Regras de tradução automática induzidas de textos paralelos envolvendo o português do Brasil

Helena de Medeiros Caseli

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

A Tradução Automática – tradução de uma língua natural (fonte) para outra (alvo) por meio de programas de computador – é uma tarefa árdua devido, principalmente, à necessidade de um conhecimento lingüístico aprofundado das duas (ou mais) línguas envolvidas para a criação de recursos como gramáticas de tradução, léxicos bilíngües, etc. Nos últimos anos, diversos trabalhos têm surgido com o intuito de diminuir o esforço no desenvolvimento de recursos para a tradução automática por meio da extração automática de conhecimento a partir de córpus paralelos alinhados, um tipo de recurso lingüístico que vem se tornando cada vez mais disponível na *web*, atualmente. Assim, o projeto ReTraTos, apresentado neste documento, visa a indução de regras de tradução a partir de córpus paralelos alinhados sentencialmente usando técnicas de Aprendizado de Máquina e EBMT (*Example Based Machine Translation*). Este é o primeiro trabalho, nesta área, envolvendo o português do Brasil (PB) e as regras de tradução serão induzidas de córpus paralelos PB-inglês e PB-espanhol. As regras induzidas poderão ser usadas em um sistema de tradução automática indireta por transferência para traduzir sentenças na língua fonte para sentenças na língua alvo.

Abstract

Machine Translation – translation from one natural language (source) into another (target) by means of computer programs – is a hard task mainly due to the need of comprehensive linguistic knowledge concerning the two (or more) languages involved with which to create resources such as translation grammars, bilingual lexicons, etc. In the latest years, much work has been carried out with a focus on diminishing efforts in the development of machine translation resources by means of automatic knowledge extration from aligned parallel corpora, a kind of linguistic resource that is currently becoming more and more available on the web. Thus, project ReTraTos, presented in this document, aims at the induction of translation rules from sentence-aligned parallel corpora using Machine Learning and EBMT (Example Based Machine Translation) techniques. This is the first work in this area for Brazilian Portuguese (BP) and the translation rules will be induced from BP-English and BP-Spanish parallel corpora. The induced rules could be used in a transfer-based machine translation system to translate sentences from a source language into a target language.

Capítulo 1

Introdução

A Tradução Automática (TA) – tradução de uma língua natural (fonte) para outra (alvo) por meio de programas de computador – é uma tarefa árdua devido, principalmente, à necessidade de um conhecimento lingüístico aprofundado das duas (ou mais) línguas para a construção de recursos como gramáticas de tradução, léxicos bilíngües, etc. A escassez de recursos lingüísticos e mesmo a dificuldade em produzi-los, geralmente, são fatores limitantes na atuação dos sistemas de TA restringindo-os, por exemplo, quanto ao domínio de aplicação. Por outro lado, com o advento da *web*, torna-se cada vez maior a quantidade de informação disponível em diversas línguas e, como conseqüência, a criação de novas técnicas e recursos para sistemas de TA faz-se necessária.

Para lidar com esse desequilíbrio entre escassez de conhecimento lingüístico e abundância de informação multilingüe, diversos métodos de TA vêm sendo propostos com o intuito de gerar, automaticamente, conhecimento lingüístico a partir dos recursos multilingües existentes em abundância e, assim, tornar a construção de tradutores automáticos menos trabalhosa.

Nesse sentido, o projeto apresentado aqui surge como uma tentativa de superar esse desequilíbrio e para compreender melhor o contexto no qual ele está inserido é necessário, antes, fazer uma breve introdução sobre as estratégias e os paradigmas utilizados pelos sistemas de TA.

Os sistemas de TA podem ser classificados de acordo com três estratégias e três paradigmas. As estratégias caracterizam o projeto de processamento, enquanto os paradigmas indicam a utilização de determinados recursos nesse processamento. Porém, é importante citar que a utilização de uma estratégia não implica a utilização de um paradigma específico e vice-versa. Além disso, várias estratégias podem ser utilizadas em conjunto com o intuito de aprimorar os resultados operacionais.

Com relação às estratégias, a tradução de uma sentença na língua fonte para uma sentença na língua alvo pode ser obtida de modo direto (apenas substituindo-se as palavras fonte por palavras alvo) ou de modo indireto (por meio de uma forma de representação intermediária entre a língua-fonte e a língua-alvo). A tradução indireta admite, ainda, duas

variações: a abordagem de TA por transferência (sintática ou semântica) e a abordagem de TA por meio de uma interlíngua (uma representação independente das línguas fonte e alvo).

Apesar de ser menos vantajosa na teoria, a abordagem por transferência tem provado que o desenvolvimento de interfaces específicas entre a língua fonte e a língua alvo, embora exija formação bilíngüe por parte do desenvolvedor, é menos complexo (e conseqüentemente menos oneroso, mais rápido e mais factível) do que os módulos de projeção para a interlíngua, nos quais os sistemas de representação possuem uma natureza muito abstrata. Além disso, tem sido observado que esses mesmos módulos de transferência podem ser otimizados para serem reaproveitados, de alguma maneira, por novas línguas incorporadas ao sistema (Martins & Nunes, no prelo). Por esses e outros motivos, a estratégia de tradução por transferência foi escolhida para ser seguida no projeto apresentado aqui.

Os paradigmas de TA, por sua vez, estão relacionados aos recursos utilizados no processamento, assim, têm-se os paradigmas lingüístico, não-lingüístico e híbrido. As técnicas do paradigma lingüístico se baseiam na teoria lingüística para empregar restrições sintáticas, léxicas e semânticas na geração da sentença alvo correspondente a uma certa sentença fonte. As técnicas do paradigma não-lingüístico não se baseiam nas teorias lingüísticas nem nas propriedades lingüísticas das línguas fonte e alvo, mas, por outro lado, dependem da existência de grandes córpus de textos bilíngües para treinamento e/ou base de exemplos. As técnicas híbridas, por fim, são aquelas que incorporam recursos lingüísticos às técnicas não-lingüísticas (Dorr et al., 1999).

Desses paradigmas, o mais relevante para o projeto aqui apresentado é o não-lingüístico (ou tradução automática baseada em córpus), que engloba as abordagens estatística (*Statistical Machine Translation* ou SMT) e baseada em exemplos (*Example Based Machine Translation* ou EBMT). Enquanto as técnicas de SMT usam medidas estatísticas para escolher as palavras mais prováveis (na língua alvo) de formar a tradução da sentença na língua fonte (a probabilidade da tradução determina a tradução), as técnicas de EBMT empregam reconhecimento de padrão para traduzir partes da sentença fonte fornecida e, assim, determinar a tradução.

Dessas abordagens, a que será mais explorada neste projeto é a EBMT. Algumas vantagens da utilização de EBMT, citadas em (Somers, 1999) e de especial importância para este projeto, são:

- os exemplos são dados reais da língua e, portanto, o uso desses exemplos leva a sistemas que cobrem as construções que realmente ocorrem e ignoram as outras que não ocorrem reduzindo, assim, a super-geração;

- o conhecimento lingüístico do sistema pode ser mais facilmente enriquecido, simplesmente adicionando-se mais exemplos;
- os sistemas de EBMT são dirigidos aos dados e não à teoria e, uma vez que não há gramáticas complexas desenvolvidas por uma equipe de lingüísticas, o problema de conflito de regra e a necessidade de ter uma visão geral da teoria e de como as regras interagem é menor;
- dependendo do modo como os exemplos são usados é possível que um sistema de EBMT para um novo par de línguas seja rapidamente desenvolvido com base em (apenas) um córpus paralelo alinhado.

Assim, um critério fundamental para a utilização de técnicas de EBMT é a existência de um córpus paralelo alinhado – um conjunto de exemplos (geralmente sentenças) escritos em uma língua fonte acompanhados de suas traduções na língua alvo. Segundo Somers (1999), para a construção desse córpus existem, basicamente, três pontos a serem definidos: tamanho do córpus, adequação dos exemplos à tarefa proposta e modo de representação desses exemplos.

O tamanho do córpus é influenciado significativamente pelo modo como os exemplos são armazenados e usados. Em (Somers, 1999) são apresentados tamanhos de córpus que variam de 7 a mais de 700.000 exemplos para diversos sistemas de TA ou sistemas que usam os exemplos como parte do processo de tradução, por exemplo, para extrair regras de tradução. Uma explicação para essa grande variedade de tamanhos é que os sistemas de indução de regras de tradução possuem a capacidade de generalizar os exemplos permitindo que bons resultados sejam obtidos com um córpus menor. Além disso, constatou-se que há um limite para a relação de que quanto maior o número de exemplos, melhor a qualidade do produto (sentença traduzida) gerado pelo sistema de TA.

Enquanto o tamanho ideal de um córpus para sistemas de EBMT é, ainda, uma questão em aberto, a adequação de exemplos reais a uma dada tarefa tem um agravante relacionado à natureza da língua natural: alguns exemplos podem ser redundantes e outros conflitantes. Por fim, existem diversos meios de se representar (ou armazenar) os exemplos como: estruturas de árvores anotadas; exemplos generalizados que vão desde exemplos literais até regras de tradução tradicionais, passando por uma representação intermediária (padrões de tradução) na qual as palavras são substituídas por variáveis; ou modelos estatísticos formados por parâmetros estatísticos pré-computados que expressam as probabilidades das correspondências bilíngües.

Embora a utilidade de exemplos de sentenças paralelas alinhadas seja inegavelmente grande, informações sobre as estruturas dessas sentenças e as correspondências existentes entre elas são, sem dúvida, muito mais relevantes para pesquisas em língua natural (Matsumoto et al., 1993).

Nesse contexto, nos últimos anos, vários métodos têm sido propostos para extrair, de forma automática, as correspondências estruturais, sintáticas e/ou lexicais dos textos alinhados e generalizá-las, quando possível, para tornar o recurso gerado ainda mais abrangente e valioso. Essas correspondências recebem o nome de regras de tradução (ou de transferência) e são usadas, em sistemas de tradução automática, para traduzir (transferir) a representação de uma sentença na língua fonte em uma representação correspondente na língua alvo (Boström, 2000).

1.1 Motivação

Tendo em vista o cenário descrito anteriormente, o projeto apresentado aqui surge como uma alternativa para o processo árduo de construção de tradutores, uma vez que propõe a indução de regras de tradução a partir de córpus paralelos alinhados sentencialmente empregando métodos empíricos para minimizar os custos de desenvolvimento.

Além dessa, outra motivação importante deste projeto está relacionada aos avanços nos estudos de TA para o português, incipiente no Brasil (e também em Portugal) mesmo frente à demanda enorme por sistemas desse tipo; contrapondo, assim, a escassez de trabalhos acadêmicos (e talvez comerciais) desenvolvidos exclusivamente para o Português do Brasil (PB). Vale ressaltar, aqui, que a TA envolvendo o PB tem ganhado força em pesquisa apenas recentemente, quando projetos mais ambiciosos como o da UNL[1] e o EPT-Web[2] – ambos sistemas que adotam a tradução por interlíngua – se propuseram a levar a cabo a tradução ao nível de um processo completo e robusto. No entanto, seus resultados são, ainda, preliminares e não garantem melhores desempenhos quando comparados aos sistemas comerciais.

Algumas análises dos sistemas de TA existentes, hoje, para o PB são apresentadas a seguir. Em (Oliveira Jr. et al., 2000), seis tradutores automáticos inglês-português-inglês[3] foram analisados na tradução de 20 passagens de texto (com uma ou mais sentenças) do jornal

[1] http://www.nilc.icmc.usp.br/nilc/projects/unl.htm (16/08/2004).
[2] http://www.nilc.icmc.usp.br/nilc/projects/ept-web.htm (16/08/2004).
[3] Os sistemas analisados em (Oliveira Jr. et al., 2000) foram: Translator Pro, Alta Vista, Intertran, GO Translator, Tradunet e Enterprise Translator Server.

brasileiro "Folha de São Paulo" e do jornal norte-americano "*The New York Times*" constatando-se que menos de 50% das saídas geradas pelos sistemas poderiam ser consideradas inteligíveis. Além disso, percebeu-se que essas deficiências não motivaram os desenvolvedores das ferramentas a procurar estratégias alternativas para superá-las, uma vez que os níveis de desempenho se mantêm, quase sempre, os mesmos.

Além dessa análise de desempenho dos sistemas, foi realizado também um levantamento dos principais problemas encontrados nos níveis lexical, sintático e semântico-pragmático. Desses, os dois primeiros são de fundamental importância para este trabalho e, por isso, são apresentados em detalhes a seguir.

No nível lexical os problemas identificados foram: dicionarização das palavras, homônimos, conotações e expressões idiomáticas. Em relação à dicionarização, constatou-se que os problemas mais freqüentes estavam relacionados a nomes próprios e palavras derivadas, como "*Hungary*" e "*Hungarian*". Quanto ao problema de palavras homônimas, bastante freqüente no português, constatou-se que apenas a dicionarização das palavras não foi suficiente para solucioná-lo sendo necessários recursos mais elaborados para desambigüação lexical de sentido. O terceiro problema refere-se ao uso conotativo de palavras em português que, por possuírem um contexto cultural muito específico, não podem ser transferidas de uma língua para outra de maneira direta. Um exemplo desse problema foi a tradução incorreta da expressão, em portugês, "pegar carona" (no sentido de "tirar proveito de") para "*to hitchhike*", em inglês, uma vez que o sentido, nesse caso, não é o literal. Por fim, os sistemas avaliados apresentaram muitos problemas na tradução de expressões idiomáticas (como "abrir mão de", "ao pé da letra" e os *phrasal verbs* do inglês) – nas quais o significado da expressão como um todo não pode ser obtido por meio da composição dos significados das palavras que a formam.

No nível sintático, foram identificados problemas como: concordância (artigo-substantivo ou substantivo-verbo), uso incorreto (tempos verbais, preposições, artigos, pronomes ou comparações) e ausência de algum componente (preposição, artigo, pronome reflexivo ou conjunção). Constatou-se, ainda, que alguns desses problemas poderiam ser solucionados com a existência de regras de geração para, por exemplo, garantir a concordância entre artigo e substantivo. Muitos outros problemas, no entanto, estão relacionados às diferenças sintáticas entre as duas línguas analisadas (português e inglês) para as quais as ferramentas de tradução não estão preparadas.

Por fim, os autores do estudo apontam três fatores principais para as deficiências encontradas: a ausência ou má qualidade dos recursos lingüísticos disponíveis; a suposição

errada de que há muita similaridade semântica (praticamente uma correspondência um-para-um) entre o português e o inglês desconsiderando-se que, em muitos casos, as estruturas semânticas são dependentes de contexto ou de cultura; e a dificuldade de geração de traduções naturais que preservem não apenas a informação da sentença, mas também a forma como esta informação é passada na língua alvo, uma vez que a forma é tão importante quanto o conteúdo propriamente dito.

Nesse sentido, vale ressaltar que o projeto apresentado aqui não visa resolver problemas de ambigüidade lexical de sentido nem construir um dicionário bilíngüe que contemple todas as palavras das línguas envolvidas. Além disso, o paradigma de EBMT se mostra muito adequado para a obtenção de regras de tradução (objetivo deste trabalho), uma vez que os exemplos são dados reais da língua e, por isso, representam as estruturas dependentes de contexto e de cultura presentes nas línguas fonte e alvo; além de preservarem a forma como a informação é transmitida na língua em questão.

Em uma outra análise do desempenho de sistemas de TA, apresentada em (Fossey et al., 2004), quatro sistemas[4] foram avaliados na tradução de 515 sentenças da primeira página do jornal "*The New York Times*" (em inglês) para o português. Nessa análise, as sentenças foram classificadas em três tipos: gramaticais corretas (sentenças que traduzem de uma forma aceitável o sentido da frase original), gramaticais incorretas (sentenças que obedecem regras gramaticais, mas não obedecem regras semânticas) e agramaticais (sentenças que não possuem nada que as identifique como uma sentença da língua portuguesa). Os resultados dessa análise mostraram que nenhum dos sistemas alcançou um número satisfatório de sentenças consideradas "gramaticais corretas", isso porque nas quatro ferramentas a somatória das sentenças "gramaticais incorretas" e "agramaticais" sempre ultrapassa 50% do número total de sentenças do córpus: Linguatec e-translation Server (66,8%), Intertran (85,9%), Systram (69,1%) e FreeTranslation (66%).

Com base nas duas avaliações de sistemas de TA apresentadas anteriormente e com o intuito de analisar mais profundamente os tipos de erros encontrados na tradução de/para o português, realizou-se uma nova análise do desempenho de sistemas de TA português-inglês-português – Systran[5] (ST), FreeTranslation[6] (FT) e TranslatorPro (TP) – e agora, também, português-espanhol-português – Universia[7] e AutomaticTrans[8]. O propósito dessa nova

[4] Em (Fossey et al., 2004), foram avaliados: Linguatec e-translation Server, Intertran, Systran e FreeTranslation.
[5] http://www.systransoft.com (16/08/2004).
[6] http://www.freetranslation.com (16/08/2004).
[7] http://tradutor.universia.net/pt/ (16/08/2004).
[8] https://www.automatictrans.es (16/08/2004).

análise era apontar as classes de problemas, principalmente no nível sintático, que necessitam de maior atenção por parte dos desenvolvedores dos sistemas de TA para esses idiomas.

Nesse experimento, 20 sentenças em PB e suas respectivas traduções para o inglês e o espanhol foram submetidas aos tradutores constatando-se que a maioria dos erros encontrados nas traduções, nos dois pares de línguas e nos dois sentidos, foram os causados pela tradução incorreta (ou a não tradução) de palavras (erro lexical) ou pelo uso incorreto (ou ausência) de preposições, artigos e tempos verbais como é apresentado na Tabela 1 (para o par PB-inglês), e na Tabela 2 (para o par PB-espanhol).

Tabela 1. Resumo dos erros encontrados no experimento realizado com sentenças PB-inglês

Sistema / Erro (%)	Português ? Inglês			Inglês? Português		
	ST	**FT**	**TP**	**ST**	**FT**	**TP**
Lexical	**27,0**	**32,8**	**23,6**	**51,5**	**32,6**	**32,5**
Uso	**51,1**	**52,3**	**54,7**	**29,1**	**18,1**	**19,1**
Preposições	28,6	45,0	33,3	38,5	48,0	50,0
Artigos	44,3	28,6	29,6	53,8	16,0	8,3
Tempos verbais	5,7	2,2	16,0	5,1	28,0	16,7
Ausência	**8,0**	**3,4**	**10,2**	**11,2**	**32,6**	**31,7**
Preposições	36,4	50,0	26,7	60,0	40,0	30,0
Artigos	54,5	33,3	60,0	33,3	53,3	65,0
Outros	**13,9**	**11,5**	**11,5**	**8,2**	**16,7**	**16,7**

De acordo com os dados da Tabela 1 é possível notar que todos os tradutores automáticos analisados apresentaram, no sentido português? inglês, mais de 50% de erro no uso, principalmente, de preposições (ST = 28,6%, FT = 45% e TP = 33,3%), artigos (ST = 44,3%, FT = 28,6% e TP = 29,6%) e tempos verbais (ST = 5,7%, FT = 2,2% e TP = 16%). No sentido inglês? português, a maior ocorrência de erro está na tradução incorreta (ou não tradução) de palavras, ou seja, erro do tipo lexical (ST = 51,5%, FT = 32,6% e TP = 32,5%), porém os erros de uso incorreto ou ausência de preposições, artigos e tempos verbais representam, juntos, mais de 40% do total, em todos os sistemas (ST = 40,3%, FT = 50,7% e TP = 50,8%).

Entre os outros tipos de erros encontrados (indicados na Tabela 1 e na Tabela 2 com a denominação "Outros") estão: ordem incorreta das palavras, concordância de gênero e número (entre substantivo e artigo, por exemplo), etc.

Tabela 2. Resumo dos erros encontrados no experimento realizado com sentenças PB-espanhol

Sistema / Erro (%)	Português? Espanhol		Espanhol? Português	
	Universia	**AutomaticTrans**	**Universia**	**AutomaticTrans**
Lexical	**19,1**	**21,4**	**34,8**	**19,5**
Uso	**38,1**	**40,5**	**39,1**	**50,0**
Preposições	41,7	52,9	61,1	50,0
Artigos	12,5	11,8	38,9	50,0
Tempos verbais	20,8	29,4	0	0
Ausência	**33,3**	**33,3**	**15,2**	**22,2**
Preposições	38,1	50,0	57,1	75,0
Artigos	61,9	50,0	42,9	25,0
Outros	**9,5**	**4,8**	**10,9**	**8,3**

Na análise dos tradutores para o par PB-espanhol, constatou-se que, no sentido português? espanhol, mais de 38% dos erros estão relacionados, principalmente, ao uso incorreto de preposições (Universia = 41,7% e AutomaticTrans = 52,9%), artigos (Universia = 12,5% e AutomaticTrans = 11,8%) e tempos verbais (Universia = 20,8% e AutomaticTrans = 29,4%). Além disso, os erros de ausência nesse sentido foram bastante freqüentes (mais de 33%), principalmente, no que diz respeito a preposições (Universia = 38,1% e AutomaticTrans = 50%) e artigos (Universia = 61,9% e AutomaticTrans = 50%).

No sentido espanhol? português, a porcentagem de erro de uso também é a maior (mais de 39%) em preposições (Universia = 61,1% e AutomaticTrans = 50%) e artigos (Universia = 38,9% e AutomaticTrans = 50%). Os erros de ausência, nesse sentido, são um pouco menores do que no sentido contrário, porém, ainda se mantém alto, especialmente, com preposições (Universia = 57,1% e AutomaticTrans = 75%) e artigos (Universia = 42,9% e AutomaticTrans = 25%).

Comparando-se a quantidade de erros, por sentença, nos pares PB-inglês e PB-espanhol é possível concluir que existem, aproximadamente e em média, 8 erros/sentença no sentido português? inglês; 7 no sentido inglês? português; 3 no sentido português? espanhol e 2 no sentido espanhol? português. Assim, o número de erros no par PB-inglês é maior (mais do que o dobro) do que no par PB-espanhol; o que pode ser facilmente justificado pela maior proximidade do português com o espanhol do que com o inglês.

Dessa maneira, além da grande motivação deste projeto relacionada ao baixo desempenho dos sistemas de TA, há ainda uma última motivação referente à utilização da abordagem de Aprendizado de Máquina (AM) juntamente com EBMT como uma tentativa de superar os problemas encontrados nos sistemas de TA atuais. A abordagem de AM tem sido adotada em diversas áreas de pesquisa, sobretudo em Processamento da Língua Natural

(PLN), onde técnicas simbólicas geralmente estão associadas a altos custos com trabalhos volumosos de especialistas nem sempre disponíveis. A pesquisa em tradução automática por AM envolvendo o PB é ainda inédita e essa é a principal motivação deste projeto.

1.2 Objetivos

Embora se saiba que a qualidade da TA comercial atual só foi atingida depois de anos de esforço na criação de regras de tradução codificadas a mão e que os sistemas cuja fonte primária de conhecimento de tradução é derivada de uma base de exemplos criada automaticamente não se mostraram capazes de igualar ou superar a qualidade dos sistemas comerciais (Richardson et al., 2001), é importante esclarecer que o projeto apresentado aqui não tem como objetivo gerar um sistema de tradução automática baseado nas regras de tradução induzidas. O objetivo deste projeto é estudar a indução de regras de tradução utilizando técnicas de AM e EBMT para atacar, em princípio, os pontos mais problemáticos encontrados nos sistemas atuais de TA.

Assim, conhecidas as principais deficiências dos tradutores automáticos disponíveis para o PB, o projeto tem como objetivo: identificar como, em que medida e a que custo a abordagem por AM contribui para minimizá-las. Para isso, pretende-se investigar as principais técnicas de indução de regras de tradução propostas na literatura; implementá-las, adaptá-las e avaliá-las para os pares PB-inglês e PB-espanhol em córpus específicos; produzir um sistema de recombinação das regras de tradução induzidas no qual elas serão aplicadas a uma sentença na língua fonte e uma saída correspondente na língua alvo será gerada possibilitando, assim, a comparação da saída desse sistema com a de outros já existentes para o PB; e, por fim, avaliar o custo e os benefícios da abordagem investigada e produzir diversos documentos (relatórios técnicos, artigos científicos, etc.) com o relato de todas as etapas deste projeto.

Assim, ao fim do projeto pretende-se obter recursos lingüístico-computacionais como: córpus de textos paralelos alinhados para os pares PB-inglês e PB-espanhol; conjuntos de regras de tradução para os pares PB-inglês e PB-espanhol no domínio dos córpus utilizados na extração; um sistema capaz de induzir novas regras a partir de novos textos paralelos, possivelmente de maneira incremental; um sistema de recombinação das regras induzidas para ser utilizado na avaliação da cobertura e precisão das mesmas; e diversos documentos.

1.3 Organização do Texto

Este texto está organizado como segue. No Capítulo 2 é apresentada uma contextualização da área de indução de regras de tradução na qual tem-se: a definição de uma regra de tradução (Seção 2.1); a descrição do processo de indução de regras de tradução e das principais técnicas empregadas em cada etapa desse processo (Seção 2.2); e a apresentação das metodologias utilizadas na avaliação das regras induzidas bem como os valores levantados, na literatura, em algumas avaliações dos métodos citados (Seção 2.3).

No Capítulo 3, o projeto ReTraTos (Regras de Tradução induzidas de Textos paralelos) é apresentado em mais detalhes com: a descrição dos recursos lingüísticos (Seção 3.1) e computacionais (Seção 3.2) disponíveis para o desenvolvimento deste projeto; algumas considerações sobre as técnicas de indução de regras de tradução (Seção 3.3) e as metodologias de avaliação das regras induzidas (Seção 3.4); e uma breve descrição do estágio que a doutoranda realizará na Universidade de Alicante (Espanha) como parte deste projeto de doutorado (Seção 3.5).

No Capítulo 4, são listadas (e situadas nos cronogramas) as atividades previstas para o desenvolvimento do projeto ReTraTos e do doutorado como um todo (Seção 4.1) bem como as atividades específicas do estágio no exterior (Seção 4.2).

Por fim, o Capítulo 5 traz algumas considerações finais.

Capítulo 2

Indução de regras de tradução

Como já mencionado no capítulo anterior, um sistema de TA requer uma grande quantidade de conhecimento de tradução – geralmente armazenado em dicionários bilíngües, bases de exemplos ou modelos estatísticos – de difícil construção e/ou manutenção. Contudo, na última década, diversas pesquisas têm se concentrado na aquisição automática desse conhecimento induzindo-o de córpus bilíngües. Nesse contexto estão inseridos os sistemas de indução de regras de tradução.

De modo geral, os sistemas de indução de regras de tradução e de TA baseada nas regras induzidas possuem a arquitetura mostrada na Figura 1 na qual a linha pontilhada indica que a utilização dos recursos lingüístico-computacionais (*parsers*, dicionários bilíngües, etiquetadores, etc.) é opcional.

Nessa arquitetura, um córpus bilíngüe alinhado, geralmente no nível sentencial, é fornecido como entrada para o módulo de indução. As regras de tradução geradas como saída, posteriormente, são utilizadas na geração das sentenças alvo correspondentes às sentenças fonte por meio de um módulo de recombinação (aplicação) dessas regras.

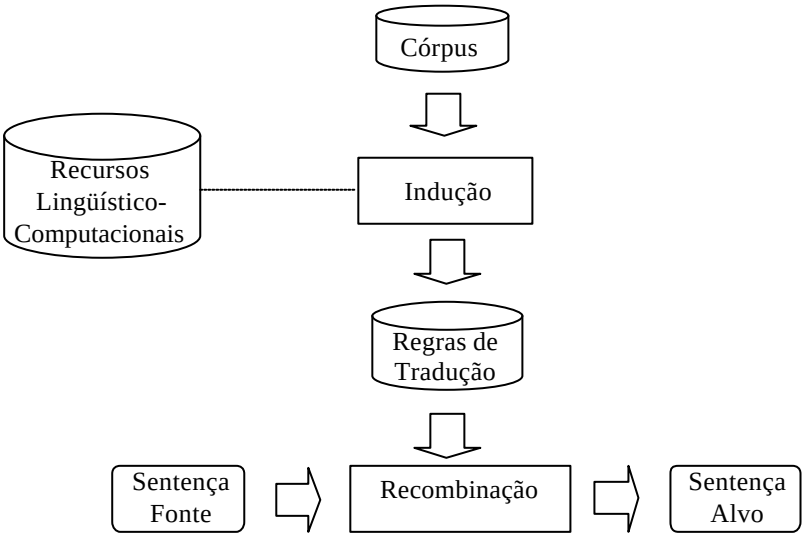

Figura 1. Arquitetura do sistema de indução de regras de tradução (McTait, 2003)

A parte variável dessa arquitetura está no módulo de indução de regras de tradução. Os sistemas de indução de regras de tradução propostos na literatura variam de acordo com diversos critérios. Um desses critérios é a utilização (ou não) de recursos lingüístico-

computacionais no processo de extração das gramáticas de tradução (como indicado pela linha pontilhada da Figura 1). Nos sistemas de EBMT "puros", a única fonte de conhecimento disponível para indução das regras é o par de textos paralelos alinhados; enquanto que em sistemas mais refinados, outros recursos lingüísticos são utilizados em menor ou maior grau. Os sistemas de EBMT também diferem quanto ao número e a qualidade dos recursos utilizados e quanto ao modo como esse conhecimento é representado, armazenado e usado para tradução (Carl, 2001).

Outro critério que diferencia os métodos de indução automática de regras de tradução é que enquanto alguns sistemas tratam as sentenças alinhadas como seqüências não-estruturadas de palavras, outros analisam sintaticamente as sentenças (*parsing*) antes de adquirir as regras de tradução (Meyers et al., 1998).

Embora existam diversos métodos de indução de regras de tradução propostos na literatura com diferentes abordagens, três etapas comuns do processo de indução podem ser identificadas na maioria dos métodos, sendo que apenas a primeira delas varia de acordo com a realização ou não de análise sintática em uma ou ambas as línguas. De modo geral, o processo de indução de regras de tradução a partir de textos paralelos alinhados sentencialmente pode ser dividido em: identificação de padrões (em sistemas que não realizam a análise sintática) ou alinhamento de árvores sintáticas (em sistemas que analisam sintaticamente as sentenças paralelas), geração das regras de tradução e filtragem e/ou ordenação das regras geradas.

A identificação dos padrões pode ser realizada, por exemplo, por meio de reconhecimento de seqüências repetidas de palavras em dois pares de exemplos (com técnicas de reconhecimento de padrão) ou por meio de correspondências lexicais existentes em um léxico bilíngüe (alinhamento lexical). O alinhamento das árvores sintáticas engloba o alinhamento dos nós com base em alinhamentos lexicais que foram extraídos de um léxico bilíngüe, gerados previamente (manual ou automaticamente) ou determinados estatisticamente durante o processo de alinhamento. Em seguida, os nós restantes são alinhados com base em regras pré-definidas, probabilidades de casamento de um nó fonte com um nó alvo, programação dinâmica, etc.

A segunda etapa – geração das regras de tradução – é realizada com base nos padrões ou alinhamentos definidos na etapa anterior. No caso dos padrões, estes são agrupados e generalizados (partes do padrão são substituídas por variáveis) considerando-se apenas a existência de similaridades, de diferenças ou de ambas (similaridades e diferenças). No caso dos alinhamentos, as regras podem ser geradas de diversas maneiras que variam de acordo

com o método e vão desde a extração de subpadrões das árvores até a expansão dos nós alinhados para a inserção de contexto.

A terceira e última etapa, presente em apenas alguns dos métodos estudados, engloba a filtragem das regras, por exemplo, para a eliminação de ambigüidades; e/ou a ordenação dessas regras de acordo com algum critério como freqüência de ocorrência, especificidade, etc.

Embora seja grande a variedade de técnicas empregadas pelos métodos de indução de regras de tradução, Menezes e Richardson (2001) apontam algumas características desejáveis para esses métodos:

- as regras devem ser induzidas com uma alta precisão;
- o método deve ser robusto em relação a erros introduzidos por recursos computacionais de análise sintática e de alinhamento sentencial/lexical, e a erros intrínsecos do córpus;
- as regras produzidas devem oferecer contexto suficiente para permitir que o sistema de TA que as utiliza escolha a melhor opção de tradução em um determinado momento.

A seguir, na Seção 2.1, são apresentados os diferentes tipos de exemplos de tradução – exemplos literais, padrões de tradução e regras de tradução – especificados na literatura, bem como o formalismo utilizado para representá-los. A próxima seção (2.2) apresenta as técnicas empregadas pelos principais métodos de indução de regras de tradução em cada uma das etapas do processo de indução. Em seguida, na Seção 2.3, tem-se uma visão geral das metodologias de avaliação empregadas, atualmente, para verificar a qualidade das regras induzidas. Por fim, na Seção 2.4, são apresentadas algumas considerações finais.

2.1 Regras de Tradução

No contexto de EBMT, os exemplos de tradução podem ser de três tipos diferentes segundo Furuse e Ida (1992). O primeiro tipo (1) consiste de exemplos literais, o segundo (2) – também conhecido como *translation template* ou padrão de tradução – consiste de um par de sentenças com palavras substituídas por variáveis, e o terceiro (3) são exemplos gramaticais ou regras de reescrita, na verdade, regras de transferência (ou tradução) do tipo encontrado em sistemas tradicionais de tradução automática baseada em regras.

I'm hungry ⇔ Eu estou com fome (1)

May I speak to X ⇔ Poderia falar com X (2)

N1 N2 for N3 ⇔ N2 de N1 para N3 (3)

N1=application/inscrição, N2=form/formulário, N3 = participation/participação

Como já mencionado no capítulo anterior, embora a utilidade de exemplos literais de sentenças paralelas alinhadas (tipo de exemplo de tradução apresentado em (1)) seja inegavelmente grande, informações sobre as estruturas das sentenças alinhadas e as correspondências existentes entre elas são, sem dúvida, muito mais relevantes para pesquisas em língua natural (Matsumoto et al., 1993). Por isso, diversos sistemas foram propostos nos últimos anos para indução de padrões ou regras de tradução (tipos (2) e (3) apresentados anteriormente).

Um padrão de tradução, segundo (McTait, 2003), pode ser definido formalmente como uma 4-tupla {S, T, A_f, A_v} com seqüências de fragmentos na língua fonte (em S) e na língua alvo (em T) separados por variáveis. Os alinhamentos entre os fragmentos e as variáveis são indicados em A_f e A_v, respectivamente. Em (4) tem-se um exemplo de um padrão de tradução no qual F_i corresponde a um fragmento de texto e V_j a uma variável.

$$F_1, V_1, F_2, V_2 ... F_p, V_p \Leftrightarrow F_1, V_1, F_2, V_2 ... F_q, V_q \quad (4)$$

Um exemplo de padrão de tradução com esse formato, para os pares de sentenças inglês-espanhol em (5), é apresentado em (6) (McTait & Trujillo, 1999):

1. The Commission gave the plan up ⇔ La Comisión abandonó el plan (5)
2. Our Government gave all laws up ⇔ Nuestro Gobierno abandonó todas las leyes

X_S gave Y_S up ⇔ X_T abandonó Y_T (6)

Nesse caso, *gave* e *up* são fragmentos na língua fonte que correspondem ao fragmento na língua alvo *abandonó*, ou seja, esses fragmentos estão alinhados e o alinhamento entre eles deve estar especificado em A_f. As variáveis também se alinham entre si, sendo X_S alinhada com X_T e Y_S alinhada com Y_T, como deve estar especificado em A_v.

Os padrões de tradução podem, ainda, conter informações morfossintáticas como os padrões apresentados em (8) gerados a partir dos pares de sentenças inglês-turco em (7) (Cicekli & Güvenir, 1996).

I give+PAST the book ⇔ kitap+ACC ver+PAST+1SG (7)
You give+PAST the pencil ⇔ kursun kalem+ACC ver+PAST+2SG

I ⇔ +1SG (8)
You ⇔ +2SG
X_S give+PAST the Y_S ⇔ Y_T+ACC ver+PAST X_T

As regras de tradução, por outro lado, podem ser compostas por informações mais complexas como as representadas no formalismo utilizado em (Lavie et al., 2004) para um método de indução de regras de tradução que realiza análise sintática. Uma regra de tradução com esse formato (veja exemplo na Figura 2 para o par de línguas inglês-hindi[9]) possui as seguintes informações:

- **Informação de tipo**: define o tipo de uma regra de tradução e, na maioria dos casos, corresponde ao tipo de um constituinte sintático. Por exemplo, as regras para sentenças são do tipo S, para sintagmas nominais (*noun phrases*) do tipo NP, etc.;
- **Informação morfossintática**: lista os componentes de uma regra (categorias lexicais, itens lexicais, etc.) tanto para a língua fonte quanto para a língua alvo;
- **Alinhamentos**: especificam como o conjunto de componentes na língua fonte se alinha com (transfere para) o conjunto de componentes na língua alvo. Além do tradicional alinhamento 1-1, alinhamentos do tipo n-0 e n-m (n, m > 1) também são possíveis;
- **Restrições do lado fonte**: fornecem informações sobre os atributos e seus respectivos valores na sentença da língua fonte. Essas restrições são usadas para restringir a aplicação de uma regra de tradução a uma dada sentença fonte de entrada;
- **Restrições do lado alvo**: são similares às restrições do lado fonte, mas em relação à língua alvo. Essas restrições são utilizadas para guiar e restringir a geração da sentença alvo correspondente à sentença fonte fornecida;
- **Restrições de ambos os lados**: informam quais valores deverão ser inseridos, na geração da sentença alvo, para substituir os valores presentes na sentença fonte.

[9] Uma língua falada na Índia.

Tal formalismo é capaz de lidar com uma variedade de divergências de tradução como: mudanças nas relações gramaticais em que, por exemplo, um objeto na língua fonte é expresso como sujeito na língua alvo; mudanças estruturais em que, por exemplo, um sintagma nominal se transforma em um sintagma preposicional em outra língua; etc. (Carbonell et al., 2002).

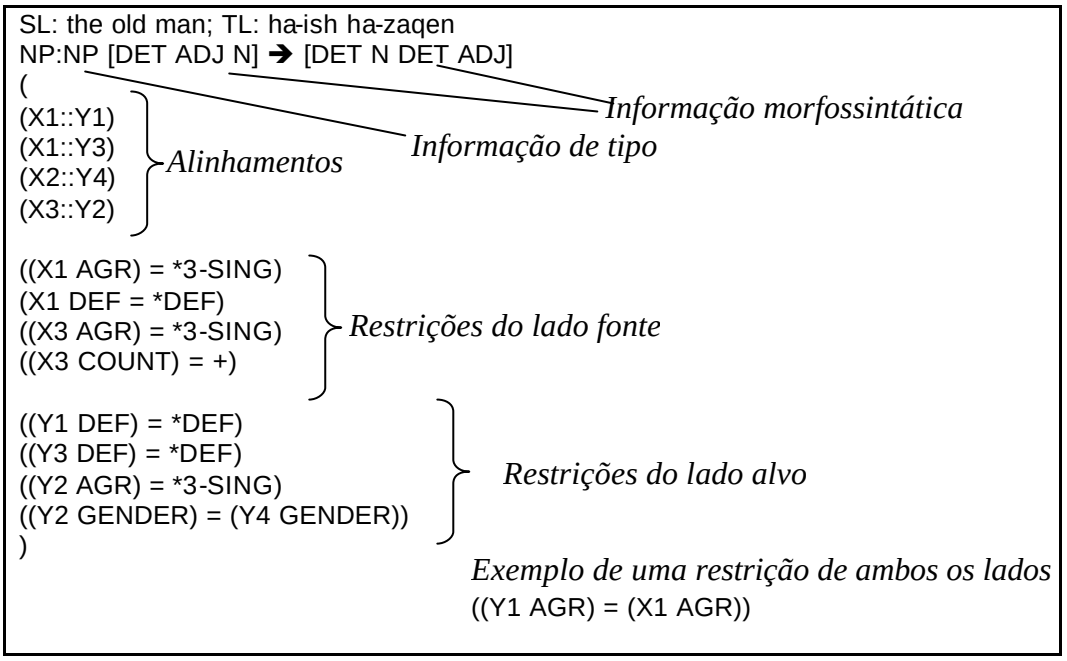

Figura 2. Exemplo um formalismo de representação de regra de tradução (Lavie et al., 2004)

Outro formalismo de representação de uma regra de tradução (agora para o par coreano-inglês), utilizado também por um método que realiza análise sintática, é apresentado na Figura 3. Esse formalismo engloba a noção de dependência sintática e identifica as variáveis pelo uso do caractere "$" prefixado.

```
(a)
        @KOREAN:
                {po} [class=vbma] (
                        s1 $X [ppca={reul}] )
        @ENGLISH:
                look [class=verb] (
                        attr at [class=preposition] (
                                ii $X ) )
        @-2xLOG_LIKELIHOOD: 12.77
```

```
(b)
        @KOREAN:
                $X [class=vbma ente={ra}]
        @ENGLISH:
                $X [class=verb mod=imp]
        @-2xLOG_LIKELIHOOD: 33.37
```

Figura 3. Outro exemplo de formalismo de representação de regra de tradução (Lavoie et al., 2002)

As regras da Figura 3 podem ser usadas para transferir a sentença em coreano *"Ci-To-Reul Po-Ra"* para a sentença em inglês *"Look at the map"*, sendo que a primeira (a) formaliza a lexicalização do predicado em inglês e a inserção da preposição correspondente, enquanto a segunda (b) define a inserção do atributo de imperativo. Cada regra, nesse formalismo, é acompanhada de seu valor de *log-likelihood* (Manning & Schutze, 1999) calculado com base nas sentenças do córpus de treinamento.

Contudo, em alguns trabalhos – como (Menezes & Richardson, 2001), (Meyers et al., 2000) e (Galley et al., 2004) – apresenta-se apenas a representação gráfica das regras como partes das árvores sintáticas geradas durante o processo de indução, como mostra a Figura 4.

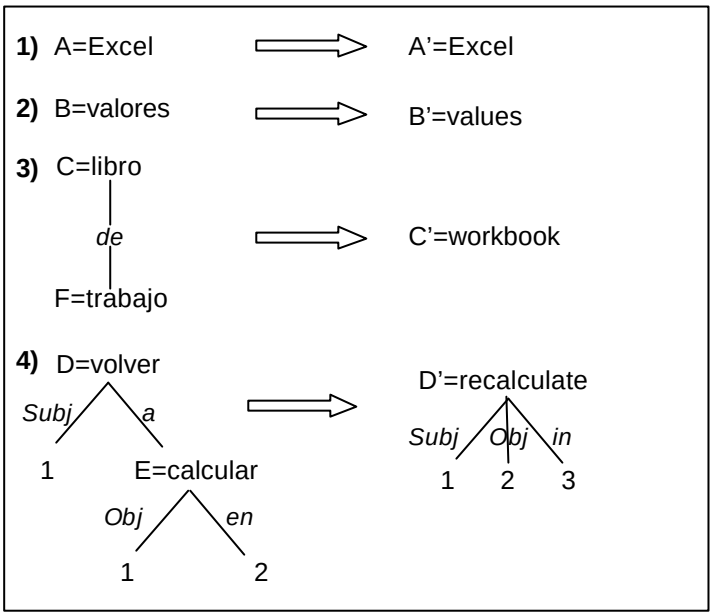

Figura 4. Exemplo de representação gráfica de regras de tradução (Menezes & Richardson, 2001)

Na Figura 4, as três primeiras regras (1, 2 e 3) podem ser usadas para preencher os nós numerados na regra 4 com as raízes das subárvores correspondentes gerando, assim, uma árvore completa usada para mapear a sentença fonte em espanhol *"Excel vuelve a calcular valores en libro de trabajo"* na sentença em inglês *"Excel recalculates values in workbook"*.

Tendo apresentado algumas definições de padrões e regras de tradução e seus respectivos formalismos de representação encontrados na literatura, de agora em diante e durante todo o projeto o termo "regra de tradução" será usado indistintamente para se referir tanto a regras quanto a padrões de tradução. Sendo assim, no contexto deste projeto, uma regra de tradução pode ser entendida como a generalização de sentenças que são traduções umas das outras, possuindo o seguinte formato:

$$A \Leftrightarrow B \qquad (9)$$

em que A é um conjunto de palavras e/ou variáveis derivadas do texto fonte (podendo conter todas as informações apresentadas na Figura 2 e até mesmo outras que se julgarem necessárias) e B, um conjunto semelhante derivado do texto alvo. O formalismo que será empregado para a representação das regras é, ainda, uma questão em aberto e será definido no decorrer deste projeto.

O símbolo ⇔ em (9) indica que as regras são bidirecionais, ou seja, as correspondências entre um conjunto de palavras e/ou variáveis na língua fonte (PB, neste projeto) e um conjunto na língua alvo (inglês ou espanhol, neste projeto) são válidas também no sentido inverso (da língua alvo para a língua fonte). Nem todos os métodos de indução de regras de tradução propostos na literatura consideram as regras como bidirecionais, mas esta é uma característica desejável e que será perseguida neste projeto.

2.2 Processo de indução de regras de tradução

A seguir são apresentadas as técnicas empregadas pelos principais métodos de indução de regras de tradução em cada uma das três etapas do processo de indução: identificação de padrões (Seção 2.2.1) ou alinhamento de árvores sintáticas (Seção 2.2.2), geração das regras de tradução (Seção 2.2.3) e filtragem e ordenação das regras geradas (Seção 2.2.4).

2.2.1 Identificação de padrões

Considere um exemplo de tradução E_a: $E_a^1 \Leftrightarrow E_a^2$ composto por um par de sentenças E_a^1 e E_a^2 que são traduções mútuas e estão escritas nas línguas L_1 e L_2, respectivamente. Na etapa de identificação de padrões, os métodos tentam identificar seqüências de palavras (padrões) em E_a buscando similaridades entre esse exemplo e outro E_b: $E_b^1 \Leftrightarrow E_b^2$ ou utilizando regras de transferência lexicais geradas previamente com base em um alinhamento lexical ou em um léxico bilíngüe.

Exemplos de métodos que se baseiam nas similaridades entre os exemplos de tradução para identificar os padrões são: (McTait & Trujillo, 1999), (McTait, 2003) e (Cicekli & Güvenir, 1996), (Güvenir & Cicekli, 1998), (Cicekli & Güvenir, 2003). Enquanto o método proposto por Brown (2001), é um exemplo de método que utiliza alinhamentos lexicais para identificar os padrões.

O algoritmo de Aprendizado de Máquina apresentado por McTait (2003) se baseia no princípio de distribuições similares de palavras (co-ocorrência e limites de freqüência). Esse

princípio pressupõe que palavras na língua fonte e na língua alvo que co-ocorrem em, pelo menos, dois pares de sentenças de um córpus bilíngüe são prováveis de serem traduções umas das outras. Dessa maneira, os itens lexicais (*tokens*) que ocorrem em pelo menos duas sentenças são armazenados juntamente com uma identificação das sentenças nas quais eles ocorrem e, em seguida, são geradas combinações para os itens lexicais recuperados (denominadas colocações).

Por exemplo, considerando-se o par de sentenças inglês-espanhol apresentado em (5) (veja Seção 2.1) os itens lexicais recuperados e as colocações geradas para esses itens lexicais são apresentados, respectivamente, em (10) e (11). As colocações são formadas quando há intersecção de pelo menos dois identificadores de sentenças nos itens recuperados, como é o caso de *gave up* em (10).

(gave) [1,2], (up) [1,2] (10)
(abandonó) [1,2]

(gave)(up) [1,2] (11)

Em (Cicekli & Güvenir, 1996), heurísticas semelhantes de similaridades e/ou diferenças entre as sentenças fonte e alvo são utilizadas: "dados dois pares de tradução, se as sentenças na língua fonte apresentam algumas similaridades, então as sentenças correspondentes na língua alvo devem possuir partes similares e estas devem ser as traduções das partes similares nas sentenças fonte; além disso, as partes diferentes restantes nas sentenças fonte devem também corresponder às diferenças nas sentenças alvo".

Para cada par de exemplos de tradução (E_a, E_b) tenta-se encontrar as similaridades entre os constituintes de E_a e E_b. Se nenhuma similaridade for encontrada, nenhum padrão será identificado; se houver similaridade, essas serão identificadas como apresentado em (12) para um par de exemplos de tradução inglês-turco apresentado, anteriormente, em (7) (veja Seção 2.1). Na próxima etapa do processo de indução são aplicadas as heurísticas de similaridades e diferenças nas seqüências de similaridades apresentadas em (12).

I give+PAST the book ⇔ kitap+ACC ver+PAST+1SG (12)
You give+PAST the pencil ⇔ kursun kalem+ACC ver+PAST+2SG

O método de indução proposto por Brown (2001) também se baseia no fato de que quando dois pares de sentenças no córpus têm alguma parte em comum, mas diferem em

alguma outra, as partes similares e diferentes correspondem a algum constituinte (sintagma nominal ou preposicional) coerente. Porém, diferentemente dos métodos apresentados até então, o algoritmo utiliza um léxico bilíngüe para identificar, inicialmente, os padrões de palavras em cada sentença.

Em seguida, com o córpus ordenado alfabeticamente pela sentença fonte, esse método inicia uma busca pelas seqüências de pares de sentenças que compartilham as *n* primeiras palavras na língua fonte; para cada seqüência encontrada cria-se um subcórpus. As sentenças em cada subcórpus são ordenadas alfabeticamente pela sentença fonte reversa e inicia-se uma busca pelas seqüências de pares de sentenças que compartilham as mesmas *m* últimas palavras na língua fonte. As partes diferentes e iguais em cada par de sentenças de uma seqüência são agrupadas em classes diferentes e o processo se repete até que nenhuma outra classe possa ser gerada.

2.2.2 Alinhamento de árvores sintáticas

Muitos dos métodos de indução de regras de tradução propostos na literatura realizam a análise sintática das sentenças nas línguas fonte e alvo (ou, às vezes, em apenas uma delas). Essa análise é efetuada de maneira automática por *parsers* específicos para os idiomas envolvidos (com ou sem treinamento prévio no domínio em questão) e pode ser seguida de uma verificação manual para a correção de possíveis erros. Com essa análise, os métodos de indução de regras de tradução podem obter, além das correspondências lexicais, regras estruturais.

A maioria dos métodos, após analisar sintaticamente os textos fonte e alvo, geram uma outra representação para as árvores. Menezes e Richardson (2001), por exemplo, extraem os lemas das palavras formando formas lógicas (FL) – grafos não ordenados que representam as relações entre os elementos mais significativos de uma sentença – como o apresentado na Figura 5. Nessa figura são exemplificadas FLs para um par de sentenças em espanhol (acima) e inglês (abaixo) nas quais os nós (em negrito) são identificados pelo lema de uma palavra lexical (*content word*) e os arcos direcionados e rotulados (em itálico) indicam as relações semânticas envolvidas. Algo semelhante é realizado por Meyers et alli (1996, 1998, 2000).

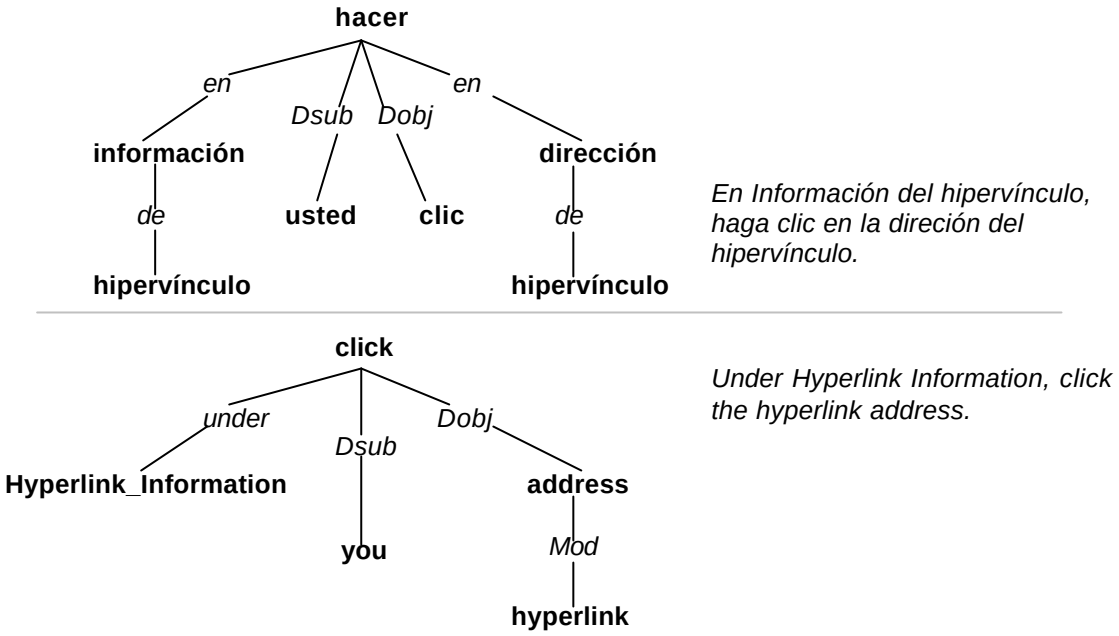

Figura 5. Formas lógicas para um par de sentenças espanhol-inglês (Menezes & Richardson, 2001)

Lavoie et alli (2001, 2002), por outro lado, convertem as árvores sintáticas fonte e alvo para uma representação de dependência sintática como mostra o exemplo da Figura 6 para o par de sentenças coreano (*Ci-To-Reul Ta-Si Po-Ra*) e inglês (*Look at the map again*).

```
(S1) {i} {Ci-To-Reul}              {Ta-Si} {Po-Ra}.
this + map-accusative + again + look-imp
(D1) {po}          [class=vbma ente={ra}] (
        s1 {ci-to} [class=nnin2 ppca={reul}] (
                s1 {i} [class=ande] )
        s1 {ta-si} [class=adco2] )
(S2) Look at the map again.
(D2) look [class=verb mod=imp] (
        attr at [class=preposition] (
                ii map [class=common_noun article=def] )
        attr again [class=adverb] )
```

Figura 6. Representação de dependência sintática para um par de sentenças coreano-inglês (Lavoie et al., 2002)

Conhecidos alguns dos principais formalismos de representação das árvores sintáticas, a primeira etapa dos métodos de indução de regras de tradução que realizam análise sintática é a de alinhamento das árvores geradas. Essa etapa, na verdade, pode ser subdividida em dois passos nos quais, primeiro, é realizado um alinhamento dos nós das árvores com base em alinhamentos lexicais extraídos de um léxico bilíngüe, gerados previamente (manual ou automaticamente) ou calculados com base em estatística. Em seguida, os nós restantes são alinhados considerando-se, por exemplo, regras de composição dos nós definidas previamente, probabilidades de casamento de um nó fonte com um nó alvo, programação dinâmica, etc.

Carl (2001) propôs um algoritmo para extração de gramáticas de tradução probabilísticas que tem como entrada um texto bilíngüe com *n* partes (sentenças) alinhadas a_1 ... a_n. Cada alinhamento a_i possui um lado esquerdo (*l*) e um direito (*r*) analisados separadamente (por um *shallow parser*) como apresentado em (13). Nesse exemplo, a, b, c, d, e são lemas de *l* e a', b', c', d', e' são lemas de *r* anotados com informação morfossintática.

a_1: (a)b(c(d(e))) ⇔ (((a')b')c')d'(e') (13)

Para cada alinhamento a_i pode-se extrair $p \times q$ correspondências (alinhamentos) lexicais C_i: {c_1 ... c_{pxq}}, em que *p* é o número de nós em *l*, ou seja, o número de parênteses no lado esquerdo do alinhamento em (13) e *q*, o mesmo número no lado direito. Os alinhamentos a_i e as correspondências lexicais c_i são generalizados na próxima etapa do processo de indução.

No método proposto por (Menezes & Richardson, 2001), o algoritmo de alinhamento primeiro tenta encontrar correspondências lexicais entre nós das formas lógicas fonte e alvo (veja Figura 5) buscando pares de tradução em um léxico bilíngüe. Em seguida, considerando-se como ponto de partida os alinhamentos lexicais encontrados, o algoritmo alinha os nós restantes utilizando uma gramática de alinhamento com 18 regras de composição codificadas manualmente. O propósito dessa gramática é garantir que apenas alinhamentos lingüisticamente significativos sejam gerados. Assim, após o processo de alinhamento das FLs da Figura 5 tem-se o resultado apresentado na Figura 7, na qual as linhas pontilhadas indicam os alinhamentos entre os nós fonte e alvo.

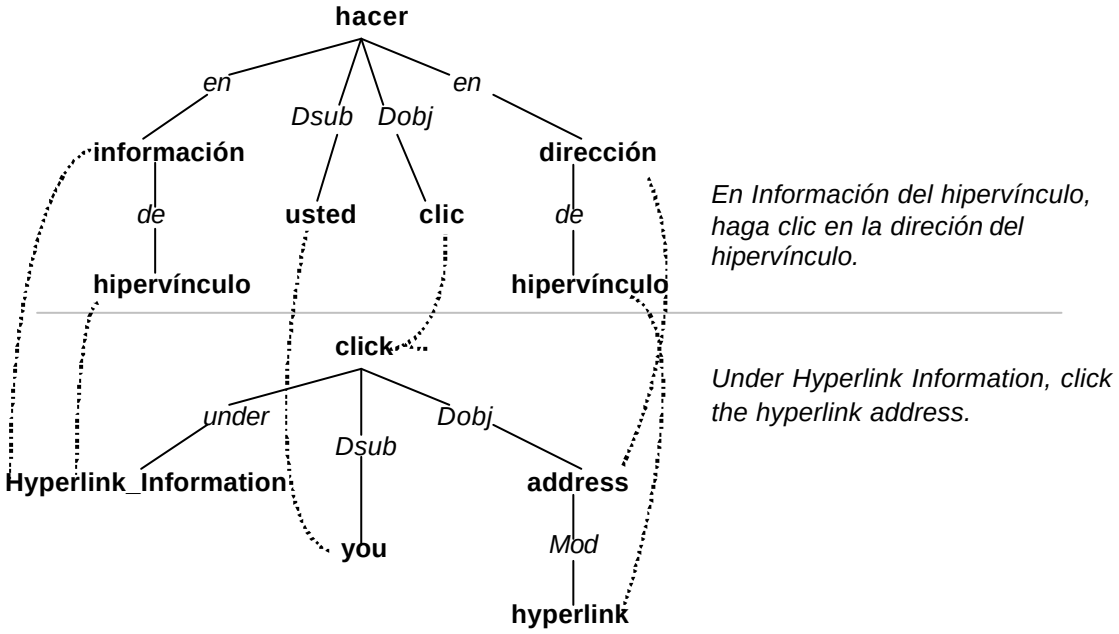

Figura 7. Alinhamentos das formas lógicas fonte e alvo da Figura 5 (Menezes & Richardson, 2001)

Em (Meyers et al., 1996, 1998, 2000), o processo de alinhamento possui uma restrição de preservação da dominância, ou seja, se um nó A domina um nó B na árvore fonte, então A' domina B' na árvore alvo, onde A' se alinha com A e B' se alinha com B. No alinhamento dos nós utiliza-se a técnica de programação dinâmica para calcular (de modo *bottom-up*) a pontuação do casamento (*matching*) de cada nó na árvore fonte com cada nó na árvore alvo, gerando uma matriz |Árvore Fonte| × |Árvore Alvo| a partir da qual as correspondências serão recuperadas para geração das regras de tradução.

O método proposto por Lavoie et alli (2001, 2002) converte as árvores sintáticas fonte e alvo dos córpus de treinamento e teste para uma representação de dependência sintática (como apresentado na Figura 6) e um dicionário de transferência lexical inicial é criado composto, inicialmente, por regras de transferência lexical (extraídas de dicionários bilíngües ou dos próprios bitextos usando métodos estatísticos) e/ou léxico-estruturais (geradas manualmente). A cada regra de transferência é atribuída uma taxa de *log likelihood* calculada com base no conjunto de treinamento (veja Figura 3).

De maneira semelhante ao sistema de Meyers et alli (1996, 1998, 2000), esse método alinha os nós das árvores fonte e alvo utilizando a técnica de programação dinâmica, porém, nesse caso, a busca pelo mapeamento menos custoso entre os nós é realizada de maneira *top-down* e bi-direcional. Nessa busca são considerados os custos de alinhar dois nós cujos lemas não estão no dicionário de transferência inicial ou que possuem *part-of-speech* (POS) ou posições relativas diferentes e o custo de remover ou inserir um nó em uma das árvores. Os alinhamentos são indicados pela inserção do atributo *aid* e alinhamentos vazios (0-1 ou 1-0) também são possíveis.

Em (Lavie et al., 2004), o sistema infere regras hierárquicas de transferência sintática baseando-se, inicialmente, nos constituintes das duas línguas alinhados lexicalmente (por um processo manual). Para isso, as sentenças de treinamento escritas na língua com mais recursos (o inglês, nesse caso) são analisadas sintaticamente e desambigüadas. Todos os componentes da regra descritos na seção 2.1 e apresentados na Figura 2 (com exceção das restrições de ambos os lados) são gerados considerando-se a estrutura sintática da língua com mais recursos, os alinhamentos lexicais e os dicionários das línguas fonte e alvo.

O método proposto por Galley et alli (2004), de modo semelhante a outros modelos estatísticos de TA como (Yamada & Knight, 2001) e (Gildea, 2003), utiliza informação sintática para extrair conhecimento aplicando cálculos estatísticos. Porém, diferentemente da maioria dos métodos estatísticos de TA, esse método não gera um modelo estatístico do

processo de tradução, mas, sim, regras simbólicas para expressar a relação entre uma árvore sintática na língua alvo e a sentença correspondente na língua fonte.

Dessa maneira, o algoritmo proposto pelos autores é bem diferente dos outros apresentados até o momento, pois parte de uma seqüência (S) de palavras na língua fonte e tenta encontrar a árvore sintática na língua alvo na qual S é mapeada. Para isso, as possíveis derivações (seqüências ordenadas de elementos que são ou um símbolo fonte ou uma subárvore alvo) de S são geradas seguindo algumas restrições. Na Figura 8 tem-se um exemplo de três possíveis derivações para a sentença, em francês, *"il ne va pas"* e a árvore para a sentença, em inglês, *"he does not go"*.

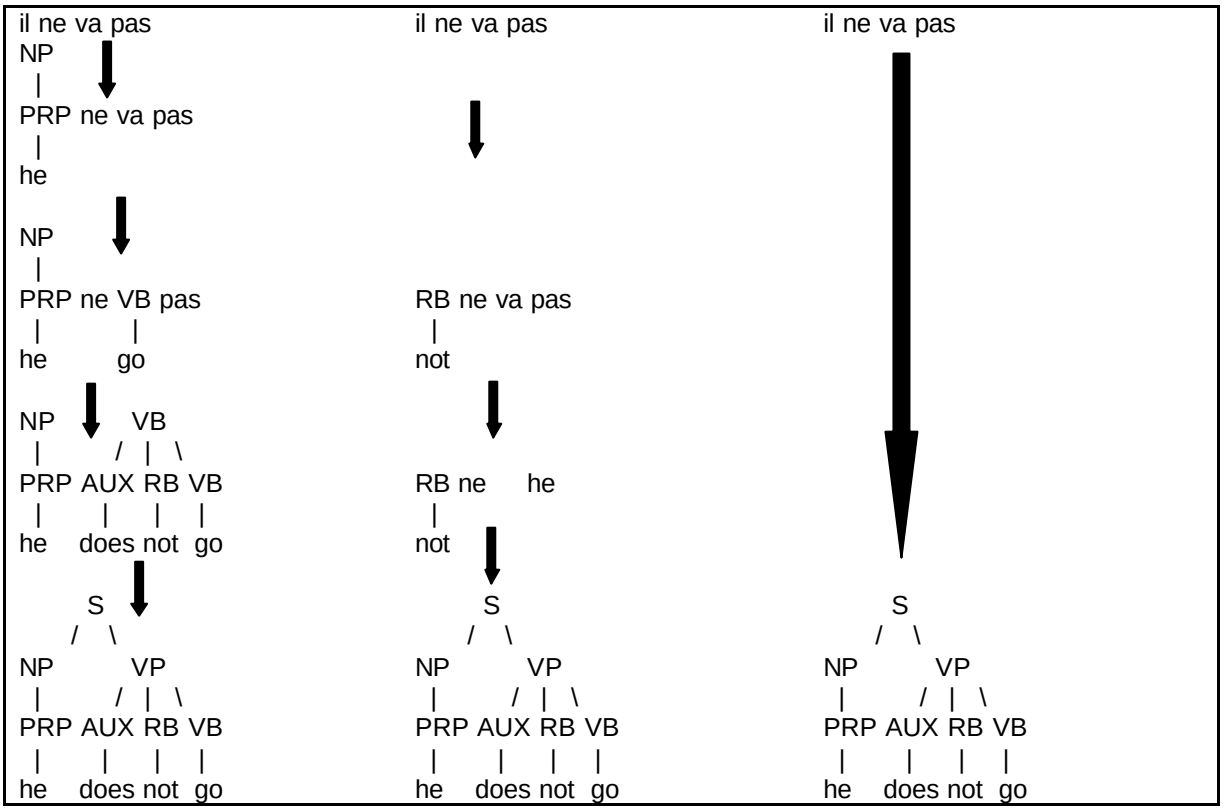

Figura 8. Três possíveis derivações de uma sentença fonte para uma árvore alvo (Galley et al., 2004)

O alinhamento, nesse caso, é realizado considerando-se que cada elemento de S deve ser substituído em exatamente um passo da derivação D (definido como *replaced(s,D)*; s ? S) e cada nó da árvore alvo T deve ser criado em exatamente um passo da derivação D (definido como *created(t,D)*; t ? T). Dessa forma, um elemento *s* de S é alinhado com um nó folha *t* de T se *replaced(s,D) = created(t,D)*. Em outras palavras, um elemento fonte é alinhado com um elemento alvo se o elemento alvo é criado durante o mesmo passo de derivação no qual o elemento fonte é substituído. Os alinhamentos induzidos pelas derivações apresentadas na Figura 8 são mostrados na Figura 9.

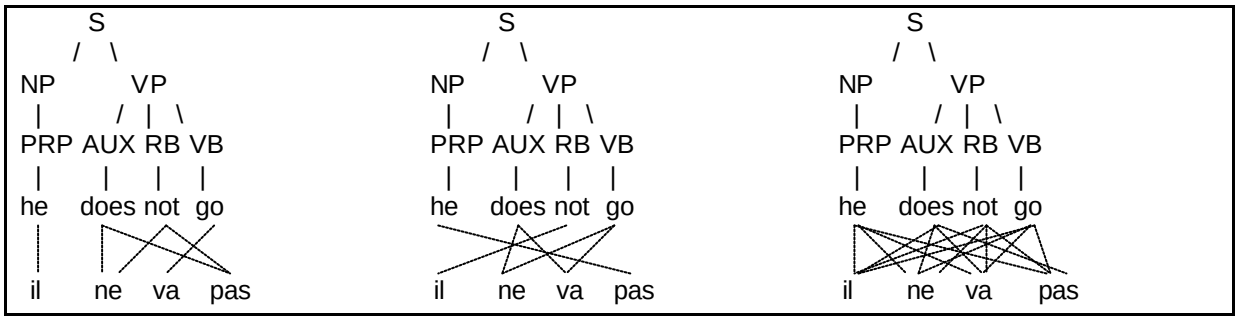

Figura 9. Alinhamentos induzidos pelas derivações da Figura 8 (Galley et al., 2004)

2.2.3 Geração das regras de tradução

Após a identificação dos padrões ou o alinhamento das árvores sintáticas, as regras são geradas aplicando diversas técnicas. Em métodos que identificam os padrões nos exemplos de tradução, as regras são geradas agrupando-se os padrões similares e/ou diferentes e, em seguida, generalizando-se esses padrões, ou seja, substituindo algumas de suas partes por variáveis. Com relação a essa substituição ela pode ser realizada com as similaridades – (Brown, 2001) –, com as diferenças – (Cicekli & Güvenir, 1996) e (Güvenir & Cicekli, 1998) – ou ambas – (McTait, 2003) e (Cicekli & Güvenir, 2003).

Em (Brown, 2001), os padrões identificados na etapa anterior são agrupados de acordo com a similaridade entre eles. As classes de padrões similares são, então, transformadas em regras de tradução pela simples substituição dos padrões por variáveis. Essas variáveis indicam a classe na qual o padrão substituído está inserido.

Nas primeiras versões do método de Cicekli e Güvenir (1996, 1998) a generalização dos padrões era feita apenas substituindo-se as partes similares, enquanto que em uma versão mais nova do método (Ciceckli & Güvenir, 2003) a heurística de diferenças também é aplicada. A heurística de similaridade permite que regras de tradução sejam inferidas quando o número de diferenças em ambos os lados de uma seqüência de similaridades é maior ou igual a 1 ($n = 1$) e, pelo menos, já foram inferidas, anteriormente, regras para *n-1* dessas diferenças. De maneira semelhante, a heurística de diferenças só permite que regras sejam inferidas se houver $n = 1$ similaridades não-vazias em ambos os lados e, pelo menos, regras inferidas para *n-1* dessas similaridades. Por exemplo, considerando-se os pares de sentenças inglês-turco apresentados, anteriormente, em (7) (Seção 2.1) e a respectiva seqüência de similaridades em (12) (Seção 2.2.1), só é possível inferir as duas primeiras regras de tradução apresentadas em (8) (Seção 2.1), aplicando-se a heurística de similaridade, se as regras em

(14) já tiverem sido inferidas anteriormente. A última regra em (8) é inferida utilizando-se a heurística de diferenças aplicada à única similaridade existente nos pares em questão.

book ⇔ kitap (14)
pencil ⇔ kursun kalem

Em (McTait, 2003), as colocações nas línguas fonte e alvo que possuem os mesmos identificadores de sentenças são consideradas traduções mútuas e, portanto, são o ponto de partida para a geração das regras. As regras são geradas utilizando-se programação dinâmica e uma métrica de similaridade bilíngüe baseada na distribuição lexical bilíngüe dos fragmentos de texto e o número de cognatos que os fragmentos compartilham. Regras complementares também são geradas com as partes diferentes do par de exemplos de tradução contanto que essas contenham itens lexicais que ocorram apenas uma vez no córpus. Um exemplo de uma regra de tradução gerada para a colocação "*gave up*" do par de sentenças em (5) (Seção 2.1) foi apresentada em (6) (Seção 2.1) enquanto que as regras de tradução complementares para esse mesmo par são apresentadas em (15).

The Commission Z_S^1 the plan Z_S^2 ⇔ La Comisión Z_T el plan (15)
Our Government Z_S^1 all laws Z_S^2 ⇔ Nuestro Gobierno Z_T todas las leyes

Em métodos que realizam a etapa de alinhamento das árvores sintáticas, as regras são geradas aplicando-se técnicas que variam desde simples cálculos estatísticos (probabilidade) ou recuperação dos alinhamentos lexicais, até processos mais complexos de expansão dos nós alinhados ou o processo inverso de extração de subpadrões nas árvores alinhadas.

Em (Carl, 2001), são atribuídos pesos para os alinhamentos (a_i) e as correspondências lexicais (c_i), com base nas suas probabilidades, para que apenas as generalizações (g_i) de maiores pesos sejam geradas. A gramática de tradução gerada é o conjunto formado pelos alinhamentos, correspondências lexicais e generalizações, por isso, outras heurísticas podem ser aplicadas para reduzir ainda mais o número de regras de tradução.

Em (Meyers et al., 2000), um processo simples também é aplicado no qual as regras de tradução são induzidas a partir das árvores sintáticas alinhadas, basicamente, recuperando-se as correspondências lexicais da matriz calculada durante o alinhamento das árvores fonte e alvo.

Em (Menezes & Richardson, 2001) o processo é um pouco mais complexo e envolve a expansão dos alinhamentos gerados na etapa anterior com tipos e quantidades variadas de

contexto utilizando construtores lingüísticos, como sintagmas nominais e verbais, para determinar as fronteiras dos contextos a serem inseridos. Assim, algumas das regras de tradução obtidas para os alinhamentos das formas lógicas apresentados na Figura 7, são apresentadas na Figura 10.

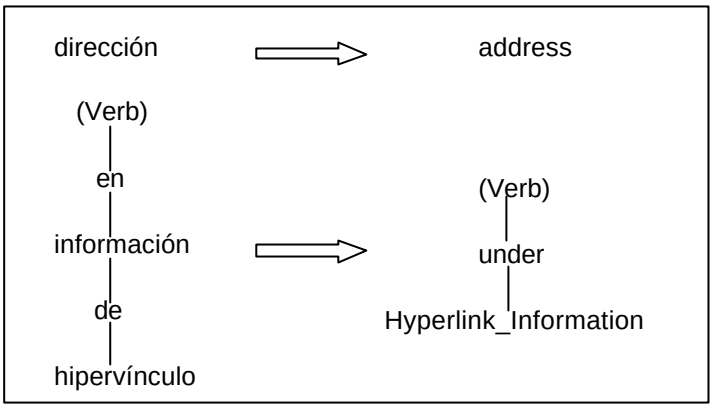

Figura 10. Regras de tradução obtidas para os alinhamentos das FLs apresentados na Figura 7 (Menezes & Richardson, 2001)

De maneira inversa, em (Lavoie et al., 2000), após o alinhamento dos nós, as regras de tradução candidatas são geradas extraindo-se subpadrões dos pares alinhados usando restrições para alinhamento e atributo, geradas manualmente. Enquanto as restrições de alinhamento definem a maioria das divergências sintáticas possíveis entre as línguas, as restrições de atributo limitam o espaço de regras de tradução que podem ser geradas a partir das subárvores que satisfazem as restrições de alinhamento. Um exemplo de uma restrição de alinhamento (que casa com os padrões estruturais da regra de transferência da Figura 3a) é apresentado na Figura 11. Essa restrição indica que qualquer par de subárvores fonte e alvo nas quais haja alinhamentos entre $X1 com $Y1 e $X2 com $Y3 pode ser usado como ponto de partida para a geração de regras de tradução.

```
@KOREAN:
        $X1 [aid=$1] (
                $R1 $X2 [aid=$2]
        )
@ENGLISH:
        $Y1 [aid=$1] (
                $R2 $Y2 (
                        $R3 $Y3 [aid=$2]
                )
        )
```

Figura 11. Restrição de alinhamento (Lavoie et al., 2002)

Após a geração das regras de tradução candidatas, o método tenta fazer a composição das regras induzidas verificando se cada subárvore pode ser obtida por uma regra gerada

previamente. Por fim, as regras são generalizadas por meio de um algoritmo denominado *Seeded Version Space Learning* que, primeiro, agrupa as regras similares (com mesmos POS, alinhamentos e tipos) e, em seguida, analisa cada grupo (*version space*) separadamente. O algoritmo tenta, repetidamente, unir duas regras com a remoção ou união de suas restrições. A regra resultante só é aceita se conseguir cobrir todos os casos cobertos pelas duas regras que ela substituirá. O processo continua até que não seja mais possível unir nenhum outro par de regras (Probst, 2002). A Tabela 3 apresenta uma regra generalizada (terceira coluna) para as Regras 1 e 2 (primeira e segunda colunas, respectivamente).

Tabela 3. Regras simples e generalizada (Carbonell et al., 2002)

Regra 1	**Regra 2**	**Regra Generalizada**
;;SL: the man	;;SL: the woman	;;SL:
;;TL: der Mann	;;TL: die Frau	;;TL:
NP::NP	NP::NP	NP::NP
[DET N] → [DET N]	[DET N] → [DET N]	[DET N] → [DET N]
(;;alignments:	(;;alignments:	(;;alignments:
(x1::y1)	(x1::y1)	(x1::y1)
(x2::y2)	(x2::y2)	(x2::y2)
;;x-side constraints:	;;x-side constraints:	;;x-side constraints:
((x1 agr) = *3-sing	((x1 agr) = *3-sing)	((x1 agr) = *3-sing)
((x1 def) = *def)	((x1 def) = *def)	((x1 def) = *def)
((x2 agr) = *3-sing)	((x2 agr) = *3-sing)	((x2 agr) = *3-sing)
((x2 count) = +)	((x2 count) = +)	((x2 count) = +)
;;y-side constraints	;;y-side constraints	;;y-side constraints
((y1 agr) = *3-sing)	((y1 agr) = *3-sing)	((y1 agr) = *3-sing)
((y1 case) = *nom)	((y1 case) = *not* *gen *dat)	
((y1 def) = *def)	((y1 def) = *def)	((y1 def) = *def)
((y2 gender) = *nom)	((y2 gender) = *f)	((y2 gender) = *f)
((y2 agr) = *3-sing)	((y2 agr) = *3-sing)	((y2 agr) = *3-sing)
((y2 case) = *nom)		
((y2 gender) = *m)	((y2 gender) = *f)	((y2 gender) = (y1 gender))

Em (Galley et al., 2004), as regras de tradução são geradas a partir das derivações aceitas para uma seqüência de palavras fonte (S) e uma árvore alvo (T) de acordo com um alinhamento (A) (veja Figura 12). O conjunto das derivações aceitas, de acordo com A, é formado pelas derivações que induzem os alinhamentos A' de tal forma que A é um subalinhamento de A', ou seja, A só alinha dois elementos se A' também os alinha. Assim, a segunda derivação apresentada na Figura 9 (veja Seção 2.2.2) para S não é uma derivação aceita de acordo com A.

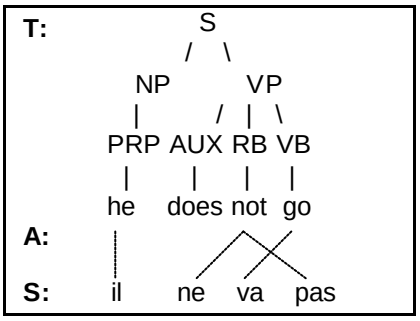

Figura 12. Uma sentença em francês (S) alinhada (A) com uma árvore sintática em inglês (T) (Galley et al., 2004)

Depois de determinado o conjunto das derivações aceitas, o próximo passo é converter cada passo de derivação em uma regra. Essas regras podem ser inferidas diretamente a partir de fragmentos de um grafo de alinhamento correspondente a S, T e A. Um grafo de alinhamento é uma T aumentada com um nó para cada elemento de S e arcos do nó folha *t* ? T para o elemento *s* ? S se há, em A, um alinhamento entre *s* e *t*. Além disso, cada nó do grafo é anotado com o conjunto de palavras (entre {}) de S que podem ser alcançadas por esse nó. A Figura 13 traz dois fragmentos do grafo de alinhamento para o exemplo da Figura 12 e as regras induzidas para eles. A entrada de uma regra é formada pelas raízes dos elementos da derivação e a saída é a árvore de símbolos com algumas folhas substituídas por variáveis Xi, em que *i* é a posição do elemento na entrada.

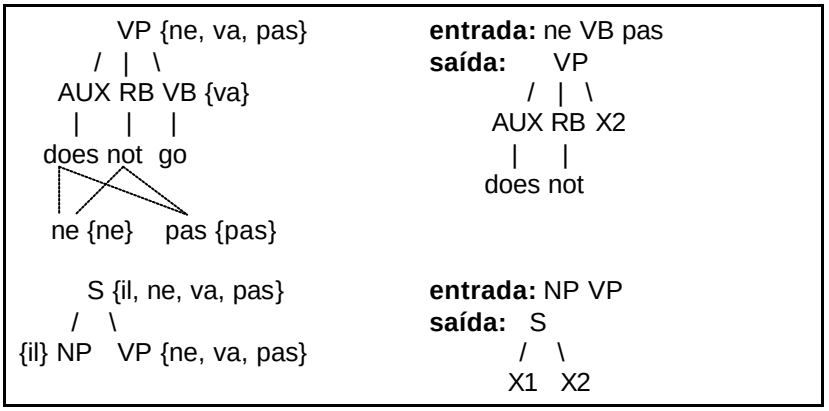

Figura 13. Dois fragmentos do grafo e as regras induzidas (Galley et al., 2004)

2.2.4 Filtragem e ordenação

Após a geração das regras de tradução, alguns métodos filtram as mesmas, por exemplo, para eliminar ambigüidades e/ou ordenam as regras geradas com base em algum critério estatístico (probabilidade, freqüência, etc.) ou de especificidade (ou generalização).

Em (Carl, 2001), a gramática gerada é filtrada para eliminar ocorrências ambíguas, ou seja, regras de tradução que possuem mesmo lado esquerdo ou lado direito. Nesse processo, apenas a regra de maior peso (calculado com base em probabilidade) é mantida para cada conjunto ambíguo. A Figura 14 apresenta um exemplo de gramática de tradução gerada após o processo de filtragem, em que *p(.)* e *w(.)* indicam, respectivamente, a probabilidade e o peso calculados para cada alinhamento (a_i), generalização (g_i) e correspondência lexical (c_i). Nesse exemplo, como as generalizações g_1 e g_2 são ambíguas, apenas a de maior peso (g_1) é mantida na gramática filtrada.

Regras de transferência induzidas	**p(.)**	**w(.)**
a_1: (dx) ⇔ (m'n')	1/4	2/4
g_1: (d*) ⇔ (m'*)	1/4	2/4
c_1: (x) ⇔ (n')	1/4	1/4
a_2: (de) ⇔ (a'b')	1/8	1/4
g_2: (d*) ⇔ (*b')	1/8	1/4
c_2: (e) ⇔ (a')	1/8	1/8
Gramática filtrada	**p(.)**	**w(.)**
g_1: (d*) ⇔ (m'*)	1/4	2/4
c_1: (x) ⇔ (n')	1/4	1/4
a_2: (de) ⇔ (a'b')	1/8	1/4

Figura 14. Gramática induzida e filtrada (Carl, 2001)

Em (Menezes & Richardson, 2001) uma filtragem das regras é realizada baseada na freqüência de ocorrência, o que, segundo os autores, melhora consideravelmente o tempo de processamento do sistema de TA que utiliza as regras induzidas.

Alguns métodos ordenam as regras geradas de acordo com a especificidade, freqüência, ou outro critério. Por exemplo, no método de Cicekli e Güvenir (2003) as regras de tradução são ordenadas por especificidade – a regra com maior número de terminais (palavras e não variáveis) na língua fonte é a mais específica – ou por algum fator de confiança – como o modelo estatístico apresentado em (Öz & Cicekli, 1998).

Em (Lavoie et al., 2002), as regras são ordenadas decrescentemente por suas taxas de *log likelihood* (o desempate é feito por uma heurística que prioriza as regras mais gerais) para serem filtradas, nessa ordem e uma de cada vez, removendo aquelas candidatas que não melhoram a precisão geral das árvores alvo produzidas. A cada iteração do processo de filtragem, uma regra candidata é adicionada provisoriamente ao conjunto de regras aceitas e o conjunto atualizado é aplicado a todas as estruturas fonte. As estruturas transferidas e as árvores alvo são comparadas e se o erro for menor do que o erro atual, a candidata permanece

no conjunto e o erro é atualizado; caso contrário, a candidata é rejeitada e removida do conjunto de regras aceitas.

2.3 Avaliação das regras de tradução

As regras de tradução resultantes do processo de indução podem ser avaliadas diretamente, por um especialista, ou indiretamente, por meio de seu uso em um sistema de TA. No primeiro caso, avaliam-se as regras de tradução resultantes do processo de indução (veja repositório da Figura 1), enquanto que, no segundo caso, as regras são usadas (recombinadas) para traduzir sentenças fonte em sentenças alvo e a avaliação é feita com base nas sentenças alvo produzidas (veja processo de recombinação ilustrado na Figura 1).

Tradicionalmente, em ambos os casos, o processo é trabalhoso e necessita da ajuda de especialistas para se determinar, por exemplo, a precisão e/ou cobertura das regras de tradução ou a aceitabilidade das sentenças alvo geradas. Uma alternativa para tornar a avaliação indireta menos trabalhosa é realizá-la automaticamente por meio de alguma métrica capaz de julgar a qualidade da sentença alvo com base em uma ou mais sentenças de referência (consideradas corretas). Assim, a seguir são apresentadas as diferentes metodologias de avaliação de regras de tradução: direta, indireta não-automática e indireta automática.

2.3.1 Avaliação direta

Na avaliação direta das regras de tradução, um tradutor humano especialista nas duas línguas envolvidas é responsável por analisar as regras induzidas e julgá-las segundo sua cobertura, relevância, precisão ou qualquer outro critério de interesse. A avaliação direta é mais trabalhosa do que a avaliação indireta pois, além de ser especialista nas duas línguas, o tradutor humano deve também estar familiarizado com o formalismo de representação das regras.

Uma alternativa para a avaliação direta – utilizada em (Carl, 2001) – é considerar como língua alvo a mesma língua fonte e verificar, nesse caso, a porcentagem de regras geradas com o lado esquerdo (fonte) igual ao lado direito (alvo). Dessa maneira, a qualidade da gramática de tradução é avaliada contando-se o número de regras com lados direito e esquerdo iguais. Porém, essa alternativa se limita a verificar a precisão das regras e não pode

ser usada para analisar cobertura, relevância ou outros critérios. Os resultados da avaliação do método de Carl (2001) são apresentados na Seção 2.3.4.

2.3.2 Avaliação indireta não-automática

Na avaliação indireta não-automática, o processo é um pouco menos trabalhoso já que, nesse caso, o tradutor humano não precisa estudar o formalismo de representação das regras de tradução para analisá-las. Além disso, também não é necessário (em muitos casos) que ele tenha conhecimento das duas línguas, bastando que seja especialista apenas na língua alvo.

Assim, as regras de tradução são utilizadas em um sistema de TA e o tradutor humano precisa julgar, por exemplo, se a sentença alvo gerada para uma dada entrada na língua fonte é adequada (ou não) – como em (Öz & Cicekli, 1998) – ou, ainda, se é melhor do que outra gerada por um sistema em comparação (desenvolvido com outra tecnologia) – como o Babelfish[10], por exemplo, em (Menezes & Richardson, 2001) e (Lavoie et al., 2002). Os resultados das avaliações desses métodos são apresentados na Seção 2.3.4.

2.3.3 Avaliação indireta automática

Na avaliação indireta automática, a sentença alvo (candidata) gerada pelo sistema de TA com regras induzidas é comparada com uma ou mais sentenças de referência (consideradas corretas) por meio de uma métrica. Essa metodologia é a que tem sido mais aplicada, atualmente, para avaliar os sistemas de TA.

Alguns estudos – (Doddington, 2002), (Turian et al., 2003) e (Finch et al., 2004) – sobre o número de sentenças de referências que devem ser utilizadas em uma avaliação indireta automática constataram que quanto maior o número de referências, melhor a performance da avaliação. Outra constatação foi a de que a métrica NIST (apresentada em detalhes a seguir) tem uma melhora gradual de performance quando até 4 referências são usadas, mas com mais de 4 referências sua performance começa a cair.

Os trabalhos mais recentes em avaliação de sistemas de TA utilizam métricas que estão se tornando padrão, como BLEU (Papineni et al., 2002) e NIST (Doddington, 2002); além das tradicionais precisão (*precision*), cobertura (*recall*) e medida-F (*F-measure*) (Melamed et al., 2003). A seguir é apresentada uma breve descrição sobre cada uma dessas métricas.

[10] http://world.altavista.com (16/08/2004).

BLEU

A métrica BLEU (cujo nome provém de *BiLingual Evaluation Understudy*) avalia a saída de um sistema de TA medindo a precisão dos *n*-gramas (de tamanhos que variam de 1 a 4, nesse caso) das sentenças alvo geradas automaticamente em relação a um conjunto de traduções de referência. A idéia por trás dessa métrica é a de que uma boa tradução tem mais *n*-gramas em comum com as sentenças de referência do que uma tradução ruim (Finch et al., 2004).

A BLEU é calculada como a média geométrica da precisão de *n*-grama, multiplicada pela penalidade de brevidade (*brevity penalty*, ou BP) que penaliza sentenças muito menores do que a(s) referência(s). Dessa maneira, a melhor candidata deve ser similar à(s) referência(s) em tamanho, escolha e ordem das palavras. A BLEU é calculada como mostra (16), em que o valor de *N* proposto pelos autores é 4 e p_n e BP são calculados como em (17) e (18), respectivamente.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} \frac{1}{N} \ln\, p_n\right) \tag{16}$$

$$p_n = \frac{\sum_{w_1 \ldots w_n \in C} Count_{clip}(w_1 \ldots w_n)}{\sum_{w_1 \ldots w_n \in C} Count(w_1 \ldots w_n)} \tag{17}$$

em que C é a candidata a tradução, Count(w_1 ... w_n) é o número de vezes que o *n*-grama w_1 ... w_n ocorre na candidata a tradução C e $Count_{clip}$(w_1 ... w_n) é o número de vezes que o *n*-grama w_1 ... w_n casa com um *n*-grama de referência, limitado pelo número máximo de vezes que ele ocorre em qualquer uma das referências.

$$BP = \begin{cases} 1 & \text{se } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{se } c = r \end{cases} \tag{18}$$

em que *c* é o tamanho da candidata C e *r* é o tamanho médio das referências para essa candidata.

A medida de precisão (p_n) captura dois aspectos da tradução: adequação e fluência. Uma tradução que utiliza as mesmas palavras (1-grama) que as referências tende a satisfazer a adequação enquanto que a existência de seqüências maiores de *n*-gramas em comum está relacionada à fluência (Papineni et al., 2002).

O valor da métrica BLEU, para uma candidata, varia entre 0 e 1, sendo que quanto mais próximo do 1, melhor é a sentença candidata em relação às sentenças de referência.

Em dois dos métodos de indução de regras de tradução apresentados anteriormente – (Lavoie et al., 2002) e (Lavie et al., 2004) – essa métrica foi utilizada para avaliar o desempenho, como apresentado em detalhes na Seção 2.3.4.

NIST

A métrica NIST, assim como a BLEU, também se baseia em precisão de *n*-gramas (que variam, nesse caso, de 1 a 5), porém ela emprega a média aritmética das quantidades de *n*-gramas ao invés da média geométrica como faz a BLEU. Outra diferença entre essas duas métricas é que, na NIST, os *n*-gramas são ponderados por pesos de acordo com a contribuição de informação que fornecem ao invés de simplesmente serem contados como acontece na BLEU (Finch et al., 2004).

A NIST representa a informação média, por palavra, dada pelos *n*-gramas na candidata que casam com um *n*-grama de uma das referências no conjunto de referências. A penalidade de brevidade (BP') da NIST, em relação à BP da BLEU, penaliza mais seriamente as candidatas muito pequenas e menos as candidatas mais próximas das referências, em tamanho. Assim, a NIST é calculada como mostra (19) em que C, c, r e , Count(w_1 ... w_n) são os mesmos definidos para a BLEU e $N = 5$. Info[11] e BP' são mostradas em (20) e (21), respectivamente.

$$NIST = BP \times \sum_{n=1}^{N} \sum_{w_1 \ldots w_n \in C} \frac{Info(w_1 \ldots w_n)}{Count(w_1 \ldots w_n)} \tag{19}$$

$$Info(w_1 \ldots w_n) = \log_2 \left[\frac{número\_de\_ocorrências\_de\_w_1 \ldots w_{n-1}}{número\_de\_ocorrências\_de\_w_1 \ldots w_n} \right] \tag{20}$$

$$BP' = \begin{cases} 1 & \text{se } c > r \\ \exp\left(b \ln^2 \left(\frac{c}{r} \right) \right) & \text{se } c = r \end{cases} \tag{21}$$

em que ß é selecionado de tal forma que quando c = 2r/3 BP' = 0,5.

O valor da NIST é sempre positivo e quanto maior ele for, melhor é a candidata em relação às referências; porém não há um limite fixo para o valor máximo dessa métrica.

Em (Lavie et al., 2004) o sistema de TA com as regras induzidas, além de ser avaliado com a métrica BLEU, também foi avaliado usando a métrica NIST, como apresentado em detalhes na Seção 2.3.4.

[11] As quantidades de *n*-gramas usadas para calcular os pesos de informação são derivadas do conjunto de referência.

Precisão, cobertura e medida-F

Embora as métricas apresentadas anteriormente sejam úteis na comparação da qualidade das sentenças alvo geradas por diferentes sistemas de TA, é difícil entender o que elas significam, ou seja, o que significa, por exemplo, um valor de 0,112 para BLEU ou 5,32 para NIST? Nesse sentido, em (Melamed et al., 2003), os autores demonstram como sistemas de TA podem ser avaliados em termos das métricas bem conhecidas precisão e cobertura. Os autores sustentam que essas métricas podem ser interpretadas graficamente de maneira intuitiva, o que torna mais fácil o entendimento dos problemas dos sistemas de TA avaliados e de como esses problemas podem ser solucionados.

Precisão, cobertura e medida-F são utilizadas há muitos anos para avaliar diversos sistemas de PLN em áreas como recuperação de informação e alinhamento de textos paralelos. Precisão e cobertura são calculadas comparando-se os itens candidatos (Y) com os itens de referência (X) como mostram as equações (22) e (23), e a medida-F (24) é a combinação das duas métricas anteriores. Assim, a precisão demonstra o número de itens candidatos corretos (|X n Y|) em relação à quantidade total de itens candidatos (|Y|), enquanto a cobertura indica o número de itens candidatos corretos (|X n Y|) em relação à quantidade total de itens de referência (|X|).

$$precisão(Y \mid X) = \frac{|X \cap Y|}{|Y|} \tag{22}$$

$$cobertura(Y \mid X) = \frac{|X \cap Y|}{|X|} \tag{23}$$

$$medida - F = 2\frac{cobertura \times precisão}{cobertura + precisão} \tag{24}$$

A precisão verifica a capacidade do sistema em traduzir corretamente as sentenças, enquanto a cobertura indica a capacidade do sistema em traduzir corretamente o maior número possível de sentenças do conjunto de teste. A medida-F, por sua vez, representa a combinação das duas métricas anteriores. Os valores para essas três métricas variam entre 0 e 1, sendo que um valor próximo do 1 significa uma boa qualidade do sistema avaliado.

Além de ser uma métrica bem conhecida e mais fácil de compreender, a medida-F mostrou-se, em alguns casos, mais confiável do que a BLEU e a NIST para avaliar os sistemas de TA nos experimentos apresentados em (Turian et al., 2003). Em outra avaliação apresentada em (Finch et al., 2004), constatou-se, também, que a medida-F é a melhor métrica quando são usadas quatro referências ou mais.

Dentre a métricas utilizadas para avaliar o sistema de TA em (McTait, 2003), a cobertura foi a mais explorada. Em (Brown, 2001), os autores também utilizaram cobertura para avaliar seu sistema, porém de uma maneira diferente da apresentada em (23); e em (Meyers et al., 2000), os autores utilizam a medida-F para avaliar o sistema, mas com outra denominação (*accuracy*). Os resultados das avaliações desses métodos, com essas métricas, são apresentados na Seção 2.3.4.

2.3.4 Avaliação dos métodos de indução de regras de tradução

A seção anterior apresentou as diferentes maneiras de se avaliar os métodos de indução de regras de tradução. Nessa seção são apresentados os resultados das avaliações dos métodos citados na Seção 2.2 de acordo com o tipo de avaliação empregado.

Com relação à metodologia de avaliação direta das regras induzidas, entre os métodos citados, apenas em (Carl, 2001) ela foi empregada. Nessa avaliação foram utilizados 4.997 alinhamentos com análise sintática parcial (*shallow parsing*) e a qualidade da gramática de tradução foi avaliada contando-se o número de regras com lados direito e esquerdo iguais, já que as línguas fonte e alvo utilizadas na avaliação eram as mesmas. A gramática gerada possuía 4.506 regras de tradução (814 generalizações e 3.692 regras de transferência lexical e alinhamentos) das quais 3.698 (82%) tinham lados esquerdo e direito iguais e 808 (18%) diferentes. Com 50% do córpus coberto por um léxico, o número de regras com lados esquerdo e direito diferentes caiu de 18% para 3,4%. Assim, conclui-se que esse método obteve uma precisão de 82 a 96,6% na avaliação direta das regras de tradução.

Entre os métodos de indução de regras de tradução que tiveram suas regras avaliadas indiretamente por um especialista humano (avaliação indireta não-automática), estão aqueles que foram avaliados individualmente – (Öz & Cicekli, 1998) – e os que foram comparados com outros sistemas de TA – (Menezes & Richardson, 2001) e (Lavoie et al., 2002).

Em (Öz & Cicekli, 1998), a avaliação do sistema foi realizada com o intuito de verificar a precisão das regras induzidas. Para isso, as 4.723 regras de tradução extraídas para sentenças em inglês-turco foram ordenadas de acordo com fatores de confiança calculados por meio de um modelo estatístico gerado para um córpus de treinamento (composto por 488 sentenças). Em seguida, as regras ordenadas foram utilizadas em um sistema de TA constatando-se que 60% das 5 primeiras traduções geradas para cada sentença de entrada estavam corretas.

Em (Menezes & Richardson, 2001), para o treinamento do método foram utilizados 161.606 pares de sentenças espanhol-inglês (a maioria proveniente de manuais técnicos) dos quais foram extraídos 58.314 regras de tradução em um processo que levou 74 minutos em um PC de 800MHz (35,6 sent/s). O desempenho do sistema foi comparado ao do Babelfish por um tradutor humano em um córpus de teste de tamanho variável (de 200 a 500 sentenças do mesmo domínio nunca vistas pelo sistema e selecionadas aleatoriamente). As traduções geradas pelo sistema foram consideradas melhores do que as geradas pelo Babelfish em 46,5% dos casos. Porém, é importante citar que o Babelfish não foi treinado para o domínio em questão.

Em outra avaliação indireta, apresentada em (Lavoie et al., 2002), foram utilizados 1.483 pares de árvores sintáticas geradas automaticamente para sentenças dos manuais de treinamento do *U.S. Defense Language Institute* em coreano-inglês. Para o conjunto de teste foram selecionados aleatoriamente 50 pares (com no mínimo 5 nós) e os outros 1.433 pares foram usados no treinamento do sistema.

O desempenho de duas versões do sistema – uma apenas com o léxico (Lex) e outra com o léxico e as regras induzidas (Lex+Ind) – e um sistema comercial, o Babelfih, foi avaliado por dois tradutores humanos nativos da língua inglesa. Os especialistas humanos analisaram a qualidade das sentenças alvo geradas e constataram que em 46% dos casos o Lex+Ind foi melhor do que o Babelfish, empatando com este em 27% dos casos e sendo pior nos 27% restantes. Em relação à versão que utilizava apenas o léxico (Lex), o Lex+Ind foi melhor em 41% dos casos, empatou em 41% e foi pior nos 18% restantes.

O método de Lavoie et alli (2002) também foi avaliado automaticamente calculando-se o valor da métrica BLEU ((16) na Seção 2.3.3). Os valores obtidos para essa métrica, assim como na avaliação indireta não-automática, apontam a versão Lex+Ind (BLEU = 0,0950) como a de melhor desempenho, seguida pelo sistema Babelfish (BLEU = 0,0802) e a versão Lex (BLEU = 0,0767).

Outro exemplo de avaliação indireta automática que utiliza a métrica BLEU é o de (Lavie et al., 2004), na qual comparou-se o desempenho de um sistema de TA que usa as regras inferidas com dois sistemas desenvolvidos pelo mesmo grupo de pesquisa: um estatístico e outro baseado em exemplos. Os três sistemas foram treinados com um córpus hindi-inglês com 17.589 sentenças e sintagmas alinhados lexicalmente e um léxico hindi-inglês com 23.612 pares de tradução. O teste foi realizado com 258 sentenças e as métricas BLEU e NIST ((19) na Seção 2.3.3) foram calculadas para cada sistema. Os resultados mostram que o sistema com regras induzidas (BLEU = 0,112 e NITS = 5,32) foi melhor do

que o estatístico (BLEU = 0,102 e NIST = 4,70) e o baseado em exemplos (BLEU = 0,058 e NIST = 4,22).

Em (McTait, 2003), também foi realizada uma avaliação indireta automática porém, desta vez, para se verificar a cobertura das regras geradas por três versões do sistema: a versão básica (sem recursos lingüísticos além do córpus paralelo); a versão com informação morfológica (regras morfológicas, lista de lemas, etc.) e a versão com etiquetação de POS (usando o etiquetador TreeTagger (Schmid, 1994)). Segundo o autor, o conhecimento lingüístico foi adicionado como uma tentativa de aumentar a precisão dos padrões de tradução e a cobertura geral no conjunto de exemplos. Para verificar essa hipótese, foi realizado um treinamento com cada versão do sistema usando 2.500 pares de sentenças inglês-francês do córpus da "*World Health Organisation AFI*"[12]. A versão básica extraiu 9.327 padrões; a versão com informação morfológica extraiu 9.610 (0,03% a mais do que a versão básica) e a versão com etiquetação POS extraiu 7.237 (22,4% a menos do que a versão básica).

O teste foi realizado com um conjunto de 1.000 sentenças nunca vistas pelas versões do sistema. Considerando-se apenas traduções completas, o melhor desempenho foi o da versão com informação morfológica (cobertura = 33,9%), seguido pela versão básica (cobertura = 32,2%) e a versão com etiquetação de POS (cobertura = 27,2%). Segundo o autor, o baixo desempenho da versão com etiquetação de POS, provavelmente, foi devido aos erros do etiquetador.

Outra avaliação indireta realizada com o intuito de analisar a cobertura das regras de tradução foi apresentada em (Brown, 2001). Porém, nessa avaliação, a cobertura de cada candidata a tradução foi calculada como a porcentagem do total de palavras na sentença fonte de entrada para as quais o sistema gera, pelo menos, uma palavra alvo como tradução. O sistema foi analisado para córpus inglês-espanhol e inglês-francês com tamanhos variados e os resultados mostram que o sistema de indução combinado com um passo posterior de agrupamento (*clustering*) obteve bons resultados nos dois pares de línguas, chegando a 92,34% de cobertura em um córpus com 1.107.000 exemplos inglês-francês e 89,44% em um córpus com 1.000.000 de exemplos inglês-espanhol. Para córpus menores, o desempenho foi de 77,70% de cobertura para um córpus com 107.000 exemplos inglês-francês e de 72,23% para um córpus com 104.000 exemplos inglês-espanhol.

Em uma avaliação indireta automática apresentada em (Meyers et al., 2000), foram realizados dois experimentos com conjuntos diferentes de pares de sentenças espanhol-inglês

[12] http://www.who.int (16/08/2004).

extraídos do *Microsoft Excel Help Text*. No Experimento 1 utilizou-se um conjunto com 1.155 pares alinhados manualmente, enquanto que no Experimento 2 foram utilizados 2.617 pares alinhados automaticamente[13] e para os quais não houve uma otimização prévia (treinamento com textos de mesmo domínio do córpus de teste) do analisador sintático.

Em ambos os experimentos, os conjuntos foram divididos em 90% para treinamento e 10% para teste, variando-se o conjunto de treinamento e teste em 10 iterações (10-*fold cross validation*). Nas 10 iterações do processo de treinamento foram adquiridas, em média, 1.109 regras de tradução para o Experimento 1 e 2.191 regras para o Experimento 2. As regras de tradução geradas em cada um dos conjuntos de treinamento nas 10 iterações de cada experimento foram usadas para traduzir os conjuntos de teste correspondentes. As sentenças geradas como saída para esses conjuntos de teste foram, então, comparadas com as sentenças originais da Microsoft usando a medida-F ((24) na Seção 2.3.3), que os autores denominaram de *accuracy*. Os valores obtidos para essa métrica nos experimentos 1 e 2 foram 70,9 e 62,6%, respectivamente.

Por fim, o método de (Galley et al., 2004) foi avaliado de uma maneira completamente diferente. Partindo-se de córpus inglês-chinês (FBIS, com oito milhões de palavras em inglês) e inglês-francês (Hansard), as sentenças em inglês foram analisadas sintaticamente e um alinhamento lexical foi gerado descartando os alinhamentos diferentes de 1-1. Verificou-se, então, se era possível transformar as árvores fonte (em inglês) em sentenças alvo (em chinês ou francês) e quantas expansões da regra eram necessárias para isso. Como resultado, constatou-se que foi possível transformar 100% das árvores em sentenças alvo com 17 a 43 expansões das regras, sendo 11,8s o tempo de processamento para regras com até 50 expansões. Além disso, os autores concluiram que as regras de tradução geradas pelo método são sintática e lexicalmente motivadas; o método é independente do mecanismo de alinhamento lexical utilizado e é bem adequado para lidar com erros introduzidos pelas ferramentas automáticas de alinhamento e análise sintática. Contudo, por se tratar de um método estatístico, o tamanho do córpus paralelo utilizado no processo de extração de regras de tradução deve ser muito maior (na ordem de milhões de palavras) do que o utilizado por outros métodos de indução de regras de tradução não estatísticos.

Embora os métodos de indução de regras de tradução citados na Seção 2.2 tenham sido avaliados utilizando diversas metodologias (avaliação direta, avaliação indireta não-

[13] O sistema utilizado pelos autores do experimento para alinhar as sentenças está descrito em: MEYERS, A.; KOSADA, M.; GRISHMAN, R. (1998). A multilingual procedure for dicionary-based sentence alignment, In: *Proceedings of AMTA'98: Machine Translation and the Information Soup*. p.187-198.

automática ou automática) e métricas (precisão, cobertura, BLEU, NIST, etc.), em córpus de idiomas, gêneros e tamanhos muito variados, é possível identificar alguns pontos importantes nas avaliações apresentadas nesta seção. Para tanto, a Tabela 4 resume os resultados obtidos nas avaliações dos métodos agrupando-os de acordo com a metodologia de avaliação empregada (D – direta, I – indireta não-automática e IA – indireta automática) e a realização (S) (ou não, N) de análise sintática (AS).

Tabela 4. Resumo das avaliações apresentadas nesta seção

AS	Referência	Tipo	Métrica	Valores
S	Carl, 2001	D	Precisão	82 a 96,6%
N	Öz & Cicekli, 1998	I	Precisão	60% das 5 primeiras traduções estavam corretas
N	McTait, 2003	IA	Cobertura	27,2 a 33,9%
N	Brown, 2001	IA	Cobertura[14]	72,23 a 92,34%
S	Menezes & Richardson, 2001	I	Sistema com indução (SI) X Babelfish	SI melhor em 46,5% dos casos
S	Lavoie et al., 2002	I	Sistema com indução (SI) X Babelfish	SI melhor em 46% dos casos
S	Lavoie et al., 2002	IA	BLEU	0,0950
S	Lavie et al., 2004	IA	BLEU NIST	0,112 5,32
S	Meyers et al., 2000	IA	Medida-F	62,6 a 70,9%
S	Galley et al., 2004	I	Quantidade de árvores transformadas em sentenças alvo	100% com 17 a 43 expansões das regras

Antes de comentar os valores apresentados na Tabela 4, são necessárias algumas considerações. Com relação aos valores relatados em (Brown, 2001), além da métrica cobertura utilizada nessa avaliação ser mais tolerável do que a cobertura tradicional (apresentada em (23) na Seção 2.3.3), o tamanho do córpus usado no processo de indução foi muito maior do que o utilizado nos outros métodos, por exemplo, 1.107.000 exemplos inglês-francês para se atingir a cobertura de 92,34% e 107.000 exemplos nos mesmos idiomas para se alcançar 77,70% de cobertura.

Com base nos valores da Tabela 4 é possível constatar que os métodos que realizam análise sintática foram avaliados com metodologias e métricas que permitem compará-los com outros sistemas disponíveis comercialmente, como é o caso do Babelfish. Talvez, por esse motivo, os métodos com análise sintática, aparentemente, possuem melhor desempenho do que os métodos que não realizam essa análise. Como já mencionado anteriormente, os

[14] É importante lembrar que a cobertura, nesse caso, foi calculada como a porcentagem do total de palavras na sentença fonte de entrada para as quais o sistema gera, pelo menos, uma palavra alvo como tradução.

valores de BLEU e NIST não são de fácil compreensão, mas pode-se dizer que nas avaliações com essas métricas os sistemas de TA que utilizavam as regras induzidas se saíram melhor do que o Babelfish (Lavoie et al., 2002) e sistemas estatístico e baseado em exemplos (Lavie et al., 2004).

Assim, não é possível, nesse momento, apontar um método de indução de regras de tradução como o melhor, nem mesmo dizer qual é o estado da arte em termos de precisão, cobertura ou alguma outra métrica, nessa área.

Capítulo 3

Projeto ReTraTos

Tendo em vista o cenário apresentado anteriormente, este capítulo visa delimitar o ambiente de desenvolvimento do projeto ReTraTos (Regras de Tradução induzidas de Textos Paralelos), bem como especificar os recursos que poderão ser utilizados durante este projeto.

O Núcleo Interinstitucional de Lingüística Computacional (NILC), grupo de pesquisa no qual o projeto ReTraTos se desenvolve, é um grupo interdisciplinar dedicado à pesquisa e ao desenvolvimento de sistemas de PLN. Esse grupo de pesquisadores de lingüística e computação tem desenvolvido recursos e aplicativos para o processamento do Português do Brasil desde 1993, dos quais podem ser citados léxicos computacionais, *parsers*, revisor gramatical, etiquetadores, analisadores morfológicos, base de dados lexicais, sumarizadores, alinhadores sentenciais e lexicais, etc.[15]

O projeto de mestrado da doutoranda, o PESA (*Portuguese-English Sentence Alignment*), também desenvolvido no NILC, teve como objetivo estudar e implementar as principais técnicas de alinhamento sentencial de textos paralelos propostas na literatura e avaliá-las para o par PB-inglês com córpus de gêneros científico, jurídico e jornalístico. Desse projeto resultaram quatro córpus de textos paralelos PB-inglês alinhados sentencialmente[16] e diversas ferramentas de processamento e alinhamento sentencial dos textos paralelos. O conhecimento e as ferramentas produzidos durante o projeto PESA serão utilizados na produção de outros córpus alinhados sentencialmente também para o par PB-espanhol, durante o projeto ReTraTos.

Além do PESA, referente ao alinhamento sentencial, também foi desenvolvido no NILC um projeto semelhante visando o alinhamento lexical, o PEWA (*Portuguese-English Word Alignment*), do qual resultaram córpus paralelos alinhados lexicalmente (gerados com base nos córpus alinhados sentencialmente do PESA) e ferramentas automáticas para o alinhamento lexical, que também poderão se usados no projeto ReTraTos.

Em relação à área de TA, um dos projetos desenvolvidos no NILC merece atenção especial: o projeto EPT-Web cujo objetivo é implementar um tradutor inglês-português de

[15] Informações sobre os recursos desenvolvidos pelo NILC podem ser obtidas na página: http://www.nilc.icmc.usp.br (16/08/2004).
[16] Página dos córpus paralelos: http://www.nilc.icmc.usp.br/nilc/tools/parallelcorpora.htm (16/08/2004).

páginas da *web* que utiliza a interlíngua UNL (*Universal Networking Language*) para traduzir as primeiras páginas da versão eletrônica do jornal norte-americano *"The New York Times"*. Os resultados relatados no EPT-Web (atualmente em fase de avaliação) serão utilizados na comparação com os resultados gerados pelo projeto ReTraTos, uma vez que o tradutor EPT-Web será um dos poucos recursos existentes e desenvolvidos especificamente para o PB.

Nesse contexto, são apresentados neste capítulo os recursos lingüísticos (Seção 3.1) e computacionais (Seção 3.2) disponíveis, atualmente, para o desenvolvimento do projeto ReTraTos, bem como algumas considerações sobre as técnicas de indução (Seção 3.3) e as metodologias de avaliação (Seção 3.4) que poderão ser utilizadas nesse projeto. Por fim, algumas considerações sobre o estágio de um ano que a doutoranda realizará na Universidade de Alicante (Espanha), como parte do projeto ReTraTos, são apresentadas na Seção 3.5.

3.1 Recursos lingüísticos

Como apresentado no Capítulo 1, um recurso lingüístico indispensável para a extração de regras de tradução usando a abordagem de EBMT e técnicas de Aprendizado de Máquina são os córpus paralelos alinhados sentencialmente. Atualmente, o NILC dispõe de córpus paralelos alinhados sentencialmente para o par PB-inglês resultantes do projeto PESA e diversos textos paralelos PB-inglês e PB-espanhol a serem alinhados.

Os córpus PB-inglês alinhados sentencialmente resultantes do projeto PESA podem ser divididos em três conjuntos diferentes, de acordo com o gênero: científico, jurídico e jornalístico.

O córpus científico é composto por 65 pares de textos paralelos (resumos e *abstracts*) provenientes de trabalhos acadêmicos da área de computação desenvolvidos no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP) em São Carlos, apresentados na forma de artigos publicados em revistas especializadas, monografias de qualificação de mestrado, dissertações de mestrado e teses de doutorado. Esses trabalhos pertencem a subdomínios variados da computação como: banco de dados, computação de alto desempenho, computação gráfica e processamento de imagens, engenharia de software, hipermídia, inteligência computacional, matemática computacional, sistemas digitais, sistemas distribuídos e programação concorrente. Os textos desse córpus são relativamente pequenos (aproximadamente 150 palavras) e possuem, em média, 6 sentenças com 25 palavras cada.

O córpus jurídico é composto por 4 pares de textos paralelos extraídos da documentação oficial da ALCA[17] (Área de Livre Comércio das Américas). Trata-se de declarações e atas de cúpulas de tamanhos maiores (cerca de 22.050 palavras cada) do que os textos do córpus anterior, cada um composto por cerca de 91 sentenças com 245 palavras cada, em média. Outros 8 pares de textos paralelos PB-inglês de documentos oficiais da ALCA estão disponíveis para serem alinhados totalizando 55.379 palavras (veja Tabela 6).

O córpus jornalístico, por sua vez, é composto por 8 pares de textos paralelos (artigos) extraídos do jornal *"The New York Times"*[18] das seções: artes, negócios, saúde e internacional. Cada texto do córpus jornalístico possui, em média, 30 sentenças e cada sentença possui, em média, 25 palavras (ou seja, 750 palavras por texto). De maneira semelhante ao córpus jurídico, existe, ainda, mais um par de textos paralelos do mesmo jornal (com 1.213 palavras)' pronto para ser alinhado (veja Tabela 6).

As quantidades de pares de textos paralelos, de palavras (*tokens*) e de alinhamentos sentenciais, em cada um desses córpus, são apresentadas na Tabela 5.

Tabela 5. Característica dos córpus paralelos PB-inglês alinhados sentencialmente disponíveis no NILC, hoje

Córpus / Quantidade	Científico	Jurídico	Jornalístico
Pares de textos paralelos	65	4	8
Palavras em PB	11.278	11.199	6.008
Palavras em inglês	10.164	10.848	5.759
Total de palavras	**21.442**	**22.047**	**11.767**
Alinhamentos sentenciais do tipo 1-1	395	362	225
Alinhamentos sentenciais diferentes de 1-1	21	1	8
Total de alinhamentos 1-1	**395**	**362**	**225**

Existem, ainda, outros córpus paralelos PB-inglês a serem alinhados provenientes dos gêneros jurídico (textos do Mercosul), literário (Contos) e jornalístico (textos da revista Fapesp).

Os textos do gênero jurídico são documentos oficiais do Mercosul (Mercado Comum do Sul) correspondentes a protocolos, tratados e declarações disponíveis na *web*[19]. São 6 pares de textos paralelos PB-inglês num total de 32.959 palavras.

[17] Disponíveis no site oficial da ALCA (http://www.ftaa-alca.org/alca_p.asp) (16/08/2004).
[18] Disponíveis na *web* em inglês (http://www.nytimes.com) e em PB (http://ultimosegundo.ig.com.br/nytimes/index.shtml) (16/08/2004).
[19] Disponíveis no site oficial do Mercosul (http://www.mercosur.org.uy) (16/08/2004).

O córpus literário é composto por contos australianos e canadenses (originalmente escritos em inglês) traduzidos para o PB[20]. São 13 pares de contos canadenses e 25 pares de contos australianos, ou seja, 38 pares de textos paralelos num total de 211.781 palavras.

Por fim, o último conjunto de textos paralelos PB-inglês disponível para ser utilizado no projeto ReTraTos é o córpus jornalístico composto por 646 pares de textos paralelos da revista da Fapesp disponível no Lácio-Web[21] correspondentes a um total de 901.054 palavras.

Assim, a Tabela 6 resume os números dos córpus paralelos PB-inglês que serão alinhados sentencialmente. Como o alinhamento ainda não foi realizado, é apresentada uma estimativa do número máximo de alinhamentos 1-1 que poderão ser gerados como sendo o menor valor entre as quantidades de sentenças em PB e em inglês.

Tabela 6. Característica dos córpus paralelos PB-inglês ainda não alinhados sentencialmente

Córpus / Quantidade	ALCA	NYT	Mercosul	Contos	Fapesp
Pares de textos paralelos	8	1	6	38	646
Palavras em PB	28.348	639	15.974	105.698	433.628
Palavras em inglês	27.031	574	16.985	106.083	467.426
Total de palavras	**55.379**	**1.213**	**32.959**	**211.781**	**901.054**
Sentenças em PB	1.266	22	930	8.147	18.758
Sentenças em inglês	1.268	20	948	7.802	18.147
Nº máximo de alinhamentos 1-1	**1.266**	**20**	**930**	**7.802**	**18.147**

Assim, o córpus disponível para a extração de regras de tradução para os idiomas PB-inglês possui cerca de 1.257.642 palavras (612.772 em PB e 644.870 em inglês) e 29.147 exemplos de tradução (pares de alinhamentos 1-1).

Com relação ao par PB-espanhol, existem atualmente no NILC pares de textos paralelos do gênero jurídico provenientes da documentação oficial da ALCA e do Mercosul. São 12 textos paralelos da ALCA num total de 84.943 palavras e 224 do Mercosul, num total de 750.336 palavras. Esses textos ainda não foram alinhados sentencialmente e, por isso, uma estimativa do número máximo de exemplos de tradução (alinhamentos 1-1) é apresentada na Tabela 7.

[20] A coleta dos textos ficou a cargo da Profa. Dra. Stella E. O. Tagnin, que também foi a responsável pela coordenação do grupo de tradução formado por alunos do Curso de Especialização em Tradução da Universidade de São Paulo. Alguns dos contos, os canadenses, foram publicados em: TAGNIN, S.E.O. (2002). *Lá do Canadá*, São Paulo: Olavobrás.

[21] http://www.nilc.icmc.usp.br/lacioweb/ (16/08/2004).

Tabela 7. Característica dos córpus paralelos PB-espanhol ainda não alinhados sentencialmente

Córpus / Quantidade	ALCA	Mercosul
Pares de textos paralelos	12	224
Palavras em PB	39.547	362.428
Palavras em espanhol	45.396	387.908
Total de palavras	**84.943**	**750.336**
Sentenças em PB	1.634	16.171
Sentenças em espanhol	1.690	16.164
Nº máximo de alinhamentos 1-1	**1.634**	**16.164**

Assim, o córpus disponível para a extração de regras de tradução disponível para os idiomas PB-espanhol possui cerca de 835.279 palavras (401.975 em PB e 433.304 em espanhol) e 17.798 exemplos de tradução (alinhamentos 1-1).

Esses são os córpus paralelos passíveis de serem usados no projeto ReTraTos para a indução de regras de tradução, porém ainda é necessário analisá-los mais profundamente para verificar se eles são representativos dos problemas que este projeto pretende, inicialmente, atacar: uso incorreto (ou ausência) de preposições, conjunções e tempos verbais.

3.2 Recursos computacionais

Os textos paralelos dos córpus citados na seção anterior passarão por diversas tarefas de pré-processamento antes ou durante a indução das regras de tradução. Duas tarefas de pré-processamento que serão realizadas com certeza são: o alinhamento sentencial e a etiquetação morfossintática. No alinhamento sentencial será utilizado um dos alinhadores resultantes do projeto PESA e, na etiquetação morfossintática, os etiquetadores disponíveis no NILC (para o PB e o inglês) e na Universidade de Alicante (para o espanhol).

De acordo com as avaliações dos alinhadores apresentadas em (Caseli et al., no prelo), o alinhador sentencial de melhor desempenho para o par PB-inglês foi o TCA (*Translation Corpus Aligner*) proposto em (Hofland, 1996). Este alinhador emprega vários critérios de alinhamento para encontrar as correspondências entre as sentenças fonte e alvo, como listas de palavras âncoras, palavras com iniciais maiúsculas (candidatas a nomes próprios), caracteres especiais (! e ?, por exemlo), palavras cognatas e tamanho de sentenças. Tal alinhador apresentou uma precisão média de 95% nos três córpus apresentados na Tabela 5 (científico, jurídico e jornalístico). Além disso, em uma análise informal de seu desempenho no alinhamento de textos paralelos PB-espanhol de gênero jurídico (da documentação oficial

da ALCA), sem a utilização de uma lista de palavras âncoras, o método apresentou um desempenho bem satisfatório exigindo um esforço de correção manual mínimo.

Com relação ao alinhamento lexical, caso seja necessário alinhar os textos dos córpus neste nível de resolução, dois alinhadores lexicais estão disponíveis no NILC e também foram avaliados para os córpus apresentados na Tabela 5, são eles: SIMR (Melamed, 2000) e LWA (Ahrenberg et al., 2002). Segundo (Caseli et al., no prelo), o primeiro apresentou uma melhor precisão (em média, 93,48%) mas uma cobertura muito baixa (em média, 18,37%) enquanto que o segundo se manteve mais estável (57,62% de precisão e 61,6% de cobertura, em média). Assim, caso seja necessário utilizar um alinhador lexical no projeto ReTraTos, este será escolhido de acordo com a necessidade de uma melhor precisão ou cobertura.

Além dos alinhadores sentencial e lexical, outros recursos computacionais que poderão ser utilizados no projeto ReTraTos são os etiquetadores morfossintáticos. Para o PB e o inglês, provavelmente, será utilizado o etiquetador MXPOST. Em uma avaliação de etiquetadores realizada em (Aires & Aluísio, 2001), o etiquetador MXPOST (Ratnaparcki, 1996) foi apontado como o de melhor desempenho para o português brasileiro (cerca de 88% de precisão[22]) em comparação outros etiquetadores – TBL (Brill, 1995), e TreeTagger (Schmid, 1994).

Embora o MXPOST tenha sido apontado como o melhor para o PB é importante citar que essa precisão está longe da margem atingida pelos etiquetadores da língua inglesa, na qual, por exemplo, o MXPOST atingiu mais de 96% de precisão (Ratnaparcki, 1996). Alguns dos fatores que causaram essa baixa precisão no experimento com o PB são: o tamanho do córpus de treinamento – de apenas 100.000 palavras enquanto que experimentos para a língua inglesa são realizados com córpus de 1 a 2 milhões de palavras – e sua qualidade – com erros de etiquetação manual e 30,54% dos textos provenientes do gênero literário que é o tipo de texto com, proporcionalmente, a maior taxa de ambigüidade do córpus, entre outros.

Recentemente, um novo treinamento dos etiquetadores, realizado durante o projeto Lácio-WEB, com um córpus de 1.100.000 palavras, para corrigir as falhas citadas acima, resultou em uma melhora na precisão como é apresentado em detalhes em (Aluísio et al., 2003).

Como se pode perceber, o NILC dispõe de recursos para o PB e o inglês, porém não há, no momento, recursos computacionais (alinhadores e etiquetadores) específicos para o

[22] A precisão, nesse caso, foi calculada como o número de palavras classificadas corretamente dividido pelo número de palavras do arquivo de teste (Aires, 2000).

idioma espanhol. Com relação ao alinhador sentencial está sendo criada uma versão do TCA para alinhar, além de textos paralelos PB-inglês, também textos paralelos PB-espanhol.

Porém, o NILC não dispõe de etiquetadores morfossintáticos nem outras ferramentas de processamento do espanhol. A necessidade de suprir essa deficiência foi uma das razões para que uma parceria (um estágio "sanduíche") com o Departamento de Linguagens e Sistemas Informáticos da Universidade de Alicante (Espanha) fosse apresentada e aceita por órgãos financiadores permitindo, assim, que a doutoranda desenvolva um estágio de um ano em dito estabelecimento de ensino.

O grupo de Alicante dispõe de diversos recursos específicos para o espanhol, dos quais o de maior interesse para o projeto ReTraTos é o etiquetador morfossintático desenvolvido como parte do projeto de implementação do tradutor espanhol-catalão interNOSTRUM[23] que, posteriormente, foi adaptado para a criação do tradutor espanhol-português (o tradutor Universia). Outras informações sobre o estágio de um ano na Universidade de Alicante são apresentados na Seção 3.5.

Diversos outros recursos computacionais desenvolvidos no NILC, ou não, também poderão ser utilizados no projeto ReTraTos, entre eles pode-se citar o léxico computacional Unitex-PB e a base de dado lexical Diadorim, ambos desenvolvidos no NILC como projetos de mestrado cujo objetivo era o tratamento do PB.

3.3 Sobre a escolha das técnicas de indução de regras de tradução

Com os recursos lingüístico-computacionais em mãos e tendo em mente os problemas que se pretende atacar, o próximo passo é identificar quais técnicas são mais adequadas para tratá-los.

Como apresentado no Capítulo 2 e, mais especificamente na Seção 2.2, é grande a variedade de técnicas empregadas pelos métodos no processo de indução de regras de tradução. A escolha da(s) técnica(s) mais adequada(s) será realizada com base nas características dos córpus de treinamento e nos recursos computacionais disponíveis. Com relação aos córpus de treinamento, por exemplo, no córpus PB-espanhol – que é composto por textos jurídicos da ALCA e do Mercosul – há uma ocorrência muito freqüente de certas

[23] No site do interNOSTRUM (http://www.internostrum.com) podem ser obtidas outras informações sobre o etiquetador e há também uma versão para a *web* para os idiomas espanhol e catalão (16/08/2004).

construções resultando em pares de sentenças bastante similares e, portanto, viabilizando a utilização de técnicas baseadas em reconhecimento de padrões (veja Seção 2.2.1). Por outro lado, uma técnica desse tipo seria inviável em córpus como o do jornal "*The New York Times*" no qual os textos não seguem uma estrutura fixa, foram escritos por diferentes jornalistas e abordam assuntos variados (artes, negócios, saúde e internacional) sendo, dessa maneira, pouco provável a existência de construções similares.

Quanto aos recursos computacionais, como apresentado na Seção 3.2, o NILC dispõe de diversas ferramentas para o processamento do PB e algumas para o inglês, porém as ferramentas para o tratamento do espanhol serão obtidas com o intercâmbio na Universidade de Alicante. Sabe-se, no momento, que há a disponibilidade de etiquetadores para os três idiomas, mas não foi possível, ainda, verificar se outras ferramentas (*parsers*, por exemplo) também estão disponíveis. Portanto, a disponibilidade de recursos computacionais é outro fator que deverá ser levado em consideração na escolha da(s) técnica(s).

Outro ponto que poderá ser explorado neste projeto é a utilização da abordagem estatística para auxiliar o processo de indução baseado em sintaxe, praticamente o sentido inverso dos métodos estatísticos baseados em sintaxe, que utilizam informações sintáticas para auxiliar o processo estatístico. Acredita-se que, de maneira semelhante ao que acontece com os métodos estatísticos baseados em sintaxe, o desempenho dos métodos sintáticos que utilizam abordagem estatística poderá ser melhorado.

3.4 Avaliação

Quanto às metodologias de avaliação, neste projeto, pretende-se utilizar os três tipos: avaliação direta, avaliação indireta não-automática e avaliação indireta automática. A avaliação direta deverá ser realizada em menor escala, ou seja, em apenas uma parte da gramática de tradução induzida; enquanto que as avaliações indiretas poderão ser realizadas em maior escala dependendo da disponibilidade de especialistas humanos para avaliar as sentenças alvo ou gerar as sentenças de referências (se for necessária mais de uma) que serão empregadas em uma avaliação indireta automática.

Com relação à avaliação indireta automática das regras, esta poderá ser realizada considerando-se como referência as sentenças dos textos paralelos já existentes e outras que serão geradas por especialistas em tradução. Provavelmente, todas as métricas apresentadas na Seção 2.3.3 serão utilizadas.

Espera-se que, no momento da avaliação das técnicas implementadas para a indução de regras de tradução, já tenham sido publicadas novas avaliações de sistemas que utilizam regras induzidas ou mesmo os que não as utilizam, mas foram projetados especificamente para o PB, como é o caso do EPT-Web. Assim, será possível realizar uma comparação entre o sistema de recombinação das regras gerado no projeto ReTraTos e outros sistemas.

3.5 Estágio no exterior

O grupo de TA da Universidade de Alicante (UA), cujo diretor é o Prof. Dr. Mikel Forcada (futuro supervisor da doutoranda em seu estágio no exterior), desenvolveu um protótipo de um tradutor bidirecional espanhol-português, o Tradutor Universia, baseado no tradutor espanhol-catalão, o interNOSTRUM, também desenvolvido por esse grupo.

O Tradutor Universia baseia-se na estratégia de TA indireta por transferência, que os autores denominam como transferência morfológica avançada, na qual é realizada uma análise morfológica e algumas operações que podem ser consideradas como análise sintática parcial criadas para resolver alguns problemas da transferência morfológica (como ambigüidade lexical categorial, concordância de gênero e número e ordem das palavras). O sistema é formado por 8 módulos (Gilabert-Zarco et al., 2003), dos quais o que será mais beneficiado pelo intercâmbio de experiências e conhecimento decorrentes do estágio da doutoranda é o módulo de transferência estrutural. Tal módulo é responsável pela detecção e tratamento de padrões de palavras que exigem atenção especial devido às divergências gramaticais entre as línguas (como troca de gênero e número, reordenações, mudanças preposicionais, etc.). A detecção desses padrões é feita a partir de um arquivo com regras escritas em uma linguagem de programação simples (Garrido-Alenda & Forcada, 2001), no qual, claramente, as regras geradas por um sistema automático de indução poderão ser incluídas.

Outro fato que demonstra a relevância do estágio é que os resultados obtidos em uma avaliação do sistema português? espanhol apresentada em (Gilabert-Zarco et al., 2003) mostram que a tradução de espanhol para português apresentou um resultado muito melhor do que na direção contrária: a taxa de erro no sentido espanhol-português foi de 3,8% enquanto que, no sentido português-espanhol, essa taxa foi de 10,4%.

O sistema desenvolvido pelo grupo da UA foi analisado em um experimento de pequeno porte (descrito na Seção 1.1) realizado, apenas, com o intuito de tentar identificar as classes de erros mais freqüentes na tradução espanhol-português. Nesse experimento, 20 pares

de sentenças português-espanhol foram traduzidos pelo sistema constatando-se que a maioria dos erros, no sentido português? espanhol, foi decorrente do uso incorreto (ou ausência) de preposições (28,57%), artigos (25,40%) e tempos verbais (7,94%). No sentido espanhol? português os erros de uso incorreto (ou ausência) de preposições e artigos também foram bem representativos, 32,61% e 21,74% do total (100%), respectivamente. Porém, nesse sentido, não se verificou a ocorrência de uso incorreto de tempo verbal; o que pode ser explicado pelo fato da língua fonte (o espanhol) possuir mais tempos verbais do que a língua alvo (o PB), o que pode ter dificultado o mapeamento entre os tempos verbais desses dois idiomas. Além disso, confirmou-se o que já havia sido constatado em (Gilabert-Zarco et al., 2003): a taxa de erro no sentido português-espanhol foi maior (26,98% maior) do que no sentido espanhol-português. A seguir são apresentados exemplos de tradução português-espanhol e espanhol-português gerados por esse sistema.

Português original	Hoje instruímos nossos Ministros Responsáveis por Comércio a iniciarem as negociações sobre a ALCA, como estabelecido na Declaração Ministerial de São José, de março de 1998.
Tradução para o Espanhol	Hoy instruimos nuestros Ministros Responsables por Comercio a inicien las negociaciones sobre la ALCA, como estabelecido en la Declaración Ministerial de São José, de marzo de 1998.
Espanhol original	Hoy instruimos a nuestros Ministros Responsables del Comercio que inicien las negociaciones correspondientes al ALCA de acuerdo com la Declaración Ministerial de San José, de marzo de 1998.
Tradução para o Português	Hoje instruímos a nossos Ministros Responsáveis do Comércio que iniciem as negociações correspondentes ao ALCA de acordo com a Declaração Ministerial de San José, de março de 1998.

Na tradução para o espanhol gerada pelo sistema Universia têm-se, por exemplo, erros de ausência de preposição ("... instruimos **a** nuestros ..."), uso incorreto de preposição ("... Responsables **del** Comercio ...") e concordância de gênero (" ... sobre **el** ALCA ...") como pode ser verificado comparando-se a tradução com a sentença original em espanhol (logo abaixo). No caso da tradução para o português, erros dos mesmos tipos também podem ser encontrados como em "... instruímos a nossos ministros ...", "... Ministros Responsáveis do Comércio ...", etc.

Nesse sentido, o sistema de tradução português-espanhol atualmente em desenvolvimento no grupo da Universidade de Alicante também poderá ser beneficiado com as regras de tradução induzidas, as quais poderão ser facilmente introduzidas no módulo de transferência estrutural. Assim, o estágio na UA compartilha o mesmo objetivo do projeto ReTraTos como um todo: avaliar a aplicação da abordagem de Aprendizado de Máquina na indução automática de regra de tradução, mas, nesse caso, apenas para o par PB-espanhol.

As atividades previstas para este estágio, bem como o cronograma de desenvolvimento do mesmo, são apresentados no próximo capítulo juntamente com a metodologia e o cronograma geral do projeto de doutorado.

Capítulo 4

Metodologia e Cronograma

Neste capítulo são detalhadas as atividades referentes ao projeto ReTraTos, ao curso de doutorado como um todo e ao estágio no exterior. Os cronogramas com as atividades previstas até o fim do doutorado e com as atividades do estágio no exterior também são apresentados. Assim, a Seção 4.1 descreve as atividades do doutorado como um todo enquanto que a Seção 4.2 especifica as atividades previstas para o estágio no exterior.

4.1 Atividades do Projeto ReTraTos

A seguir são apresentadas as atividades previstas para o desenvolvimento do projeto ReTraTos. É importante citar que as atividades seguirão a seqüência apresentada a seguir, sendo que as duas primeiras – criação e adaptação dos recursos lingüístico-computacionais e indução das regras de tradução – serão executadas em duas etapas, uma para cada par de línguas, como mostra o cronograma da Seção 4.1.5.

4.1.1 Criação e adaptação dos recursos lingüístico-computacionais

Antes de se começar a implementar e avaliar as técnicas para indução de regras de tradução é necessário preparar os recursos lingüístico-computacionais que serão utilizados por elas. Assim, a primeira tarefa desta atividade, já concluída, foi a delimitação dos gêneros e domínios dos textos a partir dos quais as regras serão induzidas e a coleta dos textos paralelos que satisfizessem essas restrições. Desse processo resultaram os córpus paralelos PB-inglês e PB-espanhol apresentados na Seção 3.1. Outros textos, de outros gêneros, principalmente para o par PB-espanhol, poderão, ainda, ser coletados para serem usados neste projeto.

Com relação aos recursos computacionais, os textos paralelos passarão por diversas tarefas de pré-processamento como o alinhamento sentencial (e possivelmente lexical) e a etiquetação morfossintática utilizando os recursos computacionais disponíveis no NILC e na Universidade de Alicante (citados na Seção 3.2).

4.1.2 Indução das regras de tradução

A indução das regras de tradução será efetuada em 4 passos:

A. Levantamento bibliográfico e seleção das técnicas de indução de regras de tradução.

Neste passo, um levantamento bibliográfico das principais técnicas de indução de regras de tradução apresentadas na literatura foi realizado com o intuito de possibilitar a seleção daquelas que apresentem melhores desempenhos, segundo a literatura consultada, e que se mostrem mais adequadas aos idiomas e aos recursos lingüístico-computacionais em questão.

B. Implementação das técnicas de indução de regras de tradução selecionadas.

Neste passo, serão implementadas as técnicas de indução de regras de tradução selecionadas no passo anterior.

C. Avaliação das regras de tradução.

Neste passo, será realizada a avaliação das técnicas implementadas, avaliando-se as regras induzidas (avaliação direta) quanto a sua precisão, cobertura, relevância ou outra métrica de interesse.

D. Alteração das técnicas de indução de regras de tradução.

O último passo consiste na adaptação, alteração e combinação das técnicas de indução de regras de tradução na tentativa de melhorar o desempenho das mesmas.

O ciclo implementação-avaliação-alteração (passos B, C e D) se repete até que seja obtida uma precisão satisfatória e/ou estável (sem melhorias significativas em relação às possíveis alterações).

4.1.3 Implementação do sistema de recombinação das regras de tradução

Nesta atividade será implementado um sistema de recombinação das regras de tradução obtidas para ser utilizado na comparação com outros sistemas de TA existentes, atualmente, para o PB.

O sistema que será implementado estará baseado em um processo simples de aplicação e recombinação das regras de tradução induzidas (atividade anterior), para que essas regras possam ser avaliadas sem que o sistema apresente muita influência sobre o desempenho das mesmas.

4.1.4 Avaliação indireta das regras de tradução

A segunda forma de avaliação das regras de tradução que será utilizada neste trabalho é a avaliação indireta (automática e não-automática), por meio do sistema implementado utilizando as regras induzidas.

Enquanto a avaliação direta acontece juntamente com a indução de regras de tradução (apresentada na Seção 4.1.2), a avaliação indireta será a última atividade do projeto de doutorado e ocorrerá concomitantemente à avaliação do sistema de recombinação. Nesse caso, ao invés de se avaliar a cobertura e/ou a precisão das regras, por exemplo, como ocorre na avaliação direta, serão avaliadas as sentenças alvo geradas pelo sistema.

4.1.5 Cronograma das atividades previstas até o término do doutorado

A doutoranda iniciou seu curso de doutorado em março de 2003 e, desde então, todas as atividades previstas para estes primeiros meses foram concluídas com êxito: obtenção de 48 créditos em disciplinas e aprovação nos exames de proficiência em língua inglesa (TOEFL, uma das exigências do curso de pós-graduação do ICMC) e em língua espanhola (Diploma de Espanhol como Língua Estrangeira – DELE – Nível Básico, um dos requisitos para a implementação da bolsa de estágio de doutorado).

O cronograma da Tabela 8 mostra as atividades (já desenvolvidas e previstas) referentes ao projeto ReTraTos e ao curso de doutorado cujo término está previsto para fevereiro de 2007. Como dito anteriormente, as atividades 4.1.1 e 4.1.2 serão executadas em duas etapas, uma para cada par de línguas, como indicado no cronograma.

Tabela 8. Cronograma do doutorado completo

Período / Atividade	2003		2004		2005		2006		2007
	1ºSem	2ºSem	1ºSem	2ºSem	1ºSem	2ºSem	1ºSem	2ºSem	Jan/Fev
Disciplinas									
TOEFL									
DELE									
4.1.1			PB-es			PB-in			
4.1.2			PB-es	PB-es	PB-es	PB-in	PB-in		
4.1.3									
4.1.4									
Qualificação									
Escrita da tese									
Defesa da tese									

4.2 Atividades do estágio no exterior

A seguir são detalhadas as atividades previstas para a realização do estágio na Universidade de Alicante – programado para o período de 05 de outubro de 2004 a 04 de outubro de 2005 – bem como o cronograma de desenvolvimento do mesmo.

Como dito na Seção 3.5, o estágio na UA tem como objetivo avaliar a aplicação da abordagem de Aprendizado de Máquina na indução automática de regra de tradução para o par PB-espanhol. Para tanto, pretende-se selecionar as técnicas de indução de regras de tradução que se mostrem mais adequadas ao PB-espanhol considerando-se o cenário já existente na UA; implementar as técnicas de indução de regras de tradução selecionadas; investigar o desempenho dessas técnicas em córpus específicos para o par PB-espanhol construídos previamente; adaptar as regras geradas para o padrão do sistema da UA; avaliar o desempenho das regras geradas no sistema espanhol; e iniciar a implementação de um sistema de recombinação das regras geradas, possivelmente, utilizando como base o sistema espanhol.

Nas próximas seções, cada uma dessas atividades é especificada em detalhes e, na Seção 4.2.6, é apresentado um cronograma previsto para elas.

4.2.1 Estudo do sistema da UA e seleção das técnicas de indução de regras de tradução para o par PB-espanhol

As técnicas de indução de regras de tradução, já estudadas previamente e apresentadas nessa monografia (Seção 2.2), serão selecionadas considerando-se o desempenho relatado na literatura e a adequação ao par PB-espanhol e aos recursos lingüístico-computacionais disponíveis nos grupos de pesquisa espanhol e brasileiro. Para isso, o protótipo desenvolvido pelo grupo de TA espanhol será estudado e as técnicas de indução que se mostrem mais aplicáveis nesse ambiente serão selecionadas para implementação.

4.2.2 Ciclo implementação-avaliação-alteração das técnicas de indução de regras de tradução para o par PB-espanhol

As técnicas de indução de regras de tradução selecionadas serão implementadas, avaliadas (avaliação direta) e, possivelmente, alteradas e aperfeiçoadas. Nessa atividade serão utilizados os recursos lingüístico-computacionais para o PB (disponíveis no NILC) e o espanhol (disponíveis no grupo de pesquisa na UA).

4.2.3 Avaliação indireta das regras geradas

As regras induzidas serão adaptadas para o padrão do sistema espanhol e serão adicionadas ao protótipo para que possam ser avaliadas (avaliação indireta) em uma aplicação prática.

4.2.4 Início da implementação de um sistema de recombinação das regras induzidas

A experiência do grupo de TA da UA será utilizada na especificação/implementação de um sistema de recombinação das regras de tradução induzidas automaticamente. Porém, devido à duração do estágio no exterior e à complexidade de implementação de tal sistema, ao fim do período de estadia na UA, é provável que o sistema ainda não tenha sido concluído.

4.2.5 Disciplinas

Durante seu estágio na UA, a doutoranda cursará algumas disciplinas de pós-graduação consideradas de extrema relevância para seu projeto de doutorado[24]. São elas:

a. Tradução Automática – Fundamentos e Aplicações (Nov/2004 a Fev/2005);

b. Técnicas Avançadas de Tradução (Mar/2005 a Jun/2005);

c. Marcação de Textos (Fev/2005 a Jun/2005).

4.2.6 Cronograma das atividades previstas para o estágio no exterior

O cronograma apresentado na Tabela 9 mostra as atividades referentes ao estágio no exterior previsto para o período de 05 de outubro de 2004 a 04 de outubro de 2005.

Tabela 9. Cronograma do estágio no exterior

Período / Atividade	2004		2005			
	Out/Nov	Dez/Jan	Fev/Mar	Abr/Mai	Jun/Jul	Ago/Set
4.2.1	■	■				
4.2.2		■	■	■		
4.2.3				■	■	
4.2.4					■	■
4.2.5.a		■	■	■		
4.2.5.b				■	■	
4.2.5.c			■	■	■	

[24] Outras informações sobre o programa de doutorado do DLSI podem ser obtidas em: http://www.dlsi.ua.es/docente/doctorado/ (16/08/2004).

Além das atividades descritas, outras, indispensáveis para o desenvolvimento do projeto também estarão presentes como: produção de relatórios técnicos e artigos científicos; implementação de ferramentas auxiliares; especificação da metodologia de avaliação, etc.

Capítulo 5

Considerações Finais

A Tradução Automática é uma das principais áreas de PLN e, há muito tempo, desperta o interesse de pesquisadores de lingüística e de computação no sentido de tentar melhorar o desempenho dos sistemas existentes ou propor novas tecnologias, com o intuito de tornar a troca de informações entre línguas distintas o menos problemático possível.

Nesse sentido, entre as abordagens propostas com o intuito de superar a barreira lingüística está a que visa a aquisição de conhecimento, de maneira automática, a partir de base de exemplos de tradução, ou EBMT (*Example Based Machine Translation*). Nos últimos anos, diversas propostas envolvendo essa abordagem, juntamente com a estratégia de tradução automática indireta por transferência, têm surgido na comunidade científica. Porém, não se tem conhecimento, até o momento, de nenhum trabalho deste tipo envolvendo o português do Brasil.

Assim, o projeto ReTraTos – que visa a indução de regras de tradução a partir de textos paralelos PB-inglês e PB-espanhol alinhados sentencialmente – surge como uma tentativa de superar alguns dos problemas encontrados na tradução de/para o PB (principalmente no que diz respeito ao uso incorreto ou ausência de preposições, artigos e tempos verbais) utilizando, para isso, técnicas de Aprendizado de Máquina e EBMT.

Com relação às pesquisas na área de indução de regras de tradução, é importante ressaltar que esta área é, ainda, muito recente e que, portanto, há muito a ser explorado em termos das técnicas e abordagens empregadas.

Além disso, as precisões apresentadas até o momento para os métodos de indução de regras de tradução mostram que a área é promissora e há muito a se avançar. Porém, vários métodos não foram completamente avaliados, nem um critério único de avaliação foi determinado e, por isso, não se pode afirmar quais são exatamente os melhores índices a se perseguir nessa área.

Referências Bibliográficas

AIRES, R.V.X. (2000). *Implementação, Adaptação, Combinação e Avaliação de Etiquetadores para o português do Brasil*. Dissertação (mestrado) – Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo, São Carlos. 154p.

AIRES, R.V.X.; ALUÍSIO, S.M. (2001). Implementação, Adaptação, Combinação e Avaliação de Etiquetadores para o Português do Brasil. In: *VI Workshop de Teses e Dissertações defendidas do ICMC/USP*, São Carlos. p.243-257.

AHRENBERG, L; ANDERSON, M.; MERKEL, M. (2002). A system for incremental and interactive word linking. In: *Third International Conference on Language Resources and Evaluation* (LREC 2002), Las Palmas. p.485-490.

ALUÍSIO, S.M.; PELIZZONI, J.M.; MARCHI, A.R.; OLIVEIRA, L.H.; MANENTI, R.; MARQUIVAFÁVEL, V. (2003). An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In: *Proceedings of PROPOR´2003, Lecture Notes on Artificial Intelligence*, v.1, Springer-Verlag, Faro, Portugal. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/publicacoes.htm>. Acesso em: 16/08/2004.

BOSTRÖM, H. (2000). Induction of recursive transfer rules. In: CUSSENS, J.; DEROSKI, S. (eds.), *Learning Language in Logic, Lecture Notes in Computer Science*, v.1925, Springer-Verlag Heidelberg. p.237-246. Disponível em: <http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Bostr=ouml=m:Henrik.html>. Acesso em: 16/08/2004.

BRILL, E. (1995). Transformation-based error-driven learning of natural language: a case study in part of speech tagging. *Computational Linguistics*, v.21, n.4, p.543-565. Disponível em: <http://www.cs.jhu.edu/~brill/papers.html>. Acesso em: 16/08/2004.

BROWN, R.D. (2001). Transfer-rule induction for example-based translation. In: *Proceedings of the MT Summit VII Workshop on Example-Based Machine Translation*, Santiago de Compostela, Spain. p.1-11. Disponível em: <http://www-2.cs.cmu.edu/~ralf/papers.html>. Acesso em: 16/08/2004.

CARBONELL, J.; PROBST, K.; PETERSON, E.; MONSON, C.; LAVIE, A.; BROWN, R; LEVIN, L. (2002). Automatic rule learning for resource-limited MT. In: *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas* (AMTA 2002), Tiburon, California. p.1-10. Disponível em: <http://www-2.cs.cmu.edu/~ralf/papers.html>. Acesso em: 16/08/2004.

CARL, M. (2001). Inducing probabilistic invertible translation grammars from aligned texts. In: *Proceedings of CoNLL-2001*, Toulouse, France. p.145-151.

CASELI, H.M.; SILVA, A.M.P.; NUNES, M.G.V. (no prelo). Evaluation of methods for sentence and lexical alignment of Brazilian Portuguese and English parallel texts. In: *Proceedings of the 7th Symposium on Artificial Inteligence (SBIA)*, São Luís.

CICEKLI, I.; GÜVENIR, H.A. (1996). Learning translation rules from a bilingual córpus. In: *Proceedings of the 2nd International Conference on New Methods in Language Processing (NeMLaP-2)*, Ankara, Turkey. p.90-97.

CICEKLI, I.; GÜVENIR, H.A. (2003). Learning translation templates from bilingual translation examples. In: CARL, M.; WAY, A. (eds.), *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Boston. p.247-278.

DODDINGTON, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *Proceedings of ARPA Workshop on Human Language Technology*, San Diego. p.128-132.

DORR, B.J.; JORDAN, P.W.; BENOIT, J.W. (1999). A Survey of Current Research in Machine Translation. In: ZELKOWITZ, M. (ed.), *Advances in Computers*, v.49, Academic Press, London. p.1-68.

FINCH, A.; AKIBA, Y.; SUMITA, E. (2004). How does automatic machine translation evaluation correlate with human scoring as the number of recerence translation increases?. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*. p.2019-2022.

FOSSEY, M.F.; PEDROLONGO, T.; MARTINS, R.T.; NUNES, M.G.V. (2004). Análise comparativa de tradutores automáticos inglês-português. *Série de relatórios do NILC*, NILC-RT-04-04, São Carlos. 18p.

FURUSE, O.; IIDA, H. (1992). Cooperation between transfer and analysis in example-based framework. In: *Proceedings of COLING-92*, Nantes, France. p.645-651.

GALLEY, M.; HOPKINS, M.; KNIGHT, K.; MARCU, D. (2004). What's in a translation rule? In: *Proceedings of the NAACL – HLT*.

GARRIDO-ALENDA, A.; FORCADA, M. L. (2001). Morphtrans: un lenguaje y un compilador para especificar y generar módulos de transferencia morfológica para sistemas de tradución automática. In: *Procesamiento del Lenguaje Natural*, v. 27. p.157-164.

GILABERT-ZARCO, P.; HERRERO-VICENTE, J.; ORTIZ-ROJAS, S.; PERTUSA-IBÁÑEZ, A.; RAMÍREZ-SANCHEZ, G.; SÁNCHEZ-MARTÍNEZ, F.; SAMPER-ASENSIO, M.; SCALCO, M. A.; FORCADA, M. L. (2003). Construcción rápida de un sistema de traducción automática español? portugués partiendo de un sistema español? catalán. In: *Procesamiento del Lenguaje Natural*, XIX Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, Alcalá de Henares, España. p.279-284. Disponível em: < http://www.dlsi.ua.es/~fsanchez/publ_es.html>. Acesso em: 16/08/2004.

GILDEA, D. (2003). Loosely tree-based alignment for machine translation. In: *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL-03)*, Sapporo, Japan. p.80-87.

GÜVENIR, H.A.; CICEKLI, I. (1998). Learning translation templates from examples. *Information Systems*, v.23. n.6, p.353-363. Disponível em: <http://citeseer.ist.psu.edu/7130.html>. Acesso em: 16/08/2004.

HOFLAND, K. (1996). A program for aligning English and Norwegian sentences. In: HOCKEY, S., IDE, N., PERISSINOTTO, G. (eds.), *Research in Humanities Computing*, Oxford, Oxford University Press. p.165-178.

LAVIE, A.; PROBST, K.; PETERSON, E.; VOGEL, S.; LEVIN, L.; FONT-LLITJOS, A.; CARBONELL, J (2004). A trainable transfer-based machine translation approach for language with limited resources. In: *Proceedings of Workshop of the European Association for Machine Translation* (EAMT – 2004), Valletta, Malta. Disponível em: <http://www-2.cs.cmu.edu/afs/cs.cmu.edu/user/alavie/www/papers/EAMT-XFER-Apr04.pdf>. Acesso em: 16/08/2004.

LAVOIE, B.; WHITE, M.; KORELSKY, T. (2001). Inducing lexico-structural transfer rules from parsed bi-texts. In: *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France. p.17-24.

LAVOIE, B.; WHITE, M.; AND KORELSKY, T. (2002). Learning domain-specific transfer rules: an experiment with Korean to English translation. In: *Proceedings of the COLING 2002 Workshop on Machine Translation in Asia*, Taipei, Taiwan. p.60-66.

MANNING, C.D.; SCHUTZE, H. (1999). *Foundations of statistical natural language processing*, MIT Press. p.172-175.

MARTINS, R.T.; NUNES, M.G.V. (no prelo). Tradução Automática: o que é e como se faz. In: LIMA, V.L.S.; VIEIRA, R. (eds.), *Engenharia de linguagem natural: introdução ao tratamento computacional da língua*, Editora Manole.

MATSUMOTO, Y.; ISHIMOTO, H.; UTSURO, T. (1993). Structural matching of parallel texts. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, Columbus, Ohio. p.23-30.

MCTAIT, K.; TRUJILLO, A. (1999). A language-neutral sparse-data algorithm for extracting translation patterns. In: *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, Chester, England. p.98-108.

MCTAIT, K. (2003). Translation patterns, linguistic knowledge and complexity in an approach to EBMT. In: CARL, M.; WAY, A. (eds.) *Recent Advances in Example-Based Machine Translation*, Amsterdam: Kluwer Academic Press, Dordrecht, The Netherlands. Disponível em: <http://www.limsi.fr/Individu/mctait/Publications/pubs.html>. Acesso em: 16/08/2004.

MELAMED, I.D. (2000). Patter recognition for mapping bitext correspondence. In: VÉRONIS, J. (ed.). *Parallel text processing*, s.l.: Kluwer Academic Publishers. p.25-47.

MELAMED, I.D.; GREEN, R.; TURIAN, J.P. (2003). Precision and recall of machine translation. In: P*roceedings of NAACL/HLT 2003*, Edmonton, Canada. p.61-63.

MENEZES, A.; RICHARDSON, S.D. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In: *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France. p.39-46.

MEYERS, A.; YANGARBER, R.; GRISHMAN, R. (1996). Alignment of shared forests for bilingual corpora. In: *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*, Copenhagen, Denmark. p.460-465.

MEYERS, A.; YANGARBER, R.; GRISHMAN, R.; MACLEOD, C.; MORENO-SANDOVAL, A. (1998). Deriving transfer rules from dominance-preserving alignments. In: *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Meeting of Association for Computational Linguistics (COLING-ACL98)*, Montréal, Quebec. p.843-847.

MEYERS, A.; KOSAKA, M.; GRISHMAN, R. (2000). Chart-based transfer rule application in machine translation. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbruken, Germany. p.537-543.

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: NUNES, M.G.V. (ed.), *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, São Paulo: ICMC/USP. Atibaia. p.85-92. Disponível em: <http://nilc.icmc.sc.usp.br/download/criticalanalysis.zip>. Acesso em: 16/08/2004.

ÖZ, Z.; CICEKLI, I. (1998). Ordering translation templates by assigning confidence factors. *Lecture Notes in Computer Science*, v.1529, Springer-Verlag. p.51-61.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W. (2002). BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics* (ACL-2002). Philadelphia, PA. p.311-318.

PROBST, K. (2002). Semi-automatic learning of transfer rules for machine translation of low-density languages. In: *Proceedings of the Seventh ESSLLI Student Session*, Trento, Italy. Disponível em: <http://www-2.cs.cmu.edu/~kathrin/>. Acesso em: 16/08/2004.

RATNAPARKHI, A. (1996). A Maximum Entropy Part-of-Speech Tagger. In: *Proceedings of the First Empirical Methods in Natural Language Processing Conference*. Disponível em: <http://www.cis.upenn.edu/~adwait/statnlp.html>. Acesso em: 16/08/2004.

RICHARDSON, S.D.; DOLAN, W.B.; MENEZES, A.; CORSTON-OLIVER, M. (2001). Overcoming the customization bottleneck using example-based MT. In: *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France. p.9-16.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. p.44-49.

SOMERS, H. (1999). Review article: example-based machine translation. *Machine Translation*, v.14, Kluwer Academic Publishers, Netherlands. p.113-157.

TURIAN, J.P.; SHEN, L.; MELAMED, I.D. (2003). Evaluation of machine translation and its evaluation. In: *MT Summit IX*, New Orleans, USA. p.386-393.

YAMADA, K.; KNIGHT, K. (2001). A syntax-based statistical translation model. In: *Proceedings of the 39th Annual Conference of the Association for Computational Linguistics*, Toulose, France. p.6-11.