Pedrolongo, T.; Martins, R. T.; Nunes, M. das G. V. (2004). Análise comparativa de tradutores automáticos inglês-português, Série de Relatórios do NILC, NILC-TR-04-04, São Carlos, Março, 18p.
- Francis, W. M.; Kucera, H. (1979). *Brown Corpus – Manual of Information*. Department of Linguistics, Brown University (<http://helmer.aksis.uib.no/icame/brown/bcm.html> [03/04/2004]).
- Frankenberg, A. Garcia; Santos, D. (2003). Introducing COMPARA: the Portuguese-English Parallel Corpus. *Corpora in translator education*, pages 71-87. Manchester.
- Gajek, O. (1991). The METAL system. *Communications of the ACM*, 34 (9), pp. 46-47.
- Goodman, K; Nirenburg, S. (1991). *The KBMT Project: A case study in Knowledge-Based Machine Translation*. Morgan Kaufmann Publishers, California.
- Hofland, K. (1996). A program for aligning English and Norwegian sentences. *Research in Humanities Computing*, pages 165-178. Oxford University Press, Oxford.
- Ide, N.; Erjavec, T.; Tufi, D. (2002). Sense Discrimination with Parallel Corpora. In *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pp. 54-60. Philadelphia.
- Koehn, P. (2002). Europarl: A Multilingual Corpus for Evaluation of Machine Translation. ([www.isi.edu/~koehn/publications/europarl](http://www.isi.edu/~koehn/publications/europarl)).
- Leacock, C.; Chodorow, M.; Miller, G.A. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24 (1), pp. 147-165.
- Leacock, C.; Towell, G.; Voorhees, E.M. (1993). Corpus-Based Statistical Sense Resolution. In *Proceedings of the ARPA Human Language Technology Workshop*, pp. 260-265. Morgan Kaufmann Publishers, San Francisco.
- Lee, Y.K.; Ng, H.T. (2002). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 41-48. Philadelphia.
- Leffa, V. J. (1998). *Textual constraints in L2 lexical disambiguation*, *System, Great Britain*, 26(2), p. 183-194.
- Miller, G.A.; Chorodow, M.; Landes, S.; Leacock, C; Thomas, R.G. (1994). Using a Semantic Concordancer for Sense Identification. In *Proceedings of the ARPA Human Language Technology Workshop - ACL*, pp. 240-243. Washington.

- Ng, H.T.; Lee, H.B. (1996). *Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-Based Approach*. In *Proceedings of the 34th Annual Meeting of Association for Computational Linguistics*, pp. 40-47. Somerset.
- Ng, H. T. (1997b). Getting Serious about Word Sense Disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, pp. 1-7. Washington.
- Oliveira Jr., O. N.; Marchi, A. R.; Martins, M. S.; Martins, R.T. (2000). *A Critical Analysis of the Performance of English-Portuguese-English MT Systems*, In: *Anais do V PROPOR*, Atibaia, p. 85-92.
- Pedersen, B. S. (1997). *Lexical ambiguity in machine translation: expressing regularities in the polysemy of Danish Motion Verbs*. PhD Thesis, Center for Sprogteknologi, Copenhagen, Denmark.
- Ratnaparkhi, A. (1996). A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in NLP Conference*. University of Pennsylvania.
- Specia, L.; Nunes, M. das G. V. (2004). O Problema da Ambigüidade Lexical de Sentido na Comunicação Multilíngüe. São Carlos. Relatório Técnico do NILC, NILC-TR-04-01.
- Stevenson, M. (2003). *Word Sense Disambiguation: The Case for Combining Knowledge Sources*. CSLI Publications, Stanford, CA.
- Witten, I.H.; Frank, E. (2000). *Data mining : practical machine learning tools and techniques with Java implementations*, Morgan Kaufmann
- Zinovjeva, N. (2000). *Learning Sense Disambiguation Rules for Machine Translation*. Master's Thesis in Language Engineering. Department of Linguistics, Uppsala University.