ALEXANDRE AGUSTINI

EXPERIÊNCIA DE UTILIZAÇÃO DO FORMALISMO "GRAMÁTICAS SÍNCRONAS DE ADJUNÇÃO DE ÁRVORES" PARA CONSTRUÇÃO DE UM MÓDULO DE TRANSFERÊNCIA ESTRUTURAL

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre.
Curso de Mestrado em Informática,
Instituto de Informática,
Pontifícia Universidade Católica do Rio Grande do Sul.
Orientadora: Dra. Vera Lúcia Strube de Lima

PORTO ALEGRE

1997

Dados Internacionais de Catalogação na Publicação (CIP)

A284e Agustini, Alexandre

Experiência de Utilização do Formalismo "gramáticas síncronas de adjunção de árvores" para construção de um módulo de transferência estrutural / Alexandre Agustini. — Porto Alegre, 1997.

110 fl.

Diss. (Mestrado) - Faculdade de Informática, PUCRS.

1.Informática - Processamento de Linguagem Natural 2.Lingüística Computacional - Tradução Automática 3.C (Linguagem de Programação) 4.Tradução Automática - C (Linguagem de Programação) I.Título.

CDD: 006.35

CDU: 681.3.06
681.3.068

Índices para o Catálogo Sistemático:

Informática - Processamento de Linguagem Natural	006.35
Lingüística Computacional - Tradução Automática	006.35
C (Linguagem de Programação)	681.3.06
Tradução Automática - C (Linguagem de Programação)	681.3.068

Bibliotecária Responsável
Alessandra Pinto Fagundes
CRB10/1244

"Computers are incredibly fast, accurate and stupid,
humans are incredibly slow, inaccurate and brilliant;
together they are powerful beyond imagination".

Albert Einstein

AGRADECIMENTOS

A todos aqueles que de uma forma ou outra estiveram presentes durante o desenvolvimento deste trabalho, com suas críticas e sugestões ou mesmo apenas por estarem presentes.

Um agradecimento muito especial à Profa. Vera Lúcia Strube de Lima pela orientação, amizade e paciência que despendeu durante todo o desenvolvimento do trabalho.

Aos colegas do grupo de linguagem natural sempre dispostos a colaborar, principalmente ao Ivan, colega inseparável de algumas decisões e da implementação realizada.

À Flávia, companheira e incentivadora de todas as horas e que durante este período foi capaz de me dar a maior alegria imaginável... a nossa *Juju*, cujo sorriso e alegria motivam nossas vidas.

SUMÁRIO

LISTA DE ABREVIATURAS

Adj	Adjetivo
Adv	Advérbio
EBNF	Forma Normal de Backus Estendida
	(Extended Backus Naur Form)
GLC	Gramática Livre de Contexto
GSC	Gramática Sensível ao Contexto
LA	Língua-alvo
LF	Língua-fonte
LTAG	Gramática Lexicalizada de Adjunção de Árvores (*Lexicalized Tree Adjoining Grammar*)
N	Nome
NP	Noum Phrase
Prep	Preposição
S	Sentença, Sentence
SRS	Sistema de Reescrita Síncrono
STAG	Gramáticas Síncronas de Adjunção de Árvores (*Synchronous Tree Adjoining Grammars*)
SV	Sintagma Verbal
TA	Tradução automática
TAG	Gramática de Adjunção de Árvores (*Tree Adjoining Grammars*)
V	Verbo
VP	Verb Phrase
VTD	Verbo Transitivo Direto

LISTA DE FIGURAS

RESUMO

A automatização da tradução tem sido um desafio constante para lingüistas e cientistas da computação nas últimas décadas. Neste período, muitos avanços foram alcançados, porém os resultados ainda não são os esperados.

É apresentado, neste trabalho, um estudo sobre a área de tradução automática, focalizando inicialmente os principais métodos utilizados na construção de sistemas automatizados de tradução: métodos diretos e métodos baseados nos conceitos de interlíngua e *transfer*.

O trabalho descreve as Gramáticas Síncronas de Adjunção de Árvores como formalismo para projeto de um módulo de transferência estrutural, que é o componente principal de sistemas de tradução automática baseados no método *transfer*.

O módulo de transferência realiza o mapeamento das discrepâncias existentes entre a representação estrutural do texto na língua-fonte e a representação correspondente na língua-alvo. Um estudo, a partir de um corpus da área econômica, é apresentado visando a definição de um conjunto de divergências estruturais existentes na tradução entre as línguas portuguesa e inglesa.

Para validação do modelo proposto, é apresentado o protótipo de uma ferramenta que realiza as transformações estruturais observadas no corpus empregado, utilizando os conceitos de Gramáticas Síncronas de Adjunção de Árvores.

ABSTRACT

TITLE: AN EXPERIMENT ON THE USE OF SYNCHRONOUS TREE ADJOINING GRAMMAR FORMALISM FOR THE CONSTRUCTION OF A STRUCTURAL TRANSFER MODULE.

Machine translation has been a challenge for linguists and computer scientists over the last decades. During this period, plenty of progress was accomplished, though the results are not the expected ones.

In this investigation, we present a study on automatic translation, starting with a review of the main methods used for the construction of automated translation systems: direct methods and methods based on interlingua and transfer concepts.

The work describes the use of the Synchronous Tree Adjoining Grammars (STAGs) formalism for the design of a structural transfer module, which is the main component of transfer-based systems.

The transfer module establishes the correspondences between the structural representation of a sentence in the source-language and the one in the target-language. A study on a corpus on economics was developed in order to define structural divergences for the translation between the Portuguese and English languages.

A prototype that performs the structural transformations found in the corpus, based on the STAGs concepts, was developed to validate the proposed model.

1. INTRODUÇÃO

1.1 Processamento da Língua Natural e Tradução Automática

Em momento algum da História houve uma necessidade tão grande de superar as barreiras lingüísticas que dividem os povos como agora. Novos mercados em todo o mundo têm criado uma demanda cada vez maior por suporte lingüístico. Somente na comunicação entre as nove línguas oficiais da Comunidade Européia, existe a necessidade de uma tradução em 72 diferentes sentidos (Vasconcellos 93).

Desta forma, faz-se necessário dispor de meios alternativos que propiciem um intercâmbio de informações mais rápido, ágil e eficiente, para dar conta do ritmo acelerado que o homem vem impondo aos avanços tecnológicos e à produção de conhecimento nas suas áreas de atuação.

Sistemas de Tradução Automática (TA) surgiram na década de 50 e constituem um dos maiores campos de pesquisa na área da Lingüística Computacional. Os primeiros trabalhos foram desenvolvidos para manipular traduções palavra-a-palavra, através de simples pesquisa a dicionários bilíngües (Niremburg et al. 92).

Embora esta abordagem seja simples, apresentou alguns resultados iniciais expressivos. Porém, logo constatou-se uma grande quantidade de problemas ao traduzir sem a utilização de conhecimentos que complementassem a pesquisa em dicionários. A tradução palavra-a-palavra não é suficiente, uma vez que as línguas apresentam várias discrepâncias, tanto em níveis léxico e sintático quanto em nível semântico.

A pesquisa na área da TA desenvolveu-se razoavelmente durante os anos 50. Porém, nessa época, tanto a Lingüística Teórica (teorias de gramáticas, semântica, pragmática e análise do discurso) quanto a Lingüística Computacional (modelagem computacional para análise sintática, interpretação semântica e geração de textos) eram incipientes. Além disto, o *hardware* se constituía em um grande limitador na tentativa de obter traduções satisfatórias. Durante a década de 60, as pesquisas praticamente desapareceram, principalmente devido a

estes problemas teóricos e de *hardware* que pareciam insolúveis com os recursos disponíveis (Dorr 93; Kopper 95).

Atualmente, há um interesse cada vez maior em pesquisas em TA. Muitos projetos estão em andamento, impulsionados por fatores tais como a disponibilidade de estudos teóricos nas áreas de morfologia, sintaxe e semântica, além do grande desenvolvimento dos recursos de *hardware*.

Duas abordagens básicas têm sido utilizadas: métodos baseados em uma representação semântica abstrata, e métodos baseados em um módulo de transferência estrutural (Hutchins & Somers 92). Os primeiros utilizam o conceito de "universais lingüísticos", uma representação suficientemente abstrata para representar todas as informações associadas a um texto, independente da língua em que foi escrito ou para a qual será traduzido. A análise do texto na língua-fonte (LF) e a geração na língua-alvo (LA) são realizadas por processos independentes. Os métodos baseados em transferência estrutural, por outro lado, utilizam representações que permitem mapear todas as informações do texto-fonte necessárias para realizar a tradução em uma língua específica. Um módulo de transferência, intermediário, é responsável por realizar as transformações na representação do texto, na LF, para uma representação estrutural na LA.

1.2 Objetivos deste trabalho

Coulthard (Coulthard 91) cita três questões básicas que devem ser resolvidas durante um processo de tradução: ambigüidade, referências anafóricas e divergências. Estas questões envolvem vários problemas que são basicamente lingüísticos, o que explica o pouco avanço computacional obtido na solução de cada uma delas.

Tomando-se especificamente as divergências, observamos que Arnold (Arnold et al. 94) as classifica em divergências léxicas e divergências sintáticas. As primeiras têm a ver com a forma como a língua classifica as palavras, conceitos escolhidos para expressar por palavras simples e aquelas que simplesmente não são lexicalizadas. As segundas ocorrem porque as línguas utilizam diferentes estruturas para o mesmo propósito e, em alguns casos, a mesma estrutura para diferentes propósitos.

Este trabalho tem por objetivo investigar a possibilidade da construção de um módulo de transferência utilizando o formalismo Gramáticas Síncronas de Adjunção de Árvores (*Synchronous Tree Adjoining Grammars*, STAGs), especificamente no que se refere à capacidade deste formalismo em modelar divergências sintáticas existentes na tradução entre as línguas portuguesa e inglesa. O conjunto de divergências a ser tratado foi definido a partir de um corpus[1] real com sentenças da área econômico-financeira.

1.3 Metodologia utilizada

O estudo do processamento da língua natural, enfatizando a complexidade de tratamento computacional em níveis léxico, sintático e semântico (Agustini 95a), e das principais abordagens utilizadas para a implementação de sistemas de tradução automática (Agustini 95b), possibilitou o embasamento necessário para a definição da proposta de trabalho a ser desenvolvido.

Para elaboração da proposta, foi analisada a ferramenta para análise sintática desenvolvida por Kipper (Kipper 94), baseada no formalismo Gramática de Adjunção de Árvores (*Tree Adjoining Grammars*, TAGs), disponível na PUCRS, e foram observadas algumas comparações deste formalismo com outros disponíveis na literatura. Em uma segunda etapa, estudou-se o formalismo STAGs, uma extensão ao formalismo TAGs que permite a execução síncrona de duas especificações TAGs. Como a experimentação deste formalismo para a construção de sistemas de tradução automática se restringe a estudos para a tradução entre as línguas francesa e inglesa (Abeillé 92a) e entre as línguas coreana e inglesa (Egedi & Palmer 94), decidiu-se pela avaliação do potencial do formalismo para a construção de um módulo de transferência entre as línguas portuguesa e inglesa.

A figura 1.1 resume a metodologia empregada para o desenvolvimento do trabalho. Em um primeiro momento, buscou-se a obtenção de um corpus que permitisse estudar fenômenos lingüísticos que ocorrem durante a tradução da língua portuguesa para a língua inglesa. Com o estudo do corpus, foram definidas a gramática virtual[2] observada neste corpus e o conjunto de problemas estruturais observados durante a tradução humana realizada.

---

[1] Denomina-se corpus a uma coleção de textos escritos em uma determinada língua.

[2] O termo "gramática virtual" refere-se ao subconjunto das estruturas sintáticas mínimas necessárias para o reconhecimento sintático do texto de entrada e geração das estruturas de saída, neste corpus.

Em um segundo momento, foi definido um módulo de transferência baseado nas características observadas no corpus, que foi utilizado para implementação de um protótipo e avaliação do formalismo STAGs. Entende-se por avaliação do formalismo a verificação da possibilidade de modelar o conjunto de casos encontrados no corpus estudado.

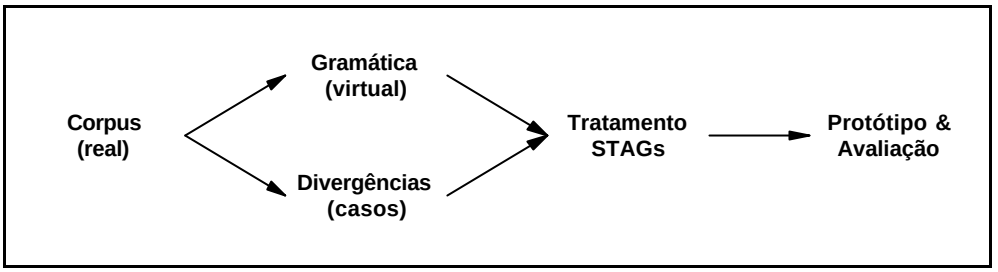


Figura 1.1 - Metodologia empregada

1.4 Estrutura do texto da dissertação

A dissertação está estruturada de forma a apresentar o problema da tradução automática sob o aspecto lingüístico, apresentado no capítulo 2, e sob o aspecto computacional, apresentado no capítulo 3.

O capítulo 4 descreve o formalismo STAGs, que será utilizado como ferramental teórico para implementação do modelo *transfer* (conforme capítulo 3) de tradução, visando a resolução dos problemas estruturais definidos no capítulo 2.

No capítulo 5 é delimitado o corpus de divergências a ser utilizado para experimentação e proposta uma arquitetura para a construção de um módulo de transferência. É descrita, ainda, a implementação do protótipo, e são apresentados exemplos de utilização.

O capítulo 6 reúne as conclusões e perspectivas para a continuidade do trabalho desenvolvido.

2. A TRADUÇÃO AUTOMÁTICA E O PROBLEMA DA DIVERGÊNCIA

2.1 Lingüística e Tradução Automática

Segundo Ladmiral (Ladmiral 79), a tradução é um caso particular de convergência lingüística: no sentido mais amplo, ela designa qualquer forma de *intermediação lingüística* que permita transmitir informação entre locutores de línguas diferentes. A tradução faz passar uma mensagem em uma língua de partida, ou língua-*fonte* (LF), para uma língua de chegada, ou língua-*alvo* (LA).

A *tradução* designa, ao mesmo tempo, um sentido dinâmico dado pela atividade do tradutor (no escopo deste trabalho, a execução do sistema de tradução automática), e um sentido estático dado pelo resultado desta atividade, ou seja, o texto-alvo em si.

Um dos princípios básicos da lingüística é que todas as línguas podem expressar tudo o que os falantes desejam comunicar. No entanto, as línguas diferem em termos do que é fácil ou difícil de expressar, do que é essencial ou opcional e do que é expresso lexical ou gramaticalmente. Por tais fatores é que a tradução entre línguas cognatas é normalmente mais simples. Há três problemas lingüísticos básicos que devem ser resolvidos durante um processo de tradução: vocabulário, gramática e simbologia.

2.1.1 Vocabulário

O ato de nomear é um problema dos falantes de todas as línguas. Em primeiro lugar, percebe-se mais distinções no mundo do que é possível diferenciar lexicalmente. Um exemplo disto é o das cores: pessoas com visão normal são capazes de diferenciar cerca de 6,5 milhões de cores diferentes; a maioria das línguas, porém, contenta-se lingüisticamente com uma dúzia de cores e raramente são utilizados mais de trinta nomes para referenciá-las.

Em segundo lugar, as diferenças se situam freqüentemente em um contínuo e, portanto, cabe à comunidade decidir a necessidade ou não de distinções lingüísticas. Por

exemplo, o espectro de cores distintas entre verde e azul em português é tratado como uma única cor em irlandês. Este fenômeno ocorre em todo o vocabulário de uma língua.

Em terceiro lugar, algumas línguas possuem vocabulários maiores que outras, e possuem áreas particulares de significado que são altamente lexicalizadas; por exemplo: o árabe da África do Norte tem mais de 100 palavras para os diferentes tipos de camelo, enquanto os esquimós têm mais de 100 palavras para a neve.

Finalmente, há itens ou conceitos específicos de uma cultura para os quais somente a língua desta cultura terá denominações.

As conseqüências destes fatores para a tradução, em nível de palavra, podem ser resumidas da seguinte forma:

- freqüentemente há uma palavra que tem a mesma amplitude de referências na outra língua. Por exemplo: doutor = *doctor*;
- haverá muitos casos, no entanto, em que uma palavra terá mais de um equivalente. Por exemplo: cravo = *nail, carnation, clove*;
- haverá casos em que duas ou mais palavras compartilham o mesmo equivalente. Por exemplo: motor/reator = *engine*;
- casos em que não existe equivalente na outra língua. Por exemplo: empreiteiras = ∅.

Durante a tradução cada um destes casos deverá ser tratado adequadamente, o que poderá corresponder a uma simples escolha de palavras ou exigir uma análise contextual mais complexa.

Além destes problemas, o vocabulário pode ser constituído de expressões fixas como *a esta altura do campeonato* ou *a língua não tem osso* e expressões idiomáticas, tais como *quebrar o galho* ou a expressão inglesa: *kick the bucket*. No processo de tradução é necessário decidir se a expressão é ou não parte essencial do texto, e se deve ou não ser revertida para um equivalente na LA.

2.1.2 Gramática

Ao examinar a gramática nota-se que os problemas encontrados são, em muito, equivalentes aos do vocabulário. Distinções feitas em uma língua não são feitas em outra, uma informação essencial em uma língua não se encontra em textos de outra.

Um exemplo destas distinções é o *gênero*. Todas as palavras em português têm gênero gramatical marcado e isto pode ou não corresponder ao sexo do referente em um caso determinado; poucas palavras em inglês têm gênero marcado, mas quando isto ocorre sempre é relacionado ao sexo do referente. Um exemplo disto é o título do conto de Hemingway, *Cat in the rain.* No texto narrativo o gato é considerado, em momentos diferentes, como macho, fêmea ou de sexo indeterminado. Esta ambigüidade é possível em inglês, mas ao traduzir o título para o português, o tradutor deverá decidir a classificação do animal em masculino ou feminino, ou prover outro processo literário.

Um outro exemplo de distinção gramatical é o tratamento das formas verbais. Em inglês a forma verbal *she liked* é ambígua, pois o tempo passado não diferencia uma única ação de um estado ou ação habitual. Em português esta distinção é feita pela escolha do tempo verbal: perfeito ou imperfeito. Assim, ao escolher-se entre *ela gostou* ou *ela gostava* para a tradução, estará sendo eliminada a ambigüidade e talvez interpretando o texto de forma incorreta.

2.1.3 Simbologismo

Um último problema é o fato de algumas palavras serem utilizadas não só como referentes, mas também como símbolos. A propriedade do símbolo pode depender de aspectos culturais ou gramaticais. Por exemplo, em um conto cujo título é *The Fox* (A raposa) o substantivo *fox*, que representa o personagem principal masculino, é uma referência metafórica a um homem, com conotações de esperteza, ligeireza e malícia. A palavra portuguesa, de gênero feminino, certamente não seria própria durante o processo de tradução.

Para a realização de um sistema de tradução completamente automatizado, todos estes problemas deverão ser tratados, constituindo em alguns casos um grande desafio computacional.

2.2 O Fenômeno da Ambigüidade

Um dos problemas mais sérios encontrados na tradução automática de textos é a dificuldade de equipar o sistema com uma representação adequada do conhecimento de mundo. O tradutor humano, quando transpõe um texto de uma língua para outra, aciona este conhecimento para resolver inúmeras ambigüidades do texto, numa operação tão rápida que parece ocorrer abaixo do nível de consciência (Leffa 95).

O problema da ambigüidade, na tradução automática, surge quando a polissemia de um termo não pode ser simetricamente transposta de uma língua para outra. Isto acontece, por exemplo, quando o número de acepções de uma palavra no dicionário é menor na LF que na LA, gerando uma ambigüidade translacional, ou seja, que surge apenas no momento da tradução (Hutchins 92). A palavra inglesa *wall*, por exemplo, pode ser traduzida para o português, como *parede* ou *muro*, obrigando o sistema a decidir entre uma ou outra acepção.

A ambigüidade pode ocorrer em diferentes níveis de complexidade, que podem ser classificados de diferentes maneiras. Uma possível classificação é proposta em três níveis: léxico, sintático e pragmático (Leffa 95, Hutchins & Somers 92).

Em nível léxico, a ambigüidade pode ocorrer de forma categórica ou semântica. A ambigüidade categórica envolve apenas a classificação gramatical da palavra. A palavra *fala*, no exemplo abaixo, pode ser classificada como um substantivo (1a) ou um verbo (1b). Este tipo de ambigüidade é de resolução relativamente fácil, pois normalmente um processo de análise sintática é capaz de determinar a categoria apropriada da palavra.

(1a) Ele perdeu a *fala*. (fala=substantivo)

(1b) Ele *fala* bem. (fala=verbo)

A ambigüidade semântica do léxico envolve homógrafos e casos de polissemia, podendo ou não ser da mesma classe gramatical. A palavra *canto*, abaixo, pode ser um lugar (2a) ou uma música (2b). No caso de serem de classes gramaticais diferentes (3a e 3b) é mais fácil de resolver que se a ambigüidade ocorrer dentro de uma mesma classe gramatical (4) (Leffa 95).

(2a) O *canto* da sala.

(2b) O *canto* gregoriano.

(3a) Eu *calo* quando necessário. (calo=verbo calar)

(3b) Tirei o *calo*. (calo=substantivo)

(4) Ela gostou da *fazenda*. (fazenda=imóvel ou tecido)

A ambigüidade em nível sintático ocorre quando duas frases têm a mesma estrutura superficial, mas partem de uma estrutura profunda diferente. A estrutura superficial pode sofrer modificações quando a frase é transposta de uma língua para outra, como nos exemplos (5) e (6) abaixo.

(5a) The girl wants the boy to read.

(5b) A menina quer que o menino leia.

(6a) The girl wants the book to read.

(6b) A menina quer o livro para ler.

A ambigüidade em nível pragmático ocorre quando duas sentenças têm a mesma estrutura superficial e a mesma estrutura profunda, mas produzem um efeito de sentido diferente. O exemplo (7), abaixo, pode referir-se a uma ameaça, uma brincadeira ou mesmo uma dúvida, tanto em inglês quanto em português; o contexto é que vai determinar o sentido. O problema é que a determinação a partir do contexto nem sempre é possível: o exemplo (8a-c), que é provavelmente uma dúvida, pode ser traduzido como uma dúvida ou como uma solicitação.

(7a) Do you know who I am?

(7b) Você sabe quem eu sou?

(8a) Would you help me?

(8b) Você me ajudaria? (dúvida)

(8c) Você poderia me ajudar? (solicitação)

2.3 Referências anafóricas

Segundo Saggion (Saggion & Carvalho 95) existem certos mecanismos utilizados pelos produtores de texto para assinalar as relações entre partes de enunciados e é por meio desses mecanismos que a estrutura do texto é construída. A coesão é o fator responsável pela textualidade, e uma seqüência de sentenças se constituirá em um texto à medida que existirem relações de coesão entre estas sentenças. *Anáfora* é um caso especial de coesão. Consiste na utilização de uma referência abreviada a uma entidade, ou entidades, na expectativa de que o ouvinte do discurso seja capaz de entender a referência e determinar a identidade da entidade.

A referência abreviada é chamada de *anáfora*, e a entidade à qual ela se refere chama-se *referente*, ou *antecedente*. O processo pelo qual determina-se o referente da anáfora é chamado de *resolução*. No exemplo (9) abaixo, o pronome *ela* refere-se à entidade *mulher* definida anteriormente, enquanto o pronome *o* refere-se à entidade *chapéu* também já definida.

(9) A mulher comprou um chapéu. *Ela* pagou-*o* com cheque.

A identificação dos antecedentes é muitas vezes fundamental para a correção do processo de tradução, principalmente quando a tradução envolve línguas que marcam o gênero dos pronomes (Hutchins & Somers 92), como é o caso do português. Nos exemplos (10) e (11) a tradução só é possível após identificar corretamente o antecedente do pronome *it*.

(10a) *The monkey* ate the banana because *it* was hungry.

(10b) *O macaco* comeu a banana porque *ele* estava com fome.

(11a) The monkey ate *the banana* because *it* was ripe.

(11b) O macaco comeu a banana porque *ela* estava madura.

A referência anafórica pode ser vista como um tipo especial de ambigüidade, em que a escolha entre os possíveis antecedentes pode envolver o conhecimento lingüístico de restrições de co-ocorrência para indicar qual o antecedente mais adequado. No exemplo (10a), a escolha de *monkey* poderia ser resultante da restrição de *hungry* se referir, normalmente, a seres animados.

2.4 O Problema da Divergência Estrutural na Tradução Automática

A divergência é definida em termos de um fenômeno *língua-a-língua*: ocorre quando uma sentença em uma LF traduz-se na LA de forma diferente (as árvores possuem estruturas diferenciadas, ou possuem estruturas similares, porém seus nodos apresentam categorias básicas diferentes). Isto significa que a divergência pode ocorrer entre duas línguas quaisquer independente da forma como a tradução é realizada (direta, *transfer* ou interlíngua, a ser visto no capítulo 3). A existência de divergências de tradução, ou seja, distinções lingüísticas entre as línguas, torna impraticável o processo de tradução direta entre as estruturas fonte e alvo.

Segundo Bonnie Dorr (Dorr 94) as divergências podem ser divididas em duas categorias: divergências sintáticas, definida a partir de propriedades específicas de cada língua e independente dos itens léxicos utilizados, e divergências léxico-semânticas, que podem ser determinadas apenas a partir de propriedades que são determinadas pelos itens léxicos utilizados.

Arnold (Arnold et al. 94), por outro lado, propõe uma classificação das divergências em léxicas e sintáticas. As primeiras têm a ver com a forma como a língua classifica as palavras, conceitos que escolhe para expressar por palavras simples e aquelas que simplesmente não são lexicalizadas. O segundo tipo ocorre porque as línguas utilizam diferentes estruturas para o mesmo propósito e, em muitos casos, a mesma estrutura para diferentes propósitos.

Em nenhum dos casos a língua portuguesa foi utilizada como objeto de estudo, o que torna bastante complexa a utilização direta destas classificações. Desta forma, decidiu-se por definir uma classificação das divergências que ocorrem na tradução entre as línguas portuguesa e inglesa, baseada na adequação de trabalhos realizados para outros pares de línguas (Abeillé et al. 90, Arnold et al. 94, Dorr 93, Dorr 94, Eynde 93, Hutchins92, Kipper 95). A classificação proposta divide as divergências em divergências léxicas, divergências sintáticas e divergências léxico-semânticas.

2.4.1 Divergência Léxica

O problema da divergência léxica ocorre nos casos em que não existe uma simetria entre itens léxicos das línguas fonte e alvo. Três casos de divergências léxicas apresentam problemas no processo de tradução automática: divergência léxica conceitual, vazio léxico da LF e vazio léxico da LA.

A *divergência léxica conceitual* representa, provavelmente, o maior problema para a tradução automática. Ela ocorre quando um conceito simples, representado por uma única palavra, corresponde a um conjunto de significados, ou mesmo de palavras, na LA. A figura 2.1 apresenta alguns exemplos deste tipo de divergência, bem como uma possível proposta de resolução.

Português	Inglês	Observações
Discussão	1. *discussion* (troca de idéias para definir algo) 2. *talks* (discussão formalmente organizada entre governantes ou instituições)	Ambas as alternativas representam o mesmo conceito; o problema básico é de estilo do texto. No corpus trabalhado, por ser aplicado à área econômica, em manchetes de jornal, a tradução *talks* mostra-se mais adequada.
Banco	1. *seat* (local para sentar) 2. *bank* (instituição financeira)	As alternativas apresentam conceitos distintos; a tradução *bank*, no corpus utilizado, seria a única aceitável.
Querer	1. *wish* (desejo de fazer algo) 2. *call for* (pedido público)	Aparentemente, uma verificação de atributos semânticos do complemento (objeto direto) da sentença pode indicar a escolha correta em cada caso.

Figura 2.1 - Exemplos de divergências léxicas conceituais

O problema do *vazio léxico*, tanto da língua-fonte quanto alvo, ocorre quando um determinado conceito, expresso por uma palavra simples ou expressão, em uma língua não possui um item correspondente na outra. Neste caso, o vazio léxico deve ser substituído pela descrição do conceito ou mesmo por uma palavra de sentido similar. Os exemplos (12) a (14) apresentam alguns destes casos.

(12) uniformização → regularization (palavra de sentido similar)

(13) empreiteiras → contract construction companies (descrição do conceito)

(14) parlamentares → members of the parliament (descrição do conceito)

Pode-se notar, pelos exemplos (13) e (14), que a herança de traços semânticos não ocorre de forma uniforme. Nestes exemplos o traço semântico *número* não permite uma regra geral de herança, exigindo um tratamento adequado a cada caso.

2.4.2 Divergência Sintática

Na sistematização das divergências sintáticas descrita por Dorr (Dorr 93, Dorr 94) é utilizada a teoria de *Government and Binding* (GB-*theory*) (Raposo 92). Neste modelo, informações independentes e específicas de uma língua são expressas, respectivamente, pelos *Princípios* e *Parâmetros* (PP).

A abordagem PP foi proposta como um modelo de como uma criança poderia adquirir habilidades que permitissem a ela usar a língua eficientemente. Neste modelo os *princípios* formam um núcleo com definições universais, enquanto os *parâmetros* contêm informações que devem ser aprendidas e são específicas a uma língua particular (Brill 93).

Utilizando estes princípios, foram definidas três classes de divergências sintáticas que podem ocorrer: ordem dos constituintes, deslocamento de preposições e omissão do sujeito. Duas outras classes, *dativo* e *movimentos de longa distância*, foram omitidas, por não ocorrerem na tradução entre as línguas portuguesa e inglesa, enfocadas neste trabalho.

2.4.2.1 Ordem dos Constituintes

A parametrização *ordem dos constituintes* determina a posição onde os especificadores (sujeito de uma sentença, por exemplo) e seus complementos (como objetos de um verbo ou advérbios) de uma frase podem ser posicionados. Os exemplos (15) e (16) apresentam divergências no posicionamento de *adjetivos* em relação a um sintagma nominal. Em Inglês, o adjetivo é normalmente posicionado à esquerda do nome, enquanto em

Português ele pode ocorrer em qualquer posição, preferencialmente à direita. O exemplo (17) apresenta um exemplo de divergência em relação ao adjunto adnominal, que em português ocorre à direita do nome, enquanto na língua inglesa ocorre de forma pré-fixada.

(15a) He is a *big* man.

(15b) Ele é um homem *alto.*

(15c) *[3]Ele é um *alto* homem.

(16a) She is a *beautiful* woman.

(16b) Ela é uma mulher *bonita.*

(16c) Ela é uma *bonita* mulher.

(17a) Presidente *da Irlanda*.

(17b) *Irish* president.

2.4.2.2 Associação de preposições

Este tipo de divergência ocorre devido ao fato de na língua inglesa a preposição governar o elemento de *trace*[4] (18b,19b), por isto não pode ser deslocada. Em Português, por sua vez, isto não ocorre, logo a preposição acompanha o pronome (18a,19a).

(18a) [*A qual* loja]*i* foi João ***Ti***[5] ?

(18b) [*What* store]*i* did John go *to* ***Ti*** ?

(19a) [*De Onde*]*i* você é ***Ti*** ?

(19b) [*Where*]*i* are you *from* ***Ti*** ?

[3] O símbolo "*", por convenção, indica uma sentença mal-formada, inválida.

[4]Segundo Chomsky, ao realizar-se um deslocamento na estrutura de uma sentença, é deixado um vestígio (elemento de *trace*) sem conteúdo fonético, na posição do constituinte movido.

[5]Na notação utilizada, ***Ti*** determina a posição original do elemento na sentença e [*x*]*i* indica o elemento que foi deslocado para gerar a sentença derivada.

2.4.2.3 Omissão do sujeito

A omissão do sujeito (sujeito nulo ou inexistente), comum na língua portuguesa, não é válida em Inglês. Por conseqüência, o sujeito deve ser derivado durante o processo de análise da LF. Os exemplos (20) e (21) apresentam dois casos de omissão do sujeito válidos na língua portuguesa e que devem ser derivados durante a tradução.

(20a) ∅ Está chovendo.

(20b) It is raining.

(21a) ∅ Foi ao cinema.

(21b) He went to the movie.

2.4.3 Divergência Léxico-Semântica

As divergências léxico-semânticas diferem das demais classes de divergências, léxicas e sintáticas, por ocorrerem a partir de propriedades específicas inseridas pela instanciação léxica realizada, ou seja, as regras que definem a divergência devem estar inseridas no dicionário bilíngüe. As divergências léxico-semânticas foram classificadas em: *conflacional*[6], estrutural, categorial, promocional e léxica.

2.4.3.1 Conflacional

Na divergência do tipo *conflacional*, ocorre a incorporação léxica de componentes de significado, ou argumentos, de uma determinada ação. Nos exemplos (22) e (23) o verbo principal em português é transitivo direto e indireto. Durante a tradução, o verbo principal e seu objeto direto são incorporados em um único item léxico (verbo transitivo direto).

(22a) Maria *passou manteiga* nas torradas.

(22b) Mary *buttered* the toasts.

(23a) Fernando Henrique *faz crítica* ao senador Dutra.

---

[6] Do termo inglês *conflation*: junção de dois conceitos com a finalidade de formar um único.

(23b) Fernando Henrique *criticizes* senator Dutra.

2.4.3.2 Estrutural

A divergência estrutural é caracterizada por uma mudança na relação existente entre os argumentos dentro da estrutura. No exemplo (24) o sintagma preposicional *na casa*, que tem função semântica de objeto indireto da sentença, é traduzido como objeto verbal (objeto direto) *the house* em inglês.

(24a) Maria entrou *na casa*.

(24b) Mary entered *the house*.

2.4.3.3 Categorial

A divergência *categorial* é caracterizada pela alteração da categoria gramatical de um determinado item léxico. No exemplo (25), o predicativo do sujeito é formado pelo nome *fome* em português, e é traduzido como o adjetivo *hungry* em Inglês.

(25a) Eu tenho *fome*.

(25b) I am *hungry*.

2.4.3.4 Promocional

De forma similar, na divergência promocional ocorre a ascensão de um determinado elemento na estrutura sintática. No exemplo (26) o verbo principal *costuma* em Português é traduzido como o sintagma adverbial *usually* em Inglês, enquanto o verbo da oração subordinada (predicativa) passou a ser o verbo principal da oração em inglês.

(26a) Jõao *costuma ir* para casa.

(26b) John *usually goes* home.

2.4.3.5 Léxica

Finalmente, a divergência léxica decorre do uso de expressões. No exemplo (27) o verbo principal em Português é *forçar*, *to force* em Inglês, mas a expressão *forçar a entrada* é traduzida de forma mais adequada como o verbo *break into*.

(27a) João *forçou a entrada* no quarto.

(27b) John *broke into* the room.

3. SISTEMAS DE TRADUÇÃO AUTOMÁTICA

3.1 Introdução

Os sistemas de tradução automática podem ser divididos em dois grandes grupos: (i) sistemas diretos, que incorporam em um único programa os conhecimentos necessários para o processo de tradução completa; (ii) sistemas indiretos - *interlíngua* e *transfer*- que dividem o processamento em estágios independentes: análise e geração.

A principal diferença entre os métodos está na complexidade de tratamento dos três componentes da tradução: análise, transferência ou conversão, e geração (síntese). A figura 3.1 apresenta esta distinção: o método direto está em um extremo, enquanto o método *interlíngua* encontra-se no outro. À medida que a análise do texto-fonte torna-se mais profunda, ela se torna mais semântica, porém a geração inversa torna-se mais complexa. No topo da figura 3.1 temos o método *interlíngua*, ou seja, de representação altamente semântica, ou independente da língua utilizada, enquanto na base estão os métodos diretos, que realizam muito pouca (ou mesmo nenhuma) análise e o processo de geração é bastante simples.

A seguir são descritos mais detalhadamente cada um destes métodos e exemplificada sua utilização.

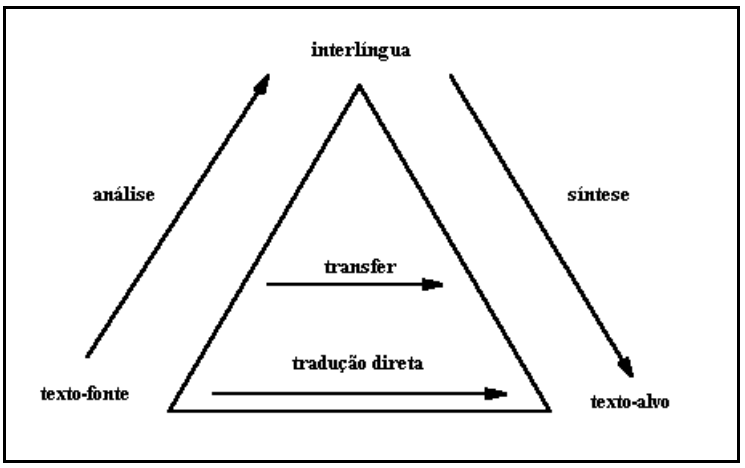


Figura 3.1 - Diagrama comparativo dos métodos de tradução

3.1.1 Métodos Diretos

Métodos diretos são característicos de sistemas projetados desde o início para traduzir de um idioma específico para outro. Sistemas desse tipo são limitados ao mínimo de trabalho necessário para realizar a tradução. Nenhuma teoria lingüística geral ou princípios de análise gramatical estão necessariamente presentes para a tradução ser realizada. Ao invés disto, estes sistemas dependem de dicionários bem-desenvolvidos, da análise morfológica e de funções de processamento de textos para chegar a traduções confiáveis do texto-fonte.

A figura 3.2 apresenta um esquema geral de tradução realizada pelos métodos diretos.

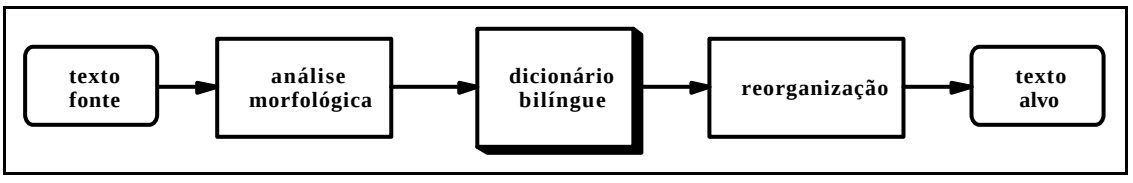


Figura 3.2 - Esquema geral dos métodos diretos

Para ilustrar o funcionamento deste tipo de método, é apresentado, a seguir, um exemplo, adaptado de (Hutchins & Somers 92), de tradução inglês-português que poderia ser gerada a partir do exemplo (28), utilizando o esquema da figura 3.2.

(28) *Any government is dependent on its supporters.*

O primeiro passo realizado é a análise léxico-morfológica do texto-fonte. O resultado deste processo é uma estrutura linear na qual são identificados os traços de cada uma das palavras, como:

- categoria gramatical (cat)={determinante, substantivo, verbo, adjetivo, preposição};
- número (num)={singular, plural}, se substantivos;
- tempo={presente, passado}, se verbos
- gênero (sex)={masc, fem, neutro}
- pessoa (pess), se pronomes.

O resultado para o exemplo (28) poderia ser a estrutura apresentada em (29).

(29)

any	Government	be	dependent	on	its	supporter
cat=det	Cat=subst Num=sg	cat=v num=sg tempo=pres	cat=Adj	cat=prep	cat=pron Num=sg pess=3 sex=neut	Cat=subs Num=pl

Este passo de análise identifica várias informações sobre as palavras isoladamente, porém não é capaz de identificar relações entre as palavras como, por exemplo, a relação entre o determinante e o substantivo, ou mesmo um sintagma nominal.

O passo seguinte realiza a pesquisa, no dicionário bilíngüe, das palavras do texto-fonte, com seus atributos, e faz a substituição por palavras da LA (preservando os atributos). Os exemplos (30a) e (30b) apresentam possíveis traduções para a sentença (28). Nestes exemplos estão grifados dois dos problemas que freqüentemente são encontrados neste tipo de abordagem: o verbo inglês *to be* pode ser traduzido, no português, tanto para *ser* como para *estar*; e os pronomes, em português, são marcados em gênero e número pelo substantivo a que se referem. Neste método, tais problemas dificilmente são resolvidos sem uma pós-revisão do texto-alvo.

(30a) qualquer governo /*é*/ dependente de /*seu*/ aliados.

(30b) qualquer governo /*está*/ dependente de /*seu*/ aliados.

Finalmente, em um último passo do processo de tradução, a maioria dos métodos diretos inclui algum tipo de identificação de contexto local, de forma a reduzir um pouco o problema da divergência entre as línguas envolvidas na tradução. Esta reorganização realiza uma verificação de palavras vizinhas, com o objetivo de detectar possíveis divergências e corrigi-las. Por exemplo: construções <adjetivo+substantivo> usualmente devem ser reescritas, em português, como <substantivo+adjetivo>.

(31) He has an *old car*.

(31a) *Ele possui um *velho carro*.

(31b) Ele possui um *carro velho*.

Outro tipo de problema dos sistemas diretos é o tratamento de expressões compostas como, por exemplo, *look after* (cuidar de), que devem ser tratadas como uma unidade. Uma alternativa utilizada é a simples inclusão das expressões no dicionário, fazendo com que o analisador léxico verifique as palavras vizinhas, com a finalidade de detectar estes casos. Isto nem sempre é válido, pois existem inúmeros casos em que a expressão pode ser utilizada de forma não contínua, como no exemplo (32), com o verbo *look up* (pesquisar):

(32) He *looked* the word *up* in the dictionary.

(32a) Ele *procurou* o termo no dicionário.

O tratamento de homógrafas constitui outro problema. Em alguns casos, este problema pode ser resolvido através de uma pesquisa envolvendo as palavras adjacentes para determinar sua classificação correta.

A palavra *empty*, por exemplo, pode ser classificada como adjetivo ou como um verbo. A seguinte regra pode ser utilizada: se *empty* é utilizado entre um determinante e um substantivo, provavelmente é um adjetivo (33a), caso contrário é um verbo (33b). Esta regra, porém, não é suficiente, uma vez que *empty* pode, ainda, aparecer entre outro adjetivo e um substantivo ou mesmo no final da frase (33c).

(33a) He is on an *empty* stomach.

(33b) Could you *empty* the trash?

(33c) The trash is *empty*.

Listar no dicionário todas as regras possíveis para cada homógrafa é um processo trabalhoso e irá exigir dicionários gigantescos, mas é a única solução possível para os métodos diretos.

A partir destas várias dificuldades, Tucker sintetiza os estágios mínimos envolvidos para permitir uma confiabilidade mínima dos métodos diretos. Estes passos, apresentados na figura 3.3, são um detalhamento da figura 3.2 e tentam minimizar os problemas apresentados acima.

(a) busca de palavras do texto-fonte no dicionário e análise morfológica;
(b) identificação de homógrafas;
(c) identificação de nomes compostos;
(d) identificação dos predicados nominais e verbais;
(e) processamento de expressões idiomáticas;
(f) processamento de preposições;
(g) identificação de sujeito-predicado;
(h) identificação de ambigüidade sintática;
(i) síntese e processamento morfológico do texto-alvo;
(j) reorganização de palavras e frases no texto-alvo.

Figura 3.3 - Passos mínimos para tradução direta (Tucker 84)

Concluindo, pode-se dizer que este tipo de abordagem apresenta resultados normalmente pobres, que exigem uma pós-revisão para a geração de um documento aceitável. Sua aplicabilidade é restrita, desta forma, a pares de línguas que apresentem poucos problemas léxicos (ambigüidade lexical) e estruturais (divergências), o que só é possível em domínios restritos e entre línguas cognatas[7].

3.1.2 Métodos *Interlíngua*

A abordagem *interlíngua* é caracterizada por sistemas nos quais a representação do *significado* da LF é independente de qualquer língua natural, sendo a única representação utilizada para geração do texto-alvo. Desta forma, a representação de uma unidade de significado qualquer será a mesma, independente da língua (ou estrutura gramatical) na qual é expressa. Subjacente a este método está a noção de *Universais Lingüísticos* buscados por lingüistas e filósofos (Araújo 93).

[7] Línguas que possuem a mesma raiz. Por exemplo, as línguas românicas que são derivadas do latim (o português, o espanhol, o francês, etc.)

A figura 3.4 apresenta um esquema geral dos sistemas baseados na abordagem *interlíngua*.

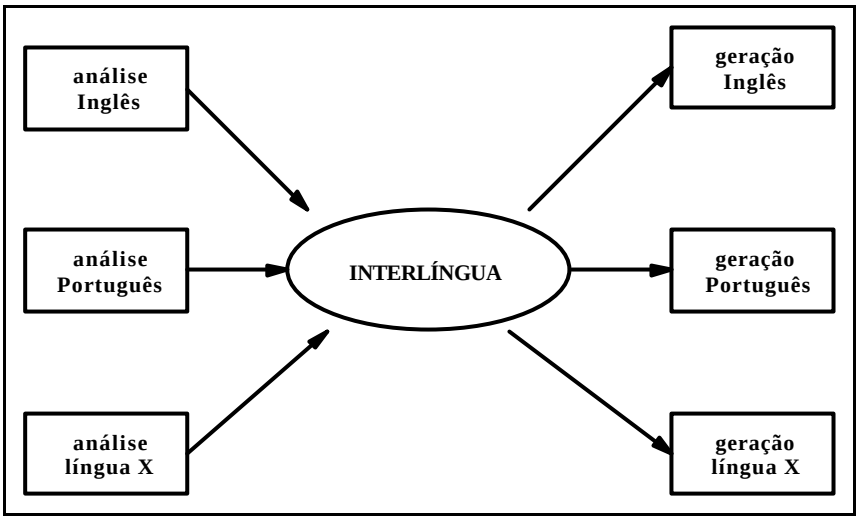


Figura 3.4 - Ambiente de tradução interlíngua (Hutchins & Somers 92)

A grande vantagem de sistemas baseados no método *interlíngua* seria a possibilidade de realizar a tradução em qualquer sentido, bastando para isto, apenas, a construção de um módulo de análise, responsável pela tradução dos textos-fonte nesta língua para a representação intermediária, e um módulo de geração da representação intermediária para o texto-alvo, conforme figura 3.5. Observa-se, porém, junto a muitos autores, que a língua não se constitui objeto de engenharia reversa, ou seja, esta bi-direcionalidade atribuída aos métodos interlíngua, normalmente, não é possível.

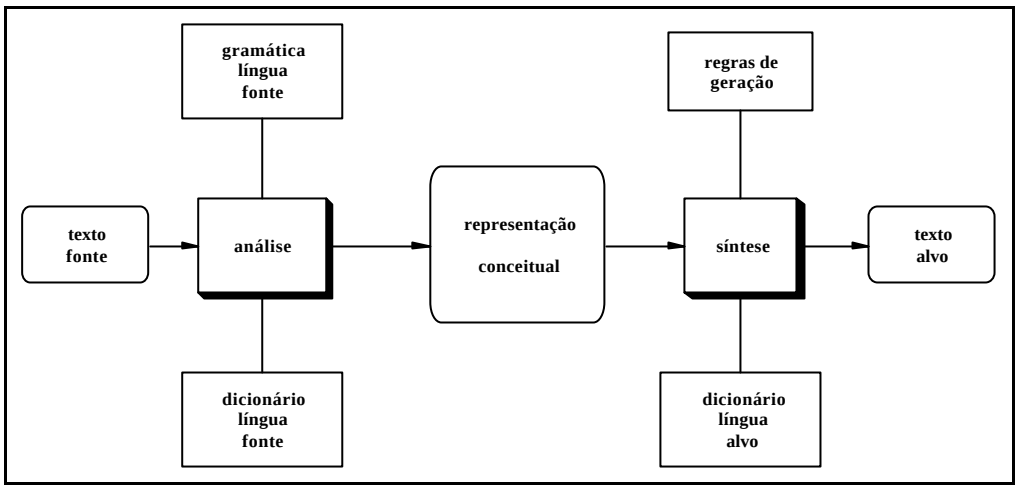


Figura 3.5 - Arquitetura geral para os métodos *interlíngua*

No método *interlíngua*, o resultado da análise do texto-fonte é uma representação independente da sentença de entrada. Esta representação é a base para a geração do texto-alvo. Uma desvantagem é a separação total entre análise e geração; não é desejável orientar a análise para uma LA específica e, ao mesmo tempo, não é possível, durante a geração, voltar para realizar alguma verificação sobre o texto-fonte.

O componente *interlíngua* deve incluir toda informação que pode ser necessária durante a fase de geração, mais precisamente, para qualquer língua envolvida no sistema, ou que possa vir a ser aí incluída.

Alguns sistemas têm fórmulas da lógica de primeira ordem como representação interlíngua. Para o exemplo (28), poderíamos ter algo como:

(34) all(X), government(X), indefinite(Y), plural(Y), support(Y,X,T), depend-on(X,Y,T), timeless(T)

Segundo Hutchins, gerar este tipo de representação tem-se mostrado mais simples que fazer a geração do texto-alvo a partir destas regras. Os exemplos abaixo (35a-c) ilustram diferentes alternativas de geração para o interlíngua (34). Este possível parafraseamento na geração normalmente levará à distorção do significado original apresentado no texto-fonte (Hutchins & Somers 92).

(35a) Every government has supporters which it depends on.

(35b) People who support all governments are depended on by them.

(35c) All governments depend on people who support them.

Inicialmente a expectativa era o desenvolvimento de uma representação universal, que poderia ser intermediária a qualquer língua natural; porém, isto depende da possibilidade de capturar e formalizar o conhecimento humano, necessário para compreensão da língua. A partir deste problema, as pesquisas se desenvolveram em torno dos métodos *transfer*, onde o conhecimento envolvido é restrito a um par específico de línguas.

3.1.3 Métodos *Transfer*

Nos métodos *transfer*, a representação do *significa*do de uma unidade gramatical difere, dependendo da língua a partir da qual esta unidade foi derivada ou para a qual será gerada. Isto implica a existência de um estágio intermediário de tradução, chamado *transfer*, o qual mapeia a representação específica do significado de uma língua para outra.

Uma sentença na LF é analisada gramaticalmente, obtendo-se uma representação interna abstrata (geralmente uma estrutura marcada com traços sintáticos e semânticos). Após, uma *transferência* é realizada, em nível léxico e estrutural, para as estruturas correspondentes na LA. Em um terceiro momento, a tradução é gerada. Três dicionários são necessários para a transferência: um dicionário da LF, um dicionário bilíngüe e um dicionário da LA. A figura 3.6 apresenta um esquema geral dos métodos *transfer*.

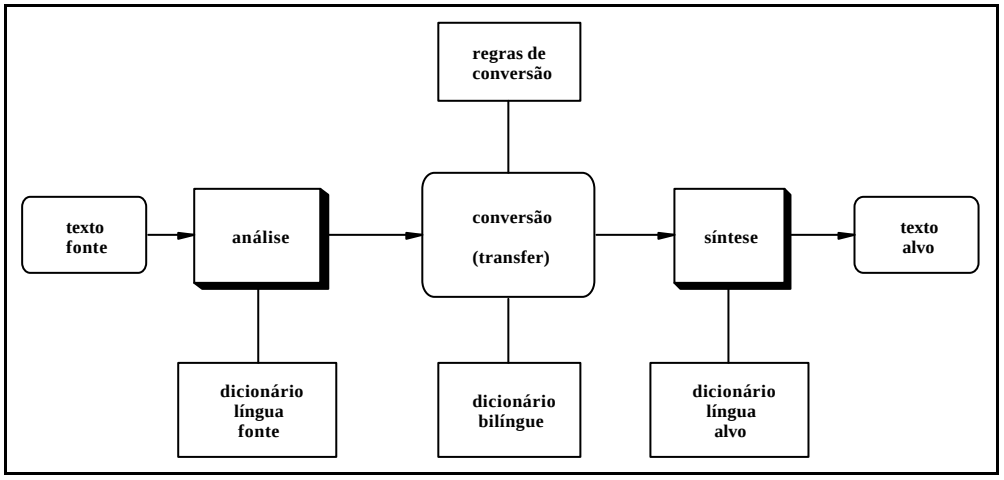


Figura 3.6 - Arquitetura geral para os métodos *transfer*

A bibliografia normalmente divide os métodos *transfer* em dois grupos básicos: escopo local e escopo global (Beardon Lumsden & Holmes 91, Hutchins & Somers 92, Whitelock & Kilby 95).

O escopo local caracteriza sistemas *transfer* nos quais as palavras são a unidade central que direciona a análise. A análise é realizada por procedimentos isolados para cada palavra, que devem determinar a sua classe gramatical, o uso possível e o sentido da palavra. As regras de transferência estão codificadas junto às palavras, no dicionário bilíngüe, possuindo todas as informações necessárias para a conversão, como por exemplo o conjunto de regras gramaticais de gramáticas categoriais e gramáticas probabilísticas.

O escopo global caracteriza sistemas *transfer* nos quais o significado de uma palavra é determinado pelo seu contexto dentro de uma análise global da sentença (ou, mais raramente, do parágrafo). A transferência é definida através de regras explícitas de tradução, sobre as gramáticas das línguas fonte e alvo.

Para o exemplo (28) e a partir do conjunto de informações léxico-morfológicas apresentadas em (29), o método *transfer* realiza a análise sintática do texto, resultando em uma representação hierárquica (ver figura 3.7) que inclui um conjunto de traços morfo-sintáticos e semânticos que serão utilizados durante o processo de transferência para a LA, conforme a figura 3.8 (a representação estrutural desta sentença é baseada no conjunto de estruturas TAG definidas em (Kipper 94)).

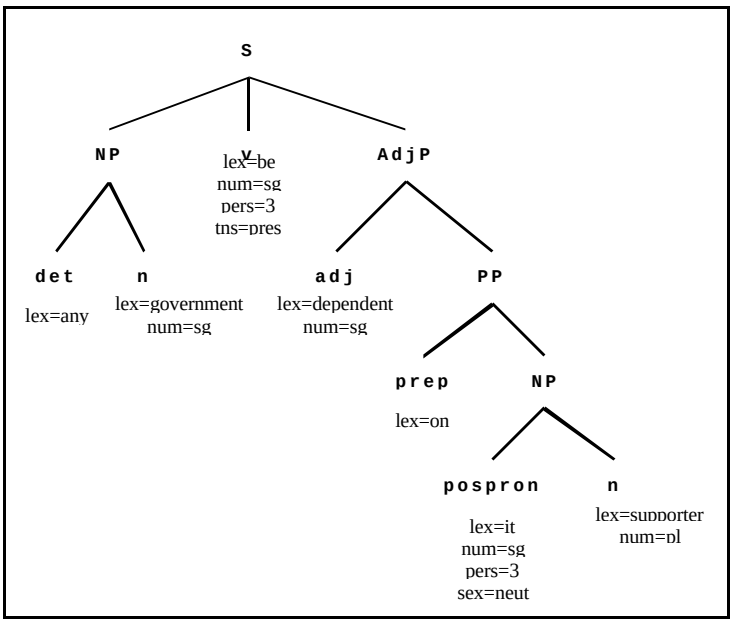


Figura 3.7 - Estrutura sintática fonte para *Any government is dependent on its supporter*

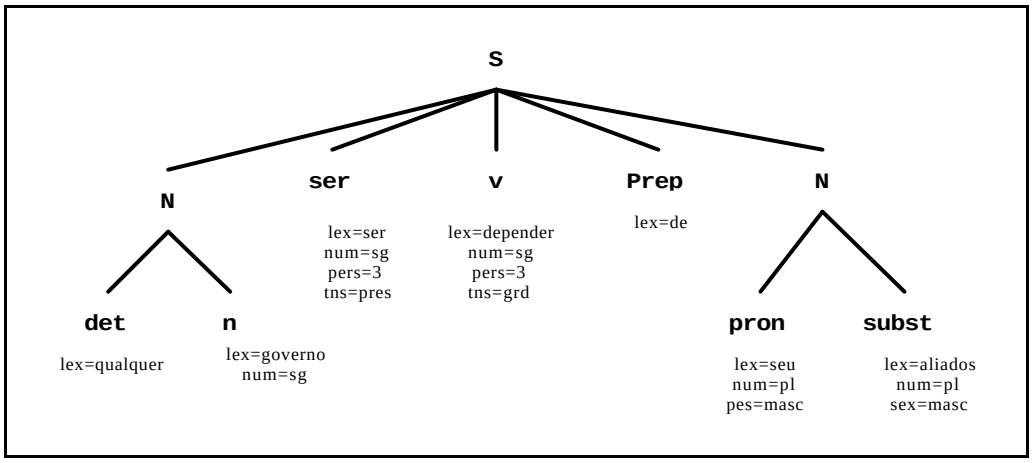


Figura 3.8 - Estrutura sintática alvo para *Qualquer governo é dependente de seus aliados*

A transferência é baseada em uma análise profunda da sentença, de forma a reconhecer seus constituintes e realizar o mapeamento. Por exemplo, na estrutura anteriormente referida o verbo *to be* (ser) é um verbo auxiliar e o adjetivo *dependent* deve ser mapeado na estrutura transposta para o português como o verbo principal. Informações de gênero e número também devem ser identificadas e herdadas pelos nodos correspondentes.

3.1.4 A técnica *shake-and-bake*

A técnica *shake-and-bake* foi proposta a partir de algumas deficiências básicas encontradas nas arquiteturas tradicionais de tradução, baseadas em métodos *transfer* ou *interlíngua* (Whitelock 92, Beaven 92).

Nos métodos *transfer*, o componente de transferência é dependente de um par de línguas específico, e deve ser totalmente escrito levando em consideração os componentes monolíngües de cada língua para permitir a compatibilidade. O número de regras pode ser bastante grande e complexo de escrever, pois devem ser consideradas as gramáticas fonte e alvo. Ainda, a portabilidade é bastante baixa, pois alterações na gramática de uma língua podem exigir grandes modificações no componente de transferência.

Nos métodos *interlíngua*, por outro lado, se o sistema é robusto, deve ser possível que qualquer expressão na LF seja tratável na geração da LA. Se o interlíngua é poderoso o suficiente para representar o significado de todas as línguas envolvidas, haverá muitas fórmulas no interlíngua que serão equivalentes às produzidas pelo processo de análise, e não será possível garantir que a fórmula correta será a utilizada no processo de geração, a menos que seja possível a realização de inferências lógicas no interlíngua. A complexidade computacional deste processo pode inviabilizar seu uso, uma vez que o problema é NP-completo (Garey & Johnson 79).

A técnica *shake-and-bake* procura superar estes problemas, oferecendo grande modularidade dos componentes monolíngües, que podem ser escritos de forma independente, utilizando apenas as construções monolíngües que serão colocadas em correspondência através de um dicionário bilíngüe.

A técnica é baseada em uma visão léxica da gramática, e a tradução é definida por meio de uma equivalência entre conjuntos de itens léxicos.

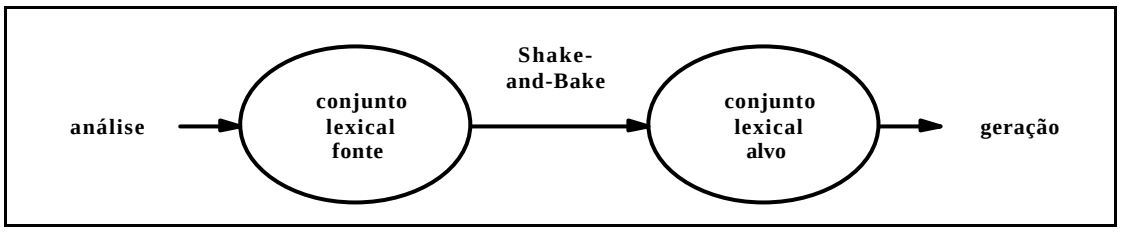


Figura 3.9 - A técnica *Shake-and-Bake*

A figura 3.9 apresenta a idéia básica de um sistema *shake-and-bake*: o processo de análise gera um conjunto de itens léxicos que mantêm informações sobre a ordem das palavras no texto-fonte. O componente de transferência combina estes itens livremente de acordo com a gramática da LA, fazendo a substituição das palavras e os reordenamentos necessários. Finalmente, a geração é responsável por dar o estilo final ao texto-alvo.

O compartilhamento de variáveis entre as duas línguas garante a manutenção das restrições semânticas, enquanto a facilidade de formular regras de equivalência entre estruturas sintáticas de línguas divergentes possibilita que esta técnica seja utilizada para tradução em ambos os sentidos.

3.1.5 Técnicas estocásticas e tradução baseada em exemplos

A idéia básica das técnicas estocásticas é que a tradução pode ser realizada a partir do uso de exemplos análogos, baseada na busca de expressões ou estruturas que já foram utilizadas em traduções anteriormente realizadas.

A base de dados de exemplos é derivada a partir de um corpus, necessariamente extenso, de textos na LF, e suas traduções em uma LA produzidas por tradutores profissionais. Este processo é chamado de alinhamento, e resulta em um conjunto de exemplos de ocorrências e suas traduções mais freqüentes. A figura 3.10 apresenta a arquitetura de um sistema de alinhamento, que realiza a leitura do corpus tentando identificar expressões e modelos de tradução.

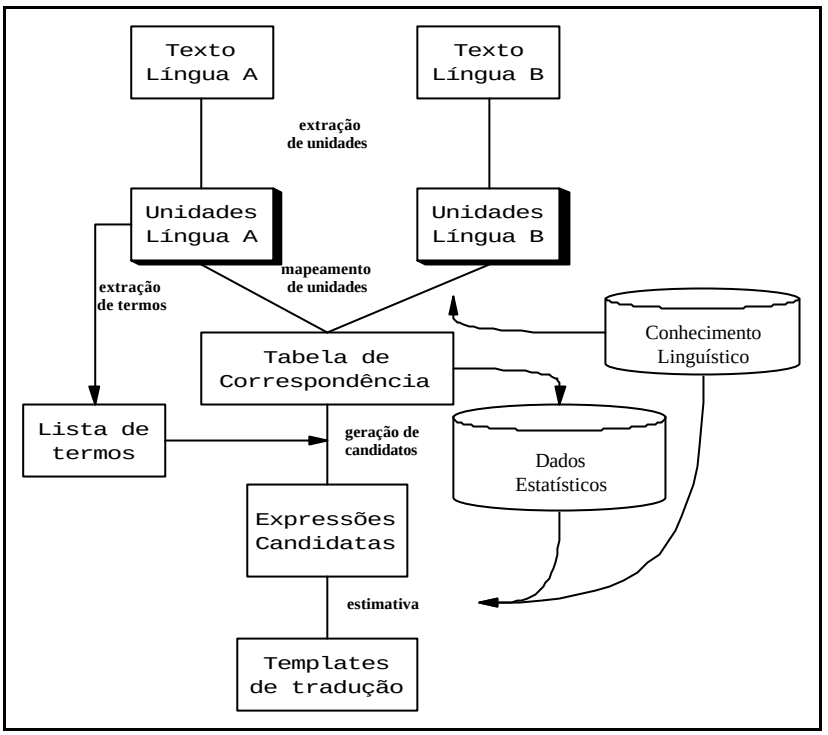


Figura 3.10 - Técnicas estocásticas - alinhamento de sentenças (Kumano 94)

Esta figura apresenta um conjunto de processos a destacar:

- Extração de unidades: unidades do texto (partes) são extraídas de ambos os corpora. As unidades podem ser palavras, expressões ou sentenças completas. O restante do texto (frases não selecionadas) é utilizado como informação de contexto.
- Mapeamento de unidades: cada unidade na LF A é mapeada em unidades da LA B, gerando uma tabela de relações entre as unidades e as suas respectivas informações contextuais. Para isto é utilizado um conceito de *grau de semelhança*, buscado de uma base de conhecimentos lingüísticos e um dicionário bilíngüe.
- Extração de termos: um conjunto de termos candidatos é extraído do texto-fonte. Este passo tenta diminuir os possíveis erros inseridos durante o mapeamento das unidades.
- Geração de candidatas: os termos selecionados para a extração são retirados do texto-alvo, e para isto é utilizado o método de *n*-gramas, onde *n* varia de acordo com o tamanho da expressão candidata e significa o

número de palavras que serão pesquisadas à direita ou à esquerda desta palavra para determinar sua correta utilização (Villavicencio 95).

- Estimativa: as expressões candidatas são avaliadas estatisticamente para obtenção das melhores alternativas.

Uma especialização deste tipo de técnica é a utilização para o aprendizado de estruturas (*templates*), de forma que a tradução possa ser realizada para pares de sentenças com estruturas similares (Filgueiras 94). Estas estruturas são obtidas a partir de estatísticas sobre a distribuição de construções-chave como, por exemplo, a similaridade de seqüências de categorias gramaticais. No exemplo (36a) é apresentado um possível *template* capaz de gerar a sentença (36b).

(36a) Remove X and replace it with Y.

⇒ Remover X e substituir por Y.

(36b) Remove *the bulb* and replace it with *a new one*.

⇒ Remover *a lâmpada* e substituir por *uma nova.*

A partir do conjunto de expressões alinhadas e dos *templates* de tradução, pode-se redefinir a arquitetura do sistema de tradução para incorporar esta técnica, de acordo com a figura 3.11.

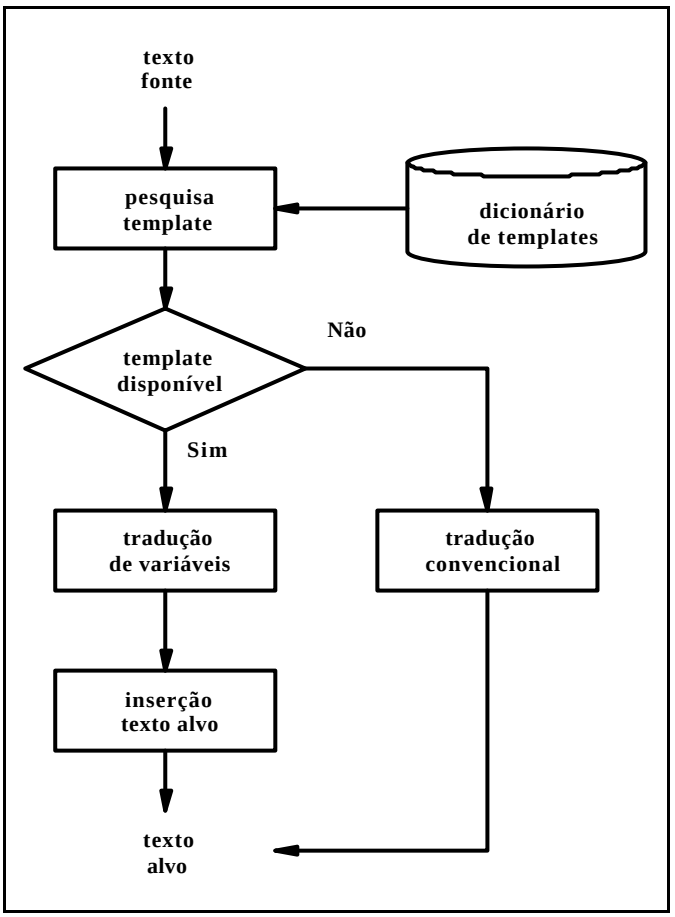


Figura 3.11 - Arquitetura de um sistema estatístico (Kinoshita et al. 94)

3.2 Alguns Sistemas baseados no Método *Transfer*

A seguir são apresentados alguns projetos que utilizam o método *transfer* para realizar a tradução, procurando enfatizar o modelo de dados utilizado e a interação com os algoritmos de análise sintática e transferência. A dificuldade na obtenção de bibliografia adequada a um estudo específico sobre o tratamento de divergências estruturais dificulta a comparação a este nível.

3.2.1 Sistema TAUM-Aviation

O sistema TAUM-Aviation foi projetado para realizar a tradução de manuais de manutenção de aeronaves entre as línguas inglesa e francesa. O sistema se processa em três estágios distintos: análise monolíngüe do texto-fonte, fase de transferência bilíngüe e geração

do texto-alvo (Whitelock & Kilby 95). Um conceito de *contexto gramatical* é utilizado para desambiguação entre diferentes acepções de uma palavra enquanto as divergências estruturais são tratadas a partir de reordenamentos da representação interna do texto.

A fase de análise é baseada no formalismo Redes de Transição Estendidas (ATNs) (Woods 73) e produz uma estrutura de dados no formato de um grafo dirigido, com vértices de entrada e saída únicos e cujos arcos são rotulados com itens léxicos, ou árvores, e "etiquetados" com traços semânticos. Esta estrutura é utilizada pela fase de transferência.

Um único dicionário de análise é utilizado, com a finalidade de obter integridade da organização e facilidade de atualização. Três tipos de regras de reescrita são manipuladas:

- regras de equivalência - define a troca de uma literal (conjunto de uma ou mais etiquetas) por uma nova literal (de etiquetas).
- regra idiomática potencial - define um arco alternativo, gerado pela substituição de um conjunto de arcos rotulados por itens léxicos por um único arco, rotulado com uma árvore sintática;
- regra idiomática fixa - mesma operação que a anterior, porém descarta a seqüência de rótulos iniciais utilizados.

A partir da análise estrutural, é gerada uma estrutura normalizada, no formato de grafo dirigido, contendo os caminhos alternativos representando cada uma das possíveis estruturas solução para cada sentença de entrada.

A transferência léxica é realizada diretamente sobre as folhas da estrutura normalizada. Este processo pode percorrer a estrutura buscando pela vizinhança do nodo, e, eventualmente, alterar a estrutura sintática. Neste momento são realizadas transformações relativas às divergências léxicas estruturais, mais especificamente ao problema do vazio léxico (ver seções 2.4.1 e 5.1(b)).

O mesmo formalismo, chamado Q-Systems, utilizado para descrever o dicionário de análise permite descrever as regras de transferência estrutural. Cada uma das estruturas sintáticas possíveis é percorrida aplicando-se as regras de transferência:

- equivalência: gera uma estrutura equivalente, com as transformações adequadas, na árvore gerada; nestas regras são modeladas as divergências que são puramente estruturais (ver seções 2.4.2 e 5.1(a)). As regras são ativadas a partir do *patter-matching* entre as estruturas do dicionário e as estruturas efetivas na árvore;
- expressões fixas: detecta a existência de subestruturas que possuem tradução divergente. Normalmente estas regras são lexicalizadas e implementam as transformações relativas às divergências léxico-semânticas (ver seções 2.4.3 e 5.1 (c, d));
- expressões potenciais: manipulam divergências potenciais, mas que o processo de transferência não possui conhecimento necessário para resolução. Estruturas alternativas são geradas para que o processo de geração escolha a mais adequada.

3.2.2 Projeto GETA-Ariane

O projeto Geta-Ariane, desenvolvido na Universidade de Grenoble - França, é derivado de um dos mais antigos projetos de tradução automática, o projeto CETA, iniciado em 1961. O objetivo do projeto é o desenvolvimento de ferramentas computacionais para descrição e aplicação de conhecimentos lingüísticos, com a finalidade específica de viabilizar a TA. A forte ligação com a área da lingüística computacional é a principal característica do trabalho desenvolvido (Boitet et al. 82, Hutchins & Somers 92, Whitelock 92).

A estrutura básica manipulada é uma árvore rotulada complexa. Cada label é composto por uma estrutura baseada no conceito de *frames*, podendo conter informações escalares, como categoria ou tempo verbal, e restrições semânticas, como subcategorização de verbos ou papel semântico desempenhado pelo nodo (sujeito, objeto, complemento).

A transferência léxica é descrita em um modelo próprio, na metalinguagem TRANSF, e possui os seguintes tipos de regras:

- formatos condicionais e de atribuição: aplicáveis respectivamente sobre os itens léxicos da LF e LA, permitem manipular os atributos semânticos determinados pelo processo de análise com a finalidade de detectar problemas de polissemia dos termos;

- funções condicionais e de atribuição: regras de transformação em nível de itens léxicos que permitem realizar a transferência bilíngüe.

A transferência estrutural é realizada a partir de um dicionário de transferência, descrito a partir do formalismo ROBRA (Boitet et al. 82), e permite descrever os seguintes tipos de regras: reordenação dos termos na LA, distribuição e eliminação de constituintes, geração de artigos, cálculo de atributos de tempo e modo corretos na LA (contrariamente a uma simples herança dos traços), integração com as regras do dicionário léxico para ativar rotinas de resolução de expressões da LF.

3.2.3 Projeto METAL

O projeto METAL é o componente de tradução de um conjunto de ferramentas para processamento da língua natural desenvolvidas pela Universidade do Texas - EUA. O componente de transferência é baseado no formalismo *Head Phrase Structure Grammars* (HPSG) (Pollard & Sag 87), com forte ênfase aos aspectos lingüísticos envolvidos no processo de tradução (Slocum 85).

As HPSGs descrevem regras de reescrita livres de contexto. Cada regra transformacional permite descrever: testes sobre os componentes individuais; testes interconstituintes; construtores de frases; e uma regra de transferência.

As regras de transferência contém regras para geração da estrutura-alvo, podendo:

- modificar a estrutura-alvo, adicionando, apagando ou reordenando nodos;
- copiar, adicionar ou apagar traços semânticos de nodos descendentes dos nodos especificados na regra;
- importar condições para os descendentes dos nodos especificados nas regras.

O algoritmo de análise ativa o processo de transferência, ou seja, de forma síncrona, sempre que as condições constantes na regra de reescrita forem satisfeitas. Inicialmente o algoritmo tenta aplicar transformações definidas pela regra transformacional associada, caso

isto não seja possível, aplica, sucessivamente, um conjunto de regras pré-definidas de transferência, a partir de um conjunto de transformações gerais.

3.3 Análise Inicial sobre os Métodos *Transfer*

O método *transfer* tem sido o modelo padrão para sistemas atuais de TA. Embora vários dos sistemas comerciais disponíveis utilizem o método *direto*, e alguns o *interlíngua* (como por exemplo o sistema KBMT (Nirenburg et al. 92)), a grande maioria dos projetos de pesquisa e sistemas comerciais são sistemas baseados em um componente explícito de transferência (Arnold et al. 94). A seguir são apresentadas algumas considerações que justificam a grande aceitação deste modelo de tradução.

Nos sistemas *transfer* é empregado um componente de transferência, um sistema de regras que relaciona palavras e estruturas de uma língua em palavras e estruturas de outra. Esta abordagem difere das demais no seguinte sentido:

- sistemas *interlíngua* dispensam este componente, realizando um mapeamento direto do texto na LF para uma representação independente da língua, e a geração desta representação para a LA;
- os sistemas *diretos* assemelham-se aos sistemas *transfer* pela utilização de regras bilíngües, mas realizam a tradução sem a utilização de representações abstratas, realizando o mapeamento diretamente entre o texto na LF para a LA;
- além destas, algumas abordagens tratam da tradução sem a utilização de regras explícitas, mas com base em estatísticas ou exemplos.

Comparando com a abordagem direta, pode-se visualizar pelo menos três vantagens dos sistemas *transfer*. Inicialmente, a extensibilidade: a adição de um novo par de línguas (ou mesmo uma nova direção de tradução) a um sistema direto significa um esforço comparável à criação de um novo sistema. Na abordagem *transfer*, por outro lado, é necessário apenas prover novas descrições das regras de transferência para cada novo "par" e prover os componentes bilíngües de análise e síntese, se ainda não estiverem disponíveis (obviamente esta vantagem não é tão explícita comparada a sistemas *interlíngua*, onde apenas novos componentes monolíngües devem ser criados).

Uma segunda vantagem é que os sistemas *transfer* oferecem uma forma bastante natural de capturar dependências estruturais. Um exemplo é o verbo inglês *to know*, que mormalmente será traduzido para *saber*, em português, se preceder um complemento sentencial, ou *conhecer*, se o complemento for um sintagma nominal. Este tipo de verificação é extremamente complexa de descrever por meio de seqüência de palavras, mas facilmente observável a partir de uma representação estrutural, utilizada nos sistemas *transfer*.

A terceira vantagem é que os mapeamentos bilíngües são muito mais fáceis de descrever, pois certos aspectos da estrutura superficial da sentença são ignorados, e as informações são organizadas em alguma forma de *estrutura de interface* (normalmente uma árvore sintática). Isto significa um conjunto de regras mais simples de construir, entender e atualizar.

Estas duas últimas vantagens podem também ser observadas em relação aos métodos baseados em corpora (estatísticos e baseados em exemplos) que trabalham diretamente sobre os textos fonte e alvo.

Contudo, como estes sistemas não trabalham com regras explícitas, os problemas não são observados em relação ao conjunto de regras, mas em relação ao número de exemplos e ao número de traduções estatísticas alternativas de um termo, necessários para geração dos dicionários. Este problema apenas não irá ocorrer se os textos forem, a priori, reconhecidos, e normalizados (ou marcados) em estruturas hierárquicas, similares às utilizadas em sistemas *transfer* (Brill 93).

Na comparação com os métodos *interlíngua*, a abordagem *transfer* é atraente, principalmente, por evitar a necessidade de projetar a representação do interlíngua. Mesmo em áreas em que a semântica seja bem-definida, existem problemas com o tamanho do vocabulário que o interlíngua irá requerer, pois deve ser capaz de representar qualquer distinção que possa ser feita em qualquer língua (pelo menos naquelas envolvidas no sistema).

Em relação às implementações estudadas, pode-se dizer que médodos baseados em gramáticas livres de contexto são de implementação complexa, principalmente devido à quantidade de regras que devem ser manipuladas para definição das dependências e, conseqüentemente, das regras de transferência. Este problema é gerado em grande parte pelo número de co-referências necessárias para modelagem das divergências.

No sistema TAUM-Aviation, há uma independência total dos níveis de análise e transferência, tanto das suas estruturas de dados quanto dos algoritmos utilizados. Desta forma, as regras de transferência contém vários testes e consistências que já foram realizadas em nível sintático. Para ativar as regras de transformação, o "co-texto" deve ser reavaliado.

No projeto GETA-Ariane, cada nível de manipulação de informações utiliza uma notação (formalismo) próprio. Desta forma, notações diferentes permitem pré-detectar expressões na LF, enquanto outro formalismo é utilizado para descrever o tratamento destas expressões na LA. O algoritmo de transferência é responsável por manipular estas diferentes representações e realizar as conversões necessárias entre estas. É interessante ressaltar que o sistema não trata o problema da polissemia das palavras, mas gera as várias acepções da palavra ambígua e deixa para o processo de geração decidir qual a mais adequada.

As HPSGs utilizadas no projeto METAL, assim como as TAGs, pertencem à classe de gramáticas denominadas Gramáticas Meio Sensíveis ao Contexto (Joshi 85), que são mais poderosas que as GLCs, mas fracamente equivalentes[8] às Gramáticas Sensíveis ao Contexto (Abeillé 88). HPSG é um formalismo lexicalizado, o que permite que toda a manipulação seja representada em nível de estrutura sintática, em que são definidas as regras de análise e transferência estrutural. Neste sistema, a transferência léxica é realizada por um módulo independente, o que exige um caminhamento na estrutura hierárquica para rever a dominância de verbos e nomes.

[8] Duas gramáticas, G_1 e G_2, são ditas *fracamente equivalentes* se a linguagem gerada de G_1, $L(G_1)$, for identica à linguagem gerada por G_2, $L(G_2)$. G_1 e G_2 serão *fortemente equivalentes* se elas forem fracamente equivalentes e para cada $w \in L(G_1) = L(G_2)$, G_1 e G_2 atribuirem a mesma descrição estrutural para w (Joshi 85).

4. O FORMALISMO GRAMÁTICAS SÍNCRONAS DE ADJUNÇÃO DE ÁRVORES

4.1 Representação da Língua Natural

Entre os pesquisadores da lingüística computacional não existe, ainda, um consenso sobre o tipo de gramática que deve ser utilizado para descrição da língua natural. Porém, algumas considerações já podem ser traçadas (Aarts 92, Agustini 95b):

- gramáticas regulares são muito simples e são capazes de representar apenas subconjuntos restritos da língua;
- gramáticas irrestritas são muito complexas, logo pouco úteis;
- gramáticas livres de contexto (GLC) geram linguagens livres de contexto, e a língua natural não é livre de contexto (Aarts 91);
- gramáticas sensíveis ao contexto (GSC) permitem descrever os problemas de dependências da língua natural, porém possuem uma alta complexidade de reconhecimento, o que torna o custo de utilização muito alto.

Por outro lado, existe um consenso de que grandes subconjuntos da língua natural são livres de contexto, e a gramática para descrever a língua natural deve ser apenas um pouco mais poderosa que uma GLC (Aarts 92).

Assim, a grande maioria das pesquisas atualmente propõe trabalhar em modelos que se situam em um nível intermediário entre as GLCs e as GSCs, tentando obter, ao mesmo tempo, uma maior capacidade de representação, com construções que permitam modelar dependências, e um modelo computacional viável.

Estas pesquisas derivaram o que passou a ser denominado Gramáticas Meio-Sensíveis ao Contexto e têm apresentado bons resultados para o processamento da língua natural (Joshi 85, Steedman 93, Abeillé 88). Dentre estes formalismos pode-se destacar as Gramáticas Categoriais (Steedman 93), Gramáticas Lineares Indexadas (Aho 69), Head

Phrase Structure Grammars (Pollard & Sag 87) e as Gramáticas de Adjunção de Árvores (TAGs) (Joshi, Levy & Takahashi 75).

4.2 Gramática de Adjunção de Árvores

O formalismo de Gramáticas de Adjunção de Árvores (TAGs) foi introduzido em 1975 por Aravind Joshi (Joshi, Levy & Takahashi 75, Joshi 85). Trabalhos têm sido desenvolvidos com a descrição de gramáticas para as línguas francesa (Abeillé 88), inglesa (Schabes et al. 88) e portuguesa (Kipper & Strube de Lima 94), entre outras. As principais características deste formalismo são:

i. TAG é um sistema de geração de árvores, a partir de um conjunto finito de árvores que definem a gramática da língua, dividido em árvores iniciais e árvores auxiliares, e duas operações de composição: adjunção e substituição;

ii. TAGs possuem um mecanismo de recursão próprio. Nas árvores elementares é estabelecido o domínio de dependências, que são relações entre os elementos das árvores elementares ou mesmo relações entre estas. A operação de adjunção possibilita a recursão, preservando estas dependências.

Formalmente, uma TAG consiste de uma quíntupla $T=(\Sigma, NT, I, A, S)$, onde:

- Σ = conjunto finito de símbolos terminais
- NT = conjunto finito de símbolos não-terminais, $\Sigma \cap NT = \varnothing$
- S = símbolo inicial, $S \in NT$
- I = conjunto finito de árvores iniciais
- A = conjunto finito de árvores auxiliares

O conjunto de árvores iniciais corresponde a estruturas lingüísticas mínimas que não contenham recursão, constituindo estruturas frasais completas de sentenças simples. Todos os nodos da fronteira contêm terminais (pré-terminais) ou não-terminais, neste caso marcados para operação de substituição.

As árvores auxiliares, por sua vez, são estruturas recursivas mínimas que possibilitam a construção de frases mais complexas. A fronteira deve conter um nodo não-terminal com o mesmo rótulo da raiz da árvore e marcado para adjunção, conforme definido a seguir; neste nodo são realizadas as operações de recursão.

O conjunto I ∪ A é chamado conjunto de árvores elementares e correspondem às estruturas lingüísticas mínimas, porém completas, que localizam as dependências como concordância e subcategorização.

Sentenças geradas a partir de uma língua definida por uma TAG são derivadas a partir da composição de uma árvore inicial, com raiz S, e árvores elementares, através das operações de substituição e adjunção.

A árvore que é resultado da composição de outras duas árvores é chamada de árvore derivada. Define-se o conjunto de árvores de uma TAG G, $\tau(G)$ como sendo o conjunto de todas as árvores derivadas a partir da árvore inicial com raiz S em I.

A adjunção é a principal operação em TAGs. Ela cria uma nova árvore a partir de uma árvore auxiliar β e uma árvore α (α é qualquer árvore, inicial, auxiliar ou derivada). Considerando-se α uma árvore contendo um nodo n nomeado por X e β uma árvore auxiliar cuja raiz é também nomeada por X (Schabes et al. 88, Joshi 92). A árvore derivada γ, obtida pela adjunção de β em α no nodo n é construída como ilustrado na figura 4.1:

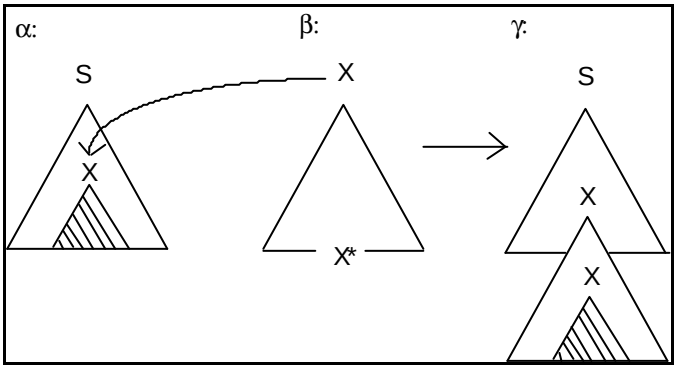


Figura 4.1 - Operação de adjunção (Joshi 92)

A operação de substituição usada em TAG é a mesma operação de reescrita das GLCs. Na operação de substituição, um não-terminal folha, marcado para substituição em uma árvore inicial, é substituído pelo nodo raiz de uma outra árvore inicial, produzindo uma nova árvore, conforme figura 4.2.

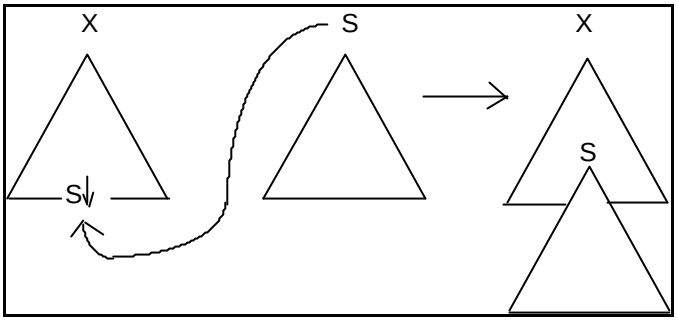


Figura 4.2 - Operação de substituição

4.2.1 TAGs Lexicalizadas

Nas gramáticas lexicalizadas (Schabes et al. 88, Schabes & Joshi 90, Joshi 92) cada estrutura elementar é associada a um item léxico, chamado de âncora da estrutura. Por lexicalizada entende-se que, em cada estrutura, existe pelo menos um item léxico que é realizado. A gramática consiste de um dicionário onde cada item léxico é associado a um número finito de estruturas para as quais este item é a âncora. Portanto, a gramática está na forma lexicalizada, se consiste de:

- um conjunto finito de estruturas, cada uma associada a um item léxico (cada item léxico será chamado de âncora para as estruturas correspondentes);
- uma ou mais operações para compor as estruturas.

Considera-se TAGs lexicalizadas (LTAGs) como uma instância de gramáticas lexicalizadas. As propriedades formais de TAGs são mantidas para LTAGs. Portanto, podemos trabalhar exclusivamente com LTAGs. Nas LTAGs as árvores elementares são associadas a itens léxicos (ditos âncoras das árvores). Cada item léxico pode conter múltiplas entradas no dicionário conforme haja possíveis categorias ou estruturas de argumentos (Schabes et al. 88).

4.2.2 TAGs com Atributos

Uma forma de tornar as gramáticas mais precisas sem aumentar em muito o número de regras é a associação de traços às palavras e definição de restrições na forma como estes traços podem interagir. Os traços podem conter diversos tipos de informações, tais como gênero e número. Por exemplo, pode-se assumir as seguintes entradas no dicionário (Beardon, lumsden & Holmes 91):

trem	= substantivo, singular
trens	= substantivo, plural
passou	= verbo, pretérito, singular
passaram	= verbo, pretérito, plural

Uma operação de unificação opera sobre dois objetos de estruturas semelhantes, como, por exemplo, descrições de objetos lingüísticos, e tenta combiná-los de forma a criar um objeto único. Se duas descrições forem compatíveis, o resultado da unificação irá conter os traços de ambos objetos (Beardon, Lumsden & Holmes 91).

Em gramáticas de unificação, uma estrutura de traços é associada a cada nodo em uma árvore de derivação a fim de descrever este nodo e suas relações com os traços dos outros nodos na árvore de derivação (Vijay-Shanker Joshi 88b). Em TAGs baseadas em traços (FTAGs), a cada nodo de uma árvore elementar são associadas duas estruturas que indicam as relações do nodo com os nodos que o dominam, chamado *top* (t, por convenção) e do nodo com os nodos que ele domina, chamados *bottom* (b, por convenção).

FTAGs atribuem duas estruturas de traços a cada nodo de adjunção em uma árvore elementar, *top* e *bottom*, e somente uma para os nodos de substituição, *top*. Quando a derivação está completa, as estruturas de traço *top* e *bottom* de todos os nodos são unificadas simultaneamente (Schabes & Joshi 90).

As noções de substituição e adjunção precisam ser ampliadas para esta nova estrutura. A estrutura de traços de um novo nodo criado por substituição herda a união dos traços dos nodos originais. O traço top no novo nodo é a união dos traços dos nodos originais, enquanto o traço *bottom* da nova árvore é simplesmente o traço *bottom* do nodo *top* da árvore a ser substituída (ver figura 4.3).

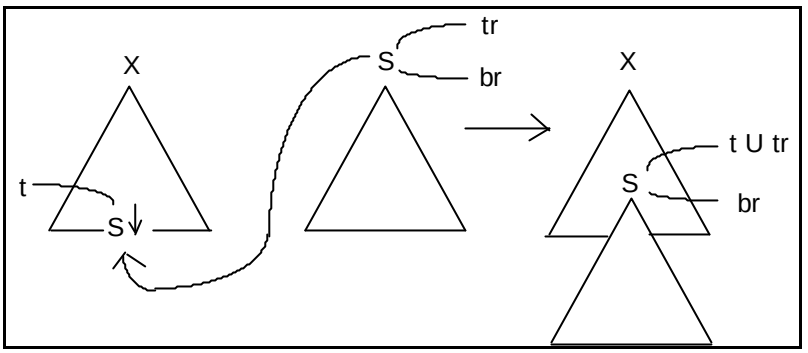


Figura 4.3 - Operação de substituição em FTAG (Becker et al. 94)

Na operação de adjunção, o traço top do nodo raiz da árvore auxiliar unifica com o traço *top* do nodo que recebe a adjunção, enquanto o traço *bottom* unifica com o traço *bottom* do nodo pé adjuntado (ver figura 4.4).

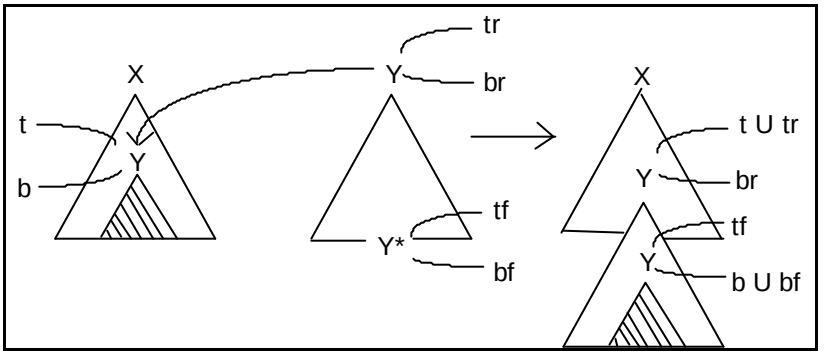


Figura 4.4 - Operação de adjunção em FTAG

A implementação do formalismo TAG em uma estrutura de unificação permite dinamicamente especificar restrições locais que, de outra forma, deveriam ser modeladas estaticamente nas gramáticas. Restrições que verbos fazem em seus complementos, por exemplo, podem ser implementadas através das estruturas de traços.

As figuras 4.5 a 4.7 apresentam um exemplo, adaptado de (Kipper 94), de reconhecimento da sentença *João parece dormir*, com todos os traços semânticos associados.

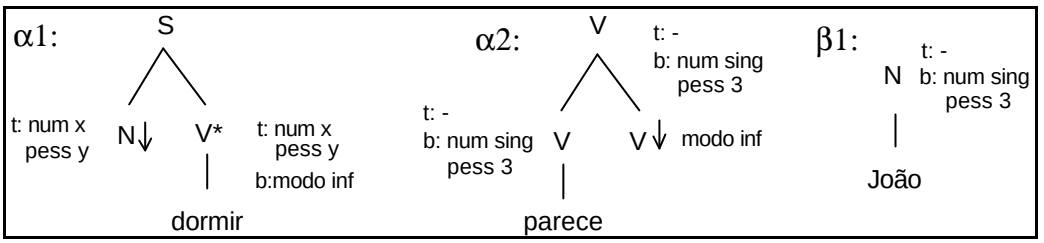


Figura 4.5 - Árvores elementares para *João parece dormir*

A árvore apresentada na figura 4.6 é gerada após a operação de adjunção sobre o verbo V, e a substituição do nome N.

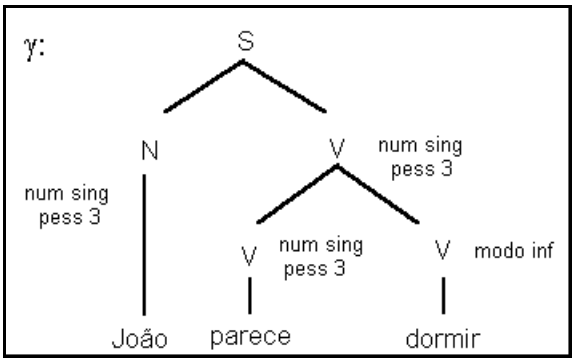


Figura 4.6 - Árvore derivada para *João parece dormir*

A árvore da figura 4.6 é submetida ao módulo de unificação semântica, resultando na estrutura sintática marcada apresentada na figura 4.7.

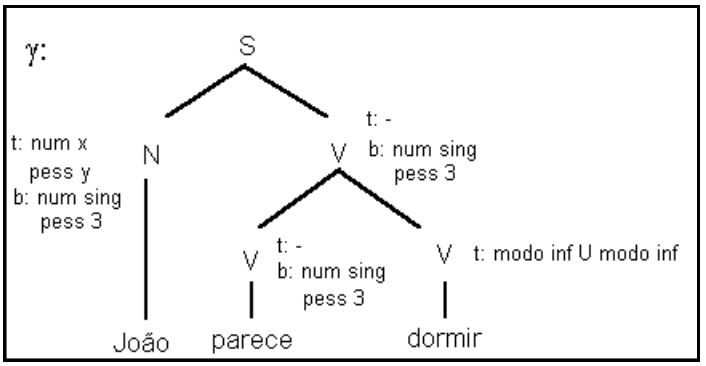


Figura 4.7 - Unificação de *João parece dormir*

4.3 Sistemas Síncronos de Reescrita

As definições abaixo formalizam o conceito de sistemas síncronos de reescrita (Harbusch & Poller 93, Rambow & Satta 96), generalizando o conceito de gramáticas síncronas definido inicialmente por Shieber e Schabes para TAGs (Shieber & Shabes 90).

Definição 1: Um Sistema de Reescrita (SR) é um par S = (A(S), P(S)) onde A(S) é um alfabeto finito que consiste de um conjunto de não-terminais N(S) e um conjunto de terminais T(S) ($A(S) := N(S) \cup T(S)$, $N(S) \cap T(S) = \varnothing$). $P(S) \subseteq A(S)^* N(S)^+ A(S)^* \times A(S)^*$ é um conjunto finito de produções.

Definição 2: Seja S = (A(S), P(S)) um SR e $w,v \in A(S)^*$. v é dito diretamente derivável de w ($w \Rightarrow v$) se e somente se (sse) existem palavras u_1, u_2, $p, q \in A(S)^*$, sendo $w = u_1 p u_2$; $v = u_1 q u_2$ e $(p,q) \in P(S)$.

Além disto, v é dito derivável de w ($v \rightarrow w$) sse existe uma seqüência de palavras $w = w_0, w_1, .., w_r = v$ com $r \in \aleph$, $w_i \in A(S)^*$ ($0 \leq i \leq r$) e $w_i \Rightarrow w_{i+1}$ ($0 \leq i \leq r-1$).

Definição 3: Um Sistema de Reescrita Síncrono (SRS) consiste de 2 sistemas de reescrita G_1 e G_2 e um conjunto de ligações L. Cada produção em $P(G_1)$ é relacionada à uma (ou mais) produção correspondente em $P(G_2)$. Uma ligação ($\in$ L) é definida entre um não-terminal de uma produção em $P(G_1)$ e um não-terminal da produção relacionada em $P(G_2)$.

Definição 4: Seja S = (G_1, G_2) um sistema de reescrita síncrono. Os termos diretamente derivável e derivável são restritos à aplicação em paralelo aos sistemas de reescrita sincronamente relacionados da seguinte forma: os dois nodos onde as operações de recursão são realizadas devem possuir uma ligação; as produções utilizadas devem estar relacionadas; após a realização da operação a ligação é eliminada.

4.3.1 STAGs para Tradução Automática

Buscando a utilização das LTAGs com Atributos para a tradução automática, foi introduzida uma nova extensão ao formalismo, chamada *Synchronous* TAGs (STAGs): um

sistema de reescrita síncrono, onde as duas línguas são representadas por LTAGs, de forma a possibilitar a inclusão de um módulo de transferência associado às operações realizadas sobre as LTAGs (Abeillé et al. 90, Shieber & Shabes 90).

O módulo de transferência é ativado sempre que uma operação realizada na derivação da sentença fonte "casar" com alguma estrutura da LTAG alvo, de acordo com uma pesquisa em um dicionário bilíngüe que controla a aplicação das regras de transferência. As estruturas pertencentes à gramática da LF são chamadas árvores sintáticas, enquanto que as estruturas pertencentes à gramática-alvo são chamadas de árvores semânticas.

Um dicionário de transferência coloca em correspondência árvores da gramática fonte, instanciadas por inserções léxicas (nodos e atributos), com árvores da gramática alvo. O método se desenvolve conforme os seguintes passos básicos (Shieber & Shabes 90): análise do texto-fonte, tradução incremental e geração. A análise é realizada de acordo com o conjunto de árvores que descrevem a gramática-fonte; cada árvore elementar da derivação é considerada com todos os atributos adquiridos pela unificação.

A árvore de derivação gerada pela análise é traduzida incrementalmente para uma árvore de derivação correspondente na LA. Cada árvore fonte é traduzida para uma árvore alvo de acordo com o dicionário de transferência. Finalmente, a sentença é gerada a partir da árvore de derivação alvo.

Ligações entre os nodos sintáticos e semânticos significam que uma operação sobre este nodo na árvore sintática tem uma combinação equivalente na árvore semântica. Assim, a representação semântica é construída de forma síncrona à derivação sintática, através da escolha de pares de árvores elementares (um nodo sintático, na árvore fonte, e um nodo semântico, na árvore alvo) das gramáticas. A figura 4.8 apresenta o algoritmo básico do processo de tradução.

1. escolher uma ligação entre dois nodos n_1 e n_2 no par <árvore sintática A_1, árvore semântica A_2>;
2. escolher um par de árvores <B_1,B_2> no dicionário bilíngüe;
3. gerar o par resultante <$B_1(A_1,n_1)$,$B_2(A_2,n_2)$>, onde B(A,n) é o resultado da realização de uma operação de adjunção ou substituição na árvore A, sobre o nodo n, utilizando a árvore B;
4. se o par (A_1, A_2) contiver outras ligações estas devem ser mantidas,

na árvore resultante.

Figura 4.8 - Algoritmo de análise das STAGs

A seguir é apresentado um exemplo de aplicação deste método para o par <inglês,francês>, proveniente de (Abeillé et al. 90). A tradução é realizada a partir do fragmento de dicionário de transferência apresentado na figura 4.9.

Inicialmente, supondo que a análise do texto-fonte derivou uma estrutura correspondente ao par (γ), conforme figura 4.9. Ao operar a árvore inicial (α) sobre o nodo da gramática inglesa NP_0 o módulo de transferência ativará a mesma operação sobre o nodo da gramática francesa NP_1.

A seguir é operado o par β na ligação $<NP_1\text{-}NP_0>$, sobre o resultado anterior, e o par ($\alpha 1$) é gerado, resultando na estrutura ($\alpha 1$) da figura 4.10.

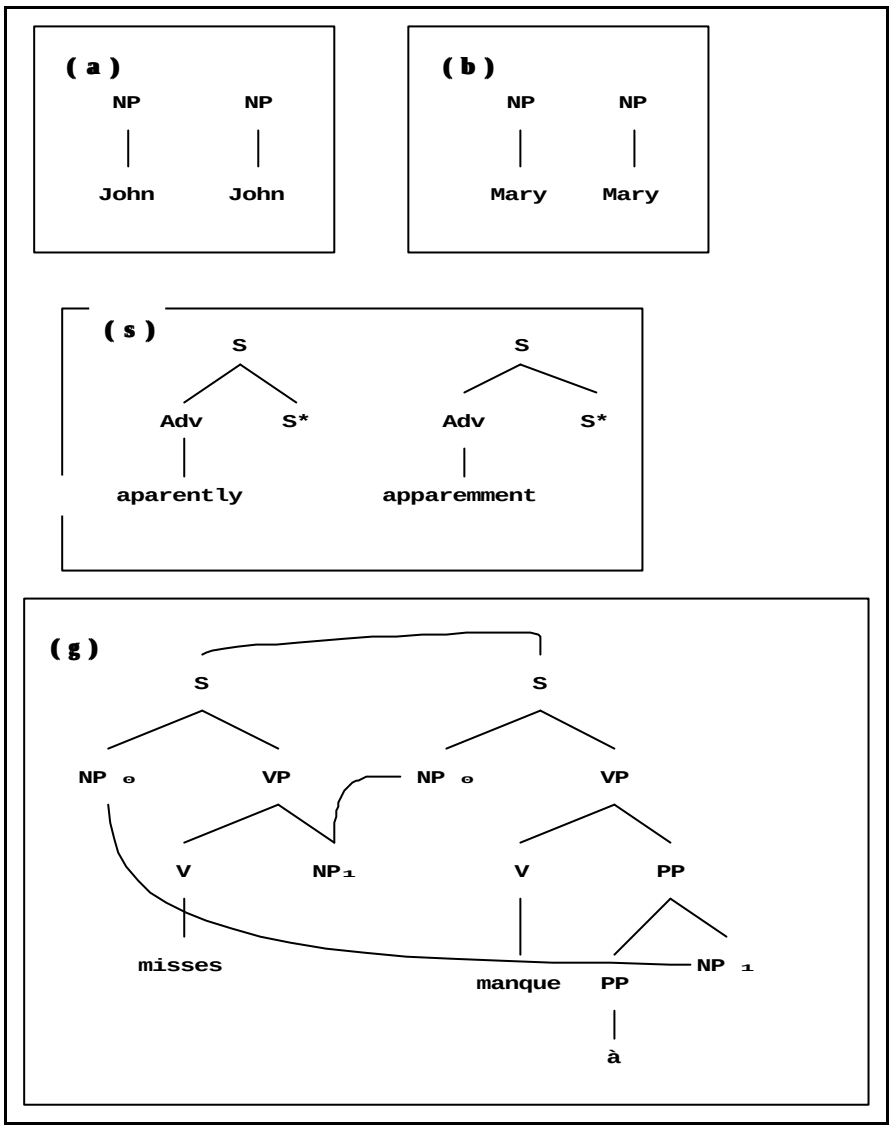


Figura 4.9 - Dicionário bilíngüe para LTAGs (Abeillé et al. 90)

Finalmente, quando é operado o par (σ) na ligação <S-S>, sobre ($\alpha 1$), o par ($\alpha 2$) é gerado e a árvore semântica (árvore de derivação na LA) está completa.

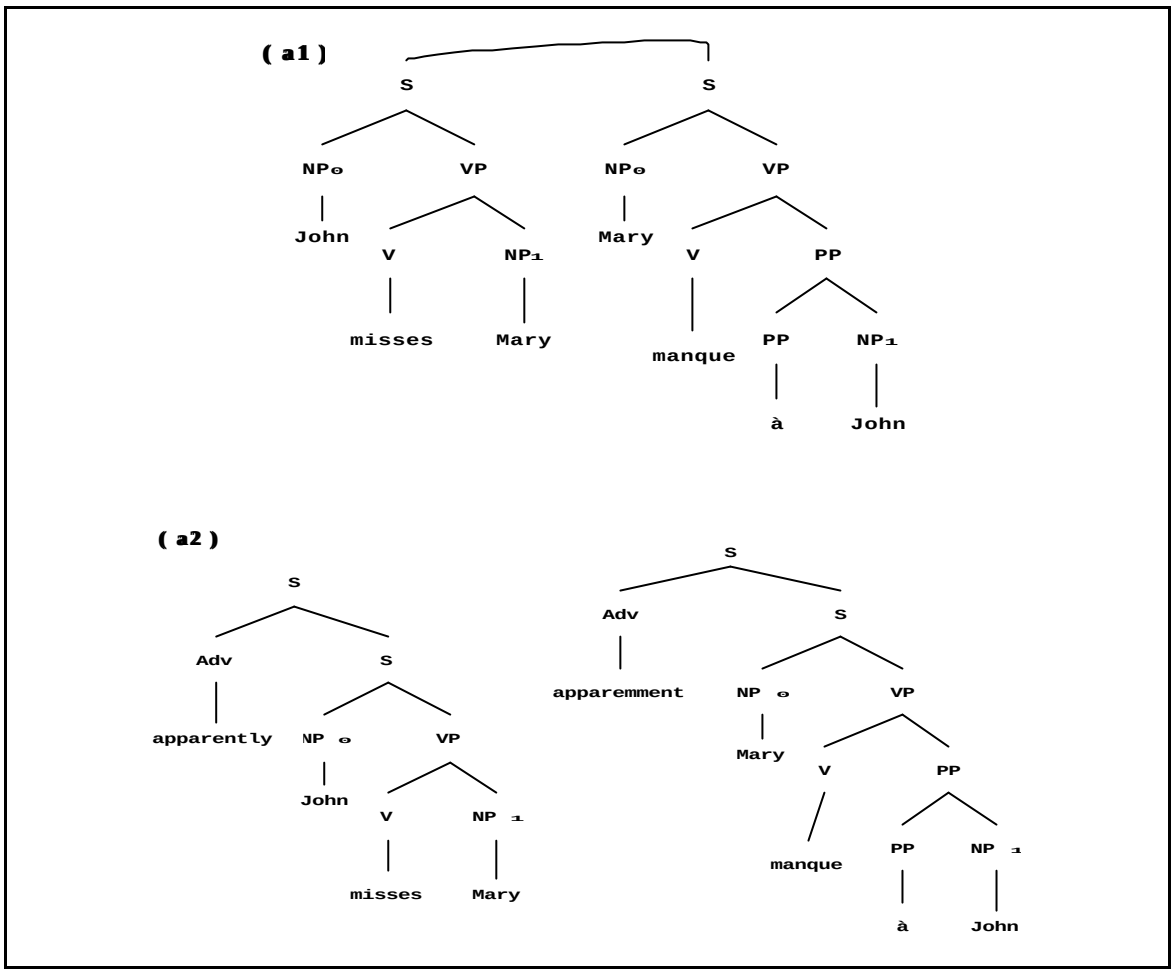


Figura 4.10 - Seqüência de operações para a sentença *Apparently, John misses Mary*

4.3.2 Apropriação ao Português

As primeiras investigações sobre a aplicabilidade do formalismo para a língua portuguesa foram realizadas por Kipper, através da descrição de uma gramática TAG para um subconjunto extenso das estruturas sintáticas; a construção de um analisador sintático-semântico e uma proposta de lexicalização do trabalho desenvolvido (Kipper 94).

STAGs, por outro lado, foi inicialmente proposto por Shieber e Schabes como um sistema para geração de estruturas semânticas (Shieber & Shabes 90) e posteriormente utilizado para implementação de módulos de transferência para tradução automática entre as línguas inglesa e francesa (Abeillé et al. 92b) e entre as línguas coreana e inglesa (Egedi & Palmer 94).

Desta forma, a apropriação do formalismo STAGs à língua portuguesa passa por dois desafios: a aplicação dos conceitos desenvolvidos por Kipper para possibilitar a criação de um módulo de transferência, e a investigação da viabilidade de utilização do formalismo, uma vez que poucos experimentos estão disponíveis na literatura.

5. PROPOSTA DE UM MÓDULO DE TRANSFERÊNCIA UTILIZANDO STAGS

5.1 Definição de um Corpus de Divergências

O primeiro passo para definir a proposta de um modelo de sistema de tradução automática se constituiu na definição e análise de um corpus bilíngüe que possibilitasse a seleção de um conjunto de divergências sintáticas existentes na tradução entre as línguas portuguesa e inglesa. A escolha de um corpus real não é uma tarefa simples, uma vez que exige sentenças completas, de grau de complexidade razoável e preferencialmente não ambíguas.

Devido à grande dificuldade em encontrar-se um corpus bilíngüe disponível com tais características, optou-se por construir um corpus piloto, realizando a tradução[9], marcação[10] e alinhamento[11] do mesmo, manualmente.

Assim, foram selecionadas aleatoriamente 200 sentenças, a partir de cerca de 3.000 manchetes (títulos) de notícias da área econômico-financeira distribuídas pela Agência Estado através do Sistema Broadcast[12] durante o primeiro semestre de 1996. Destas, foram descartadas 110 sentenças por se tratarem de referências a títulos anteriores, repetições ou por estarem incompletas (mal-formadas).

A seguir são apresentadas algumas características gerais do corpus selecionado:

- *baixo grau de ambigüidade*: por se tratar de títulos de um sistema de distribuição de notícias, não costumam ocorrer problemas de ambigüidade semântica nas sentenças trabalhadas, porém isto não elimina

[9] A tradução foi realizada por um tradutor nativo da língua inglesa e com experiência na área de aplicação.

[10] Rotular as palavras com suas características morfo-sintáticas e realizar a estruturação destes elementos em uma estrutura frasal.

[11] Relacionar os elementos marcados da língua-fonte com sua realização na língua-alvo.

[12] Sistema de distribuição eletrônica de mensagens, disponível por contrato de serviço.

a ambigüidade gerada durante o processamento computacional das mesmas. A ambigüidade gerada pelo analisador sintático levou à necessidade de estender-se o módulo de análise sintática utilizado, com a finalidade de reduzir o número de estruturas geradas, através da lexicalização de seu mecanismo de seleção de estruturas sintáticas;

- *referências anafóricas*: como as sentenças possuem conteúdo completo e são independentes (diferentemente do texto associado aos títulos), não ocorrem referências anafóricas dentro do corpus selecionado. Isto restringe bastante a complexidade do módulo de análise semântica;
- *ambigüidade translacional*: a correta seleção do item léxico é um problema crítico. Como este não é objetivo principal do trabalho, decidiu-se restringir o problema, sempre que possível. No dicionário bilíngüe são incluídas apenas as ocorrências ambíguas que efetivamente aparecem no corpus e são geradas sentenças alternativas para cada palavra ambígua encontrada. Para uma solução mais efetiva deste problema é necessário modificar o mecanismo baseado em traços, de forma a permitir a manipulação de informações semânticas mais complexas.

A partir da análise do corpus foi definido um conjunto de problemas (divergências) que devem ser tratados pelo processo de transferência. O anexo A apresenta o corpus trabalhado e algumas considerações sobre a quantidade relativa de cada classe de divergência observada. Os casos observados são listados a seguir:

a) *ordem dos constituintes: adjetivos e locuções adjetivas*

Uma divergência que ocorre com freqüência, no corpus trabalhado, é o problema da ordem dos constituintes. O primeiro caso ocorre no posicionamento do *adjetivo.* É importante ressaltar que, em todos os casos observados, o adjetivo ocorre após o nome. Assim, uma regra geral de transformação associada a esta estrutura pode ser herdada por todas palavras desta classe. A seguir são apresentadas algumas destas ocorrências:

- peso *mexicano* ⇒ *mexican* peso
- pacote *fiscal* ⇒ *fiscal* package
- iene *forte* ⇒ *strong* yen
- dolar *comercial* ⇒ *commercial* dollar

- salário *mínimo* ⇒ *minimum* salary

Um segundo problema de ordem dos constituintes ocorre em relação às *locuções adjetivas* (adjunto adnominal), conforme os exemplos abaixo:

- presidente *da Irlanda* ⇒ *Irish* president
- bolsa *de Nova York* ⇒ *New York* stock exchange
- propaganda *de páscoa* ⇒ *Easter* advertising
- funcionários *da receita* ⇒ *inland revenue* employees

É importante salientar, contudo, que no caso de complemento nominal, embora as estruturas superficiais sejam similares, esta regra não será válida, conforme os exemplos:

- liquidação *do banco Banespa*⇒ liquidation *of the Banespa bank*
- plano *de abertura econômica*⇒ plan *to open economy*
- redução *do déficit* ⇒ reduction *of deficit*
- falta *de recursos* ⇒ lack *of resources*

Pode-se dizer que a diferença entre o adjunto adnominal e o complemento nominal está associada à transitividade das palavras. O adjunto adnominal é um termo acessório (dispensável) ao sentido da frase, logo é constituinte de palavras *intransitivas*, enquanto que o complemento nominal é indispensável para a compreensão do sentido, estando associado a palavras *transitivas* (Barros 91). Desta forma, o tratamento deste tipo de divergência será feito a partir da instanciação léxica realizada e restrita por seus atributos semânticos.

b) *divergência léxica e vazio léxico*

O problema da divergência léxica conceitual (Hutchins & Somers 92, Arnold et al. 94, Leffa 95), apresentado no item 2.4.1, e definido como uma divergência léxico-semântica por Dorr (Dorr 93) (item 2.4.3.5), ocorre com bastante freqüência, possivelmente por se tratar de um domínio específico e com jargão próprio, como por exemplo:

- bolsa ⇒ stock exchange

- corretora ⇒ firm of brokers
- receita ⇒ inland revenue
- empreiteiras ⇒ contract construction companies
- metalúrgicos ⇒ metal workers

O tratamento deste tipo de divergência é complexo dentro do formalismo STAGs, pois o mesmo é baseado em uma correspondência nodo-a-nodo entre as árvores elementares (estruturas mínimas e completas) que descrevem cada uma das línguas. A correspondência nodo-a-nodo garante uma árvore de derivação alvo completa. As ligações dentro de cada árvore especificam que cada operação (adjunção ou substituição) executada em um nodo da árvore fonte será transferida ao nodo correspondente na árvore-alvo.

No caso acima esta correspondência não é válida, pois ela é estabelecida entre uma árvore elementar e um fragmento da árvore de derivação. Desta forma, os algoritmos devem ser redefinidos de forma a garantir as propriedades da árvore-alvo, que em um segundo passo será validada sob o ponto de vista estrutural. Este processo inclui verificações como: operações não podem ser realizadas sobre um mesmo nodo mais de uma vez, árvore derivada não contém nodos que não foram derivados, unificação dos atributos na árvore derivada é válida.

c) *seleção de preposições*

Segundo Celso Luft (Luft 91), a variabilidade no uso das preposições não é caprichosa, aleatória, mas semanticamente governada: são os traços semânticos da palavra regente, primários ou secundários, que comandam a ocorrência desta ou daquela preposição. Ou seja, a preposição é efeito da palavra-núcleo da estrutura, via semântica.

Desta forma, a correta seleção da preposição associada a um sintagma preposicional, como objeto indireto de um verbo ou adjuntos adverbiais, ocorre a partir de uma verificação semântica associada ao tipo do predicado, normalmente de acordo com a definição do verbo ou predicado que a determina. Por exemplo:

- *fecha em* leve queda ⇒ *closes at* slight fall
- *investe em* propaganda de Páscoa ⇒ *invests in* Easter advertising

- *derruba* bolsa *em* 311 pontos ⇒ *lowers* stock exchange *by* 311 points

d) *Conflacional*

Finalmente, um conjunto significativo de sentenças foi traduzido utilizando um estilo próprio ao domínio de aplicação (manchetes de notícias), o que sugere a necessidade de uma forma de tratamento adequado a este tipo de situação. Por exemplo:

(1) banco central *faz hoje leilão* de nbc cambial.
central bank *to auction* nbc bounds *today*.

(2) ministros *discutem hoje* a reforma tributária.
ministers *to discuss* tax payers reform *today*.

(3) metalúrgicos do rio *fazem assembléia amanhã*.
rio metal workers *to meet tomorrow*.

Alguns fenômenos podem ser observados a partir dos exemplos acima:

- nos exemplos (1) e (3) o verbo mais o seu objeto direto (complemento nominal) foram traduzidos como verbo principal em Inglês, determinando uma divergência *conflacional*, conforme item 2.4.3.1;
- a forma verbal, presente do indicativo, não foi herdada pelo processo de tradução, passando-se o verbo para o infinitivo;
- nos exemplos (1) e (2) o adjunto adverbial de tempo foi deslocado para após o predicado verbal.

O corpus utilizado não permitiu detectar um padrão de tratamento para as ocorrências de deslocamento do adjunto adverbial. O formalismo modela corretamente a ocorrência da divergência conflacional, porém não foi possível definir uma regra de tradução adequada para o caso do adjunto adverbial, que necessita, desta forma, ser tratado pós-transferência.

e) *casos não tratados*

Alguns casos observados no corpus não foram tratados por exigir um estudo além do escopo deste trabalho:

- caso possessivo: apenas uma das sentenças foi traduzida utilizando a forma possessiva da língua inglesa; como este tratamento é complexo (Eynde 93, Arnold et al. 94), descartou-se tradução e foi utilizada a forma normal da sentença em inglês;
- nomes próprios e siglas: a convenção a ser utilizada na forma de tratamento de referências a pessoas é, em muito, uma questão de estilo. Neste trabalho, decidiu-se pela utilização do *original transposto*, ou seja, o nome próprio, ou sigla, utilizado é empregado como equivalente na LA.

5.2 Modelo Computacional Proposto

A partir do estudo das classes divergências discutidas no item anterior, a figura 5.1 apresenta uma proposta de arquitetura para o módulo de transferência de um sistema de tradução automática entre as línguas portuguesa e inglesa.

Podemos observar que em relação à figura 3.5 apresentada na seção 3 alguns comentários se fazem necessários. A figura 3.5 apresentou uma arquitetura genérica de sistemas *transfer*, onde três fases distintas podem ser visualizadas: análise, transferência e geração. Para a utilização de STAGs, contudo, esta arquitetura foi modificada pois, devido à característica incremental do método, o módulo de transferência é responsável, ainda, pela manipulação da análise sintática.

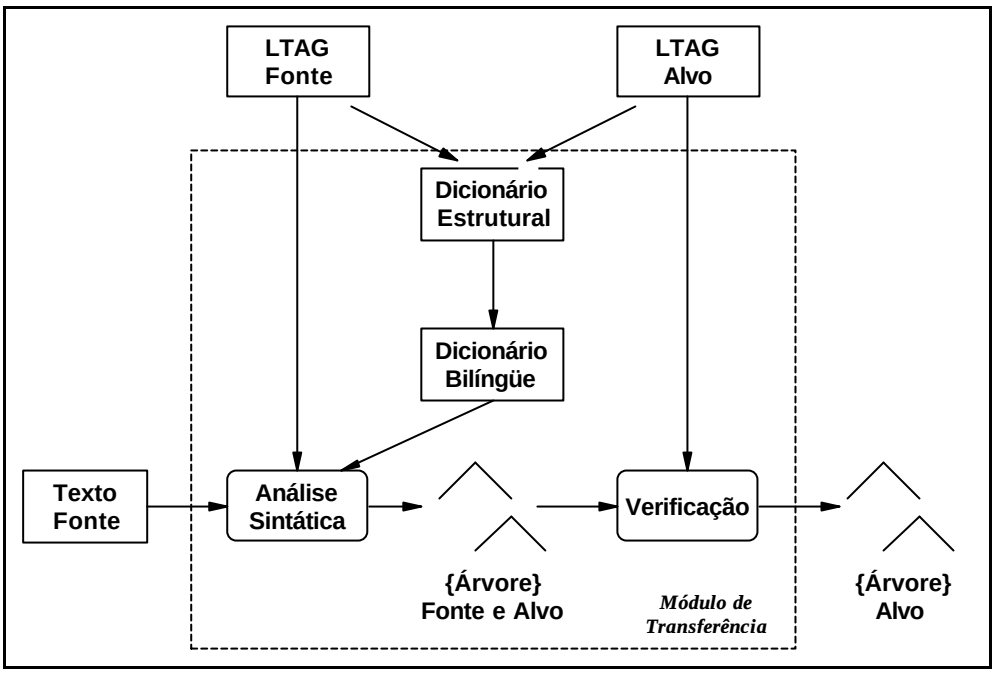


Figura 5.1 - Módulo de transferência proposto

O princípio básico do módulo de transferência é a utilização da árvore de derivação para realizar o mapeamento de uma língua para outra. Dada uma possível árvore de derivação para uma sentença na LF, a árvore alvo é construída a partir de uma correspondência nodo-a-nodo entre as árvores de derivação, isto é, a correspondência é realizada entre as árvores elementares, preservando as relações de domínio entre os nodos da árvore fonte.

O módulo de transferência recebe, como entrada, uma seqüência de itens léxicos gerada pelo analisador léxico-morfológico a partir da sentença de entrada. Um item léxico contém as informações:

- categoria gramatical;
- forma canônica (infinitivo para verbos, singular/masculino para substantivos e adjetivos, etc);
- conjunto de traços semânticos associados (gênero, número, modo, etc); e
- *lista de* árvores elementares para as quais o item subcategoriza (é o elemento âncora).

A saída corresponde à árvore de derivação na LA, com todas as modificações estruturais realizadas, e *decorada* com os traços semânticos herdados dos itens léxicos da LF.

As gramáticas FB-LTAG, fonte e alvo, são descritas de forma independente e utilizam a notação introduzida por (Kipper 94). A gramática da língua portuguesa utilizada é a descrita em (Kipper 94). Para a gramática inglesa foi adaptada a descrição apresentada em (Becker et al. 94). A figura 5.2 apresenta a gramática utilizada para descrição das LTAGs, tanto na língua portuguesa quanto na língua inglesa.

```
GrLTAG::=       { Arv }+
Arv             ::=     ÁrvoreID ";" Nome_nodo "(" Descr_arv ")"
Descr_arv       ::=     { Nome_nodo   [ "[" "t" Traços_unific "]" ]
                        [ "[" "l" Traços_unific "]" ]
                        [ "{" restrições "}" ]
                        [ Sub_árvore ]                  }+
Sub_árvore      ::=     "(" Descr_arv ")"
Traços_unif     ::=     Lista_Traços "=" Offset_nodo [ ";" Traços_unif ]
Lista_traços    ::=     TraçoID [ "," lista_Traços ]
Restrições      ::=     OpRel "(" Item_léxico { "," Item_léxico }+ ")"
Nome_nodo       ::=     Identificador Tipo_nodo
Tipo_nodo       ::=     "◊" | "↓" | "*" | ε
OpRel           ::=     "=" | "!="

Na notação EBNF utilizada:
        - { x }+        → uma ou mais ocorrências de "x";
        - [ x ]         → a ocorrência de "x"é opcional
        - "x"           → ocorrência do item léxico "x"
        - x | y         → ocorrência de "x" ou "y"
        - ε             → produção vazia
```

Figura 5.2 - EBNF para descrição das LTAGs

A figura 5.3 apresenta a descrição das árvores iniciais que subcategorizam verbos bitransitivos (com objeto direto e objeto indireto) em Português.

```
p1: S(  N0 [Num=x, Gen=y, Pess=z]
        V◊ [13] [Num=x, Pess=z, Modo=Ind]
        N1↓
        Prep(+Art[14]) [Num=a, Gen=b]
```

[13] Na notação utilizada, ◊ indica o elemento âncora da estrutura; ↓: nodo marcado para substituição; *: nodo marcado para adjunção.

[14] Na descrição apresentada por (Kipper 94a), ao realizar uma contração Prep + Art, o nodo preposição incorpora os traços semânticos do artigo.

```
N2 ↓ [Num=a, Gen=b, Pess=c])
```

Figura 5.3 - Verbos bitransitivos em Português

No exemplo da figura 5.3, 'p1' é um identificador único da árvore que está sendo descrita; os índices associados aos sintagmas indicam o papel desempenhado dentro da sentença (0: sujeito; 1: objeto direto; 2: objeto indireto); durante o processo de unificação, os traços semânticos são combinados em uma estrutura única, e as variáveis de mesmo nome devem unificar (neste caso, conter o mesmo valor). No exemplo da figura 5.3, o sujeito deve concordar em número e pessoa com o verbo, e a preposição (contração Prep + Art) deve concordar em número e gênero com o objeto indireto.

Em Inglês (ver figura 5.4), são apresentadas as árvores para verbos bitransitivos, que podem subcategorizar para duas estruturas diferentes. No caso da estrutura 'i1' a preposição associada à preposição (P2) deve ser o item léxico 'to', enquanto 'i2' subcategoriza os verbos que permitem o deslocamento do objeto indireto (PP-*shift* (Becker et al. 94)). Esta ambigüidade pode ser observada no verbo *to give* (dar), conforme exemplo abaixo:

- João deu flores a Maria
- John gave flowers to Mary
- John gave Mary flowers

```
i1:  S ( NP0
         VP (     V ◊
                  NP1 ↓
                  PP2(     P2  [ lex = #to ]
                           NP2 ↓              ) ) )

i2:  S ( NP0
         VP (     V ◊
                  NP2 ↓
                  NP1↓ ) )
```

Figura 5.4 - Verbos bitransitivos em Inglês

Um *Dicionário Estrutural*, ou dicionário de divergências sintáticas, mantém o conjunto de correspondências (ligações) nodo-a-nodo entre as gramáticas das línguas fonte e alvo. Todas as árvores elementares da LF devem estar associadas às árvores da LA. São

incluídas, ainda, informações sobre a herança dos traços semânticos. Ao construir-se este dicionário são mapeadas as divergências sintáticas.

O *Dicionário Bilíngüe* constitui o dicionário de tradução propriamente dito. Este dicionário manipula os pares de itens léxicos e aponta para uma ou mais estruturas elementares do dicionário de divergências sintáticas para as quais o item é âncora. Neste dicionário podem ser definidos fragmentos de árvores de derivação, com a finalidade de resolver as divergências léxico-semânticas. Além disto, o dicionário pode reescrever as regras contidas no dicionário de divergências sintáticas, para atender às restrições impostas pelo item léxico utilizado.

O módulo de *Análise Sintática* é responsável pela execução dos algoritmos de análise e transferência, gerando, para cada árvore solução na LF, uma ou mais árvores derivadas na LA. O algoritmo executado é o apresentado na seção 3.2, modificado para realizar a pesquisa na estrutura de dicionários proposta, e estendido para manipular transferências complexas (ver seção 5.1(b)) onde a propriedade de isomorfismo entre as representações não é mantida.

Finalmente, um módulo de *Verificação* é responsável pela ativação do processo de unificação dos traços semânticos, bem como por realizar uma verificação de consistência estrutural da árvore-alvo, descartando as árvores incompletas ou semanticamente inválidas.

A seguir é descrita em detalhe a experimentação realizada a partir desta proposta.

5.3 Experimentação realizada

Nesta seção é apresentada a implementação de um protótipo desenvolvido de acordo com o modelo proposto na figura 5.1 e levando em conta os conceitos de sincronismo introduzidos no capítulo 4.

Os dicionários são descritos utilizando-se a notação *Extended Backus Naur Form* (EBNF), enquanto que o algoritmo geral do módulo é descrito utilizando *português estruturado*.

A arquitetura funcional do módulo de transferência implementado é apresentada na figura 5.5.

O módulo de *carga* realiza a leitura dos arquivos texto que contêm a descrição EBNF dos dicionários e gramáticas utilizadas, realiza uma validação sobre os campos-chave, e gera um conjunto de estruturas compartilhadas correspondendo ao dicionários estrutural, ao dicionário bilíngüe e às listas de árvores elementares da língua fonte. A partir do texto de entrada é gerada a lista de lexicalização, composta pelas árvores elementares ancoradas pelos itens léxicos da entrada.

O *analisador sintático* utilizado é o implementado em (Kipper 94) e estendido com funções de lexicalização, o que é um requisito básico para a utilização das STAGs. Um procedimento independente de *transformação* é responsável por realizar as operações na árvore-alvo (controlado pelo analisador sintático), tendo-se, desta forma, um encapsulamento das funções de tradução.

Finalmente, um processo de *verificação* é responsável pela análise estrutural da árvore-alvo derivada, bem como a ativação da rotina de unificação. O processo é responsável por descartar as soluções inconsistentes.

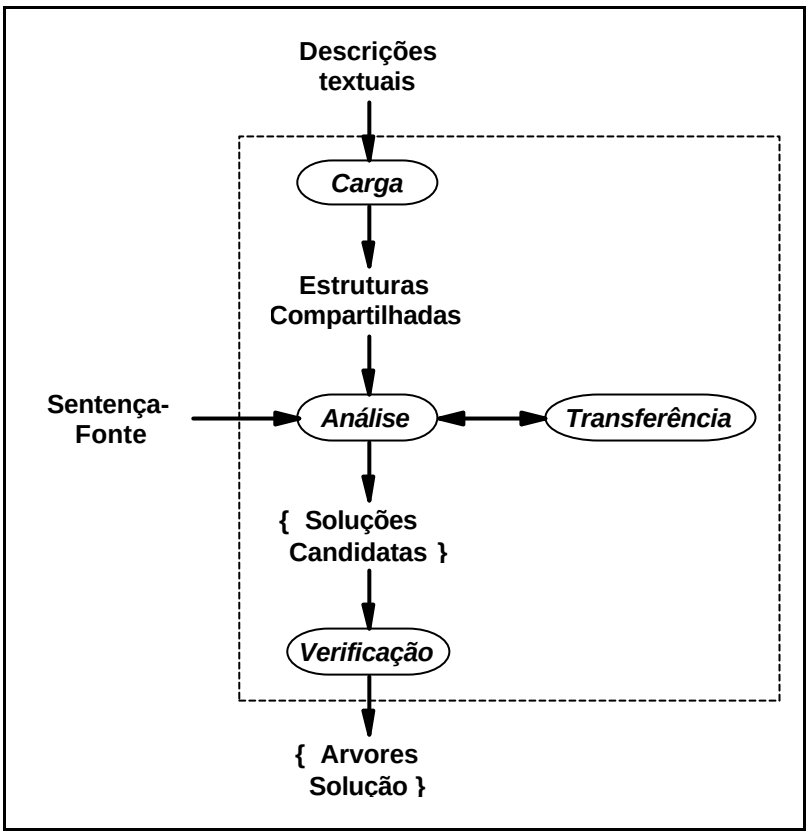


Figura 5.5 - Arquitetura do protótipo implementado

5.3.1 Dicionário Estrutural

Neste dicionário são explicitadas as ligações entre os nodos das LTAGs fonte e alvo. Isto permite modelar as divergências correspondentes à ordem dos constituintes, conforme seção 5.1(a), em dois níveis distintos: transferência estrutural simples e transferência restrita por condições semânticas, que devem ser satisfeitas para habilitar a operação de transferência. Exemplos destes casos podem ser observados na figura 5.6.

No exemplo 5.6(a), nenhuma restrição é imposta à utilização da estrutura, bastando, para tanto, que a árvore seja "ancorada" por um item léxico da classe ADJ (adjetivo). No exemplo 5.6(b), contudo, a restrição [*+NPROP*] indica que a operação só será validada caso o nome que ancora a estrutura possua o traço semântico *nome próprio*.

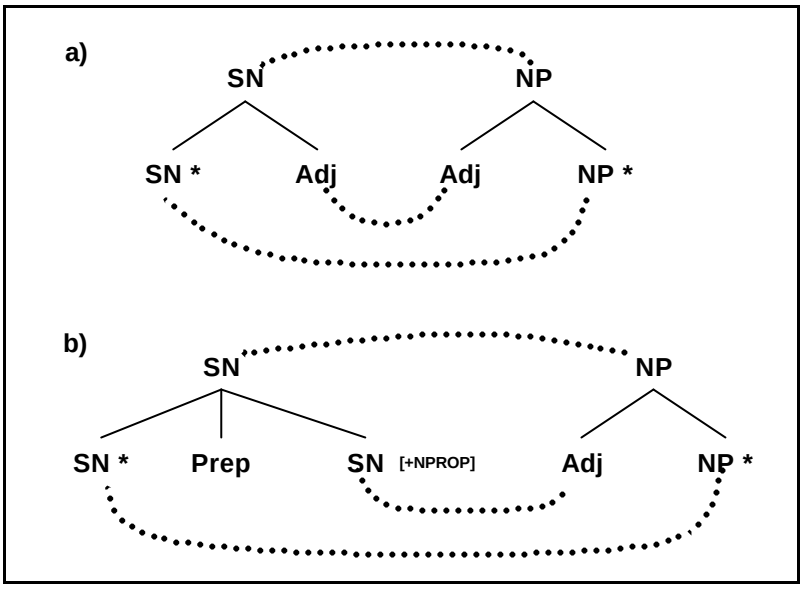


Figura 5.6 - Exemplos de entrada no dicionário estrutural

A gramática que descreve o formato deste dicionário é apresentada na figura 5.7. As regras são descritas por meio de uma correspondência posicional dos nodos das árvores, e utilizam a posição relativa do nodo dentro da estrutura para determinar as ligações entre as estruturas.

```
Dic_Transfer    ::=    "DicEstrutural"
                       { EstrutID  DescrEstr
                       }+
DescrEstr       ::=    "(" ArvFonteID "," ArvAlvoID "):""(" Corr_Estrut ")"
Corr_Estrut     ::=    [ "!" ] NodoFonte ":" NodoAlvo Parte_Semant
Parte_Semant    ::=    { [ "[" Traços_unific "]" ] [ "{" restrições "}" ] }+
Traços_unif     ::=    /* conforme definido na figura 5.2 */
Restrições      ::=    /* conforme definido na figura 5.2 */

Obs.: "!" →    indica negação da regra semântica utilizada,
               ou quebra de ligação, conforme o caso.
   NodoFonte, NodoAlvo → posição relativa do nodo na estrutura= $0, $1, $2, ...
```

Figura 5.7 - EBNF para descrição do dicionário estrutural

Nos casos de divergências estruturais observadas no corpus, o formalismo STAGs permitiu modelar corretamente as transformações necessárias na árvore derivada. Estes casos envolvem basicamente: deslocamento de adjetivos; deslocamento do advérbio de tempo; tratamento de adjuntos nominais; e diferenças na forma de estruturação das gramáticas LTAGs utilizadas.

5.3.2 Dicionário Bilíngüe

O *dicionário bilíngüe* é uma estrutura bastante complexa pois, além do dicionário de tradução propriamente dito, mantém um conjunto de informações necessárias a modelar as divergências léxico-semânticas, conforme a figura 5.8. As informações manipuladas neste dicionário são:

- item léxico na LF e seu correspondente na LA, incluindo expressões potenciais na LF (que determinam uma ambigüidade do processo de analise sintática);
- subcategorização (*Subcat*) do item léxico, na forma de uma lista de esturutras para as quais o item é elemento âncora. Um conjunto de restrições (*Restr*) pode ser definido de forma a habilitar a regra de subcategorização;
- subestruturas sintáticas, no caso de expressões da LF ou vazio-léxico da LA. Nestes casos, um fragmento de árvore sintática (*DescrEstr*) está presente no dicionário bilíngüe, que será utilizado durante o processo de derivação;
- e a regra de herança dos traços dentro das subestruturas que foram redefinidas (*Restr*).

```
Dic_Bilingue    ::=    "DicBilingue"
                         { BilingID palavraLF ":" palavraLA
                               [ "=" Subcat [ Restr ]
                               DescrEstr
                               ]
                         } +

Subcat          ::=    EstrutID [ Restr ] ";" Subcat | ε
Restr           ::=    NodoID ":" Restrições | ε
DescrEstr       ::=    /* conforme definido na figura 5.7 */
Restrições      ::=    /* conforme definido na figura 5.7 */
```

Figura 5.8 - EBNF para descrição do dicionário bilíngüe

Durante o processo de análise e tradução, dois tipos de traços semânticos são manipulados: atributos estruturais e atributos morfo-semânticos. Os atributos estruturais são restrições impostas às operações, e não necessitam ser transferidos. Os atributos morfo-semânticos são herdados pelos itens realizados dos pares de árvores lexicalizadas. Durante o desenvolvimento deste trabalho, não foram focalizadas as divergências existentes no processo de herança dos traços morfológicos, embora este estudo seja importante no processo de tradução. O dicionário bilíngüe define a herança correta para os casos onde haja conflito; nos demais casos, todos os atributos pertencentes ao item léxico da LF são transferidos para o nodo correspondente na LA.

Os atributos transferidos não são especificações absolutas para os itens léxicos na árvore alvo. O processo de geração deverá utilizá-los como restrições, ou guias, para obter a forma que melhor satisfaça os atributos. Este processo deve ignorar atributos não úteis, como, por exemplo, gênero de substantivos e adjetivos na língua inglesa, ou que sejam incompatíveis com as restrições da LA (*as bagagens*, plural em Português ⇒ *lugagge*, singular em Inglês).

A figura 5.9 apresenta dois exemplos de entrada no dicionário bilíngüe: no exemplo (a) o nome *défict* subcategoriza um complemento nominal, caso a preposição (*Prep*) pertença ao conjunto {*de, em*}. Se a restrição não for satisfeita, a operação de adjunção sobre o nodo é desfeita. Além desta restrição, o traço [#lex] indica que o item léxico (preposição) é herdado durante o processo de transferência pela subcategorização. No exemplo (b), é modelado um caso de vazio léxico da LA, contendo a subárvore que será utilizada na árvore-alvo derivada. A ligação entre os nodos indica a herança dos traços para esta subestrutura.

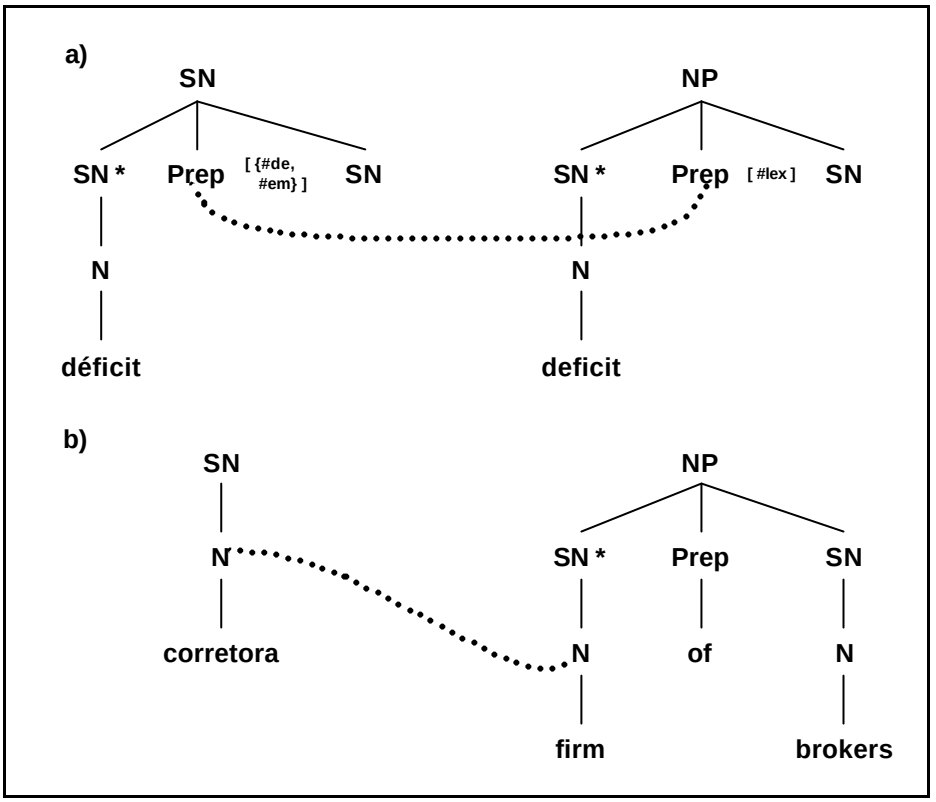


Figura 5.9 - Exemplos de entrada no dicionário bilíngüe

5.3.3 Processo de Análise e Transferência

O protótipo está implementado na linguagem de programação C (ANSI) em plataforma UNIX. O analisador sintático controla o processo e é responsável pela ativação das rotinas de transferência.

A figura 5.10 apresenta o algoritmo básico do processo de análise. A técnica utilizada implementa um algoritmo *top-down* recursivo (analisador sintático descendente recursivo), que determina algumas características importantes do módulo de transferência:

1. No momento de iniciar a derivação, as informações léxicas ainda não estão todas disponíveis. Com isto, o módulo de *carga* deve gerar uma lista de lexicalização, que é percorrida exaustivamente até que ocorra a instanciação léxica. Apenas neste momento os atributos estão disponíveis e as restrições semânticas são verificadas, descartando-se as árvores inconsistentes;

2. O algoritmo básico é bastante simples: a complexidade reside no modo de passagem dos parâmetros e controle de *backtracking*, responsáveis por manter a consistência de cada instância ativa do procedimento de análise (apenas as estruturas de dados relativas aos dicionários são globais). Restrições quanto à área de alocação da pilha do sistema (*stack*) impedem que o sistema seja executado em MS-DOS, embora alguns testes tenham sido realizados com sucesso para migração para plataforma WINDOWS.

```
Nome.......: parser_tag
Função.....: realiza análise sintática e ativação das rotinas de transferência
             para uma árvore inicial
Parâmetros:
        <Pont_ini, Pont_aux> → ponteiros para LTAGs elementares
        Texto → ponteiro para a lista de tokens
        Ini_sol → ponteiro para árvore-solução ativa
        Ini_alv → ponteiro para a árvore-alvo ativa
        Passo → controle do backtracking
        Memodic → ponteiro para o dicionário sintático
Retorno....: nenhum

Algoritmo..:
   Busca símbolo para derivar
   Se símbolo é nodo terminal descrito na árvore-solução
        Armazena informações de backtracking
        Consome símb olo
        Ativa parser recursivamente
        Desfaz consumo
        Restaura informações
   Senao_Se símbolo possui a mesma categoria do nodo na árvore solução
        Pesquisa token no dicionário bilíngüe
            Para cada entrada (token e possíveis expressões)
                Realiza a herança dos atributos
                Armazena informações de backtracking
                Consome símbolo
                Ativa parser recursivamente
                Desfaz consumo
                Restaura informações
            FimPara
   Senão_Se for possível realizar substituição ou adjunção no nodo
        Enquanto houver árvores na lista de lexicalização
        Armazena informações de backtracking
        Busca estrutura-alvo no dicionário estrutural
            Para cada entrada (estruturas sintáticas alternativas)
                Realiza operação (substituição ou adjunção)
                Ativa parser recursivamente
                Desfaz operação
                Restaura informações
```

```
            FimPara
    Senão_Se  não existem mais nodos livres na árvore-solucao
        Se processo de verificação estrutural da árvore-alvo está ok
            Lista estrutura derivada como solução (árvores fonte e alvo)
        FimSe
    FimSe
```

Figura 5.10 - Algoritmo geral do módulo de análise e transferência

Dois casos de ambigüidade na transfêrencia são manipulados durante o processo de análise sintática (ambigüidades na análise sintática da sentença de entrada são tratadas implicitamente pelo processo recursivo presente no algoritmo):

- no caso de ocorrer mais de uma árvore-alvo candidata para uma mesma árvore elementar, seria necessário manipular uma estrutura no formato de um reticulado[15] para manter todas as árvores e suas dependências. Aproveitando-se a recursão existente no algoritmo, optou-se por replicar as entradas na lista de árvores elementares (criando-se uma nova entrada para cada estrutura alternativa). Desta forma, várias árvores poderão ser geradas para uma mesma árvore solução da LF, cabendo ao processo de verificação descartar as alternativas mal-formadas ou, caso isto não seja possível, caberá ao módulo de geração escolher a mais adequada;
- no caso da ambigüidade léxica conceitual, conforme definido na seção 2.4.1, da mesma forma que no caso anterior, são replicadas as entradas na lista de árvores elementares, uma para cada possível acepção do termo. Neste caso, não será necessária a verificação, cabendo ao módulo de geração a escolha da acepção mais adequada.

5.3.4 Processo de Verificação

O *processo de verificação* realiza a consistência da árvore solução. Dois passos básicos são executados:

1. Percorre a estrutura verificando se todos os nodos não-terminais foram derivados;

15 O reticulado define uma estrutura na forma de "planos de árvores compartilhadas", onde em cada caminho está representada uma solução possível.

2. Realiza a propagação dos atributos e a unificação, de acordo com as regras da LTAG-alvo.

Caso alguma destas verificações falhe, a árvore é descartada, e o controle retorna ao analisador sintático para derivar a próxima alternativa. Se, ao final da análise sintática, nenhuma árvore for derivada, isto representa uma inconsistência dos dicionários utilizados.

5.3.5 Exemplo de Utilização

A seguir são comentados os principais passos executados pelo módulo de tradução durante o reconhecimento da sentença:

Cavallo anuncia déficit na balança comercial.

Para facilitar o entendimento, nos pares de árvores <fonte, alvo>, geradas pelo processo de análise, foram omitidas as ligações entre os nodos, mantendo-se apenas a ligação que será utilizada no passo seguinte. Da mesma forma, não estão representados os traços semânticos da árvore de derivação e herdados pela árvore alvo.

Em um primeiro passo, ocorre a carga do sistema e é gerada uma lista de lexicalização, contendo todas as estruturas elementares ancoradas pelos itens léxicos da entrada. A rotina de análise sintática é ativada para cada árvore inicial da lista ancorada pelo(s) verbo(s) da sentença, ou seja, é assumido que o verbo é o elemento que governa a sentença.

No exemplo, o par de árvores da figura 5.11 é recuperado do dicionário estrutural, de acordo com a subcategorização do verbo *anunciar* (VTD). A seguir são realizadas as operações (todas ocorrem de forma síncrona nas árvores fonte e alvo):

- substituição utilizando a regra SN [N(NProp)], que realiza a inserção léxica do nome próprio *Cavallo*;
- substituição do pré-terminal V;
- substituição do sintagma nominal, utilizando a regra SN [N()].

O resultado deste primeiro passo está apresentado na figura 5.12 (as formas derivadas dos itens léxicos foram inseridas apenas para facilitar a leitura, pois os itens léxicos inseridos estão na forma normalizada, o verbo *to announce*, por exemplo, é gerado em sua forma normalizada com os traços: [*pessoa=3; num=singular; tempo=presente, modo=indicativo; tempo=presente*].

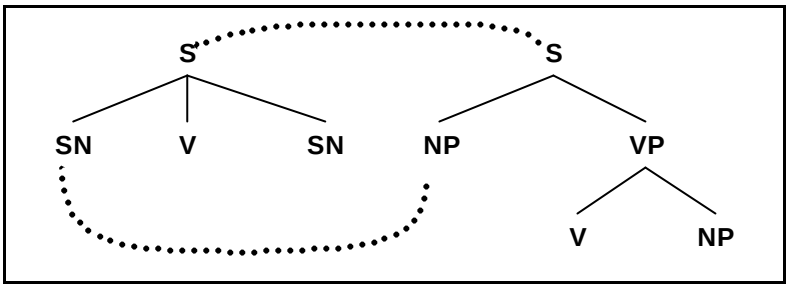


Figura 5.11 - Árvore inicial subcategorizada pelo verbo *anunciar*

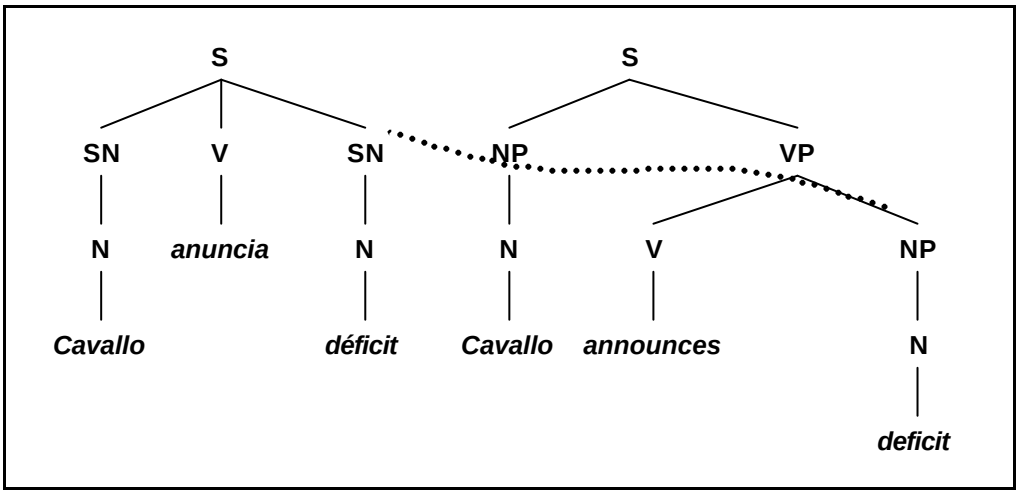


Figura 5.12 - Árvores após o primeiro passo da transformação

A partir das estruturas apresentadas na figura 5.12, é selecionado o nodo sintático que derivou o nome *déficit*. Segundo o dicionário bilíngüe, este nodo pode ter a seguinte subcategorização:

- Déficit N, __ Prep NP [Prep = { #de, #em }]

Esta regra habilita a adjunção de um complemento nominal, utilizando o par de estruturas elementares apresentado na figura 5.13, com a restrição de que a preposição seja instanciada para *de* ou *em* (caso contrário, o par de árvores derivadas será descartado durante o processo de verificação).

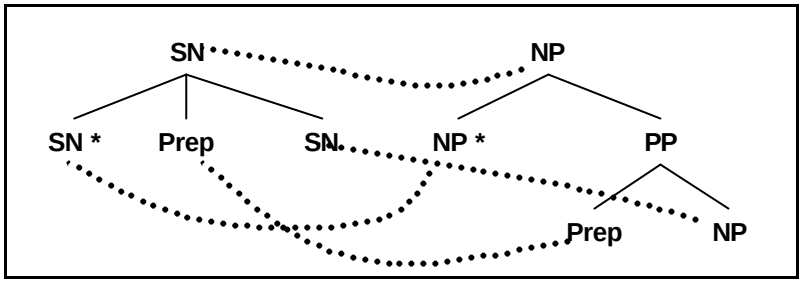


Figura 5.13 - Dicionário estrutural para adjunção de complemento nominal

Realizando a adjunção da estrutura apresentada na figura 5.13, sobre o par de árvores da figura 5.12, e em seguida realizando as operações de substituição dos itens léxicos, obtém-se o par de árvores da figura 5.14.

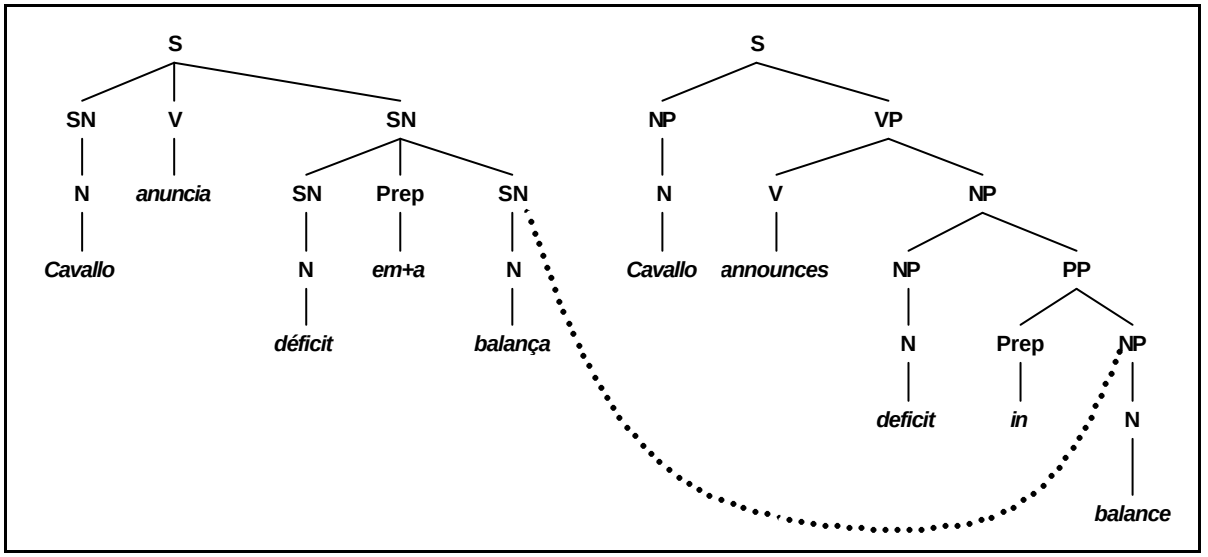


Figura 5.14 - Árvore de derivação após adjunção do complemento nominal

Finalmente, em um último passo, ocorre a manipulação do adjetivo *comercial*. A entrada no dicionário bilíngüe deste item aponta para a estrutura de subcategorização apresentada na figura 5.15.

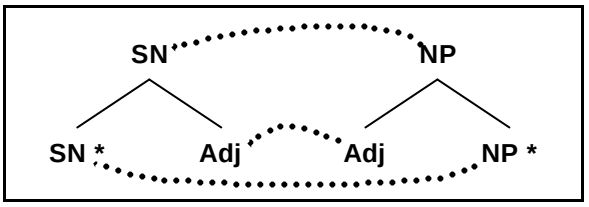


Figura 5.15 - Estrutura sintática para a divergência de ordem do adjetivo

Ao realizar a adjunção do par de árvores apresentado na figura 5.15 sobre a árvore da figura 5.14, e ao realizar a instanciação léxica, obtém-se o par de árvores apresentadas na figura 5.16. Como nenhuma outra operação é possível sobre esta árvore, a análise sintática é concluída e o processo de verificação é ativado sobre a árvore-alvo (caso seja possível unificar os traços, a árvore é listada como solução).

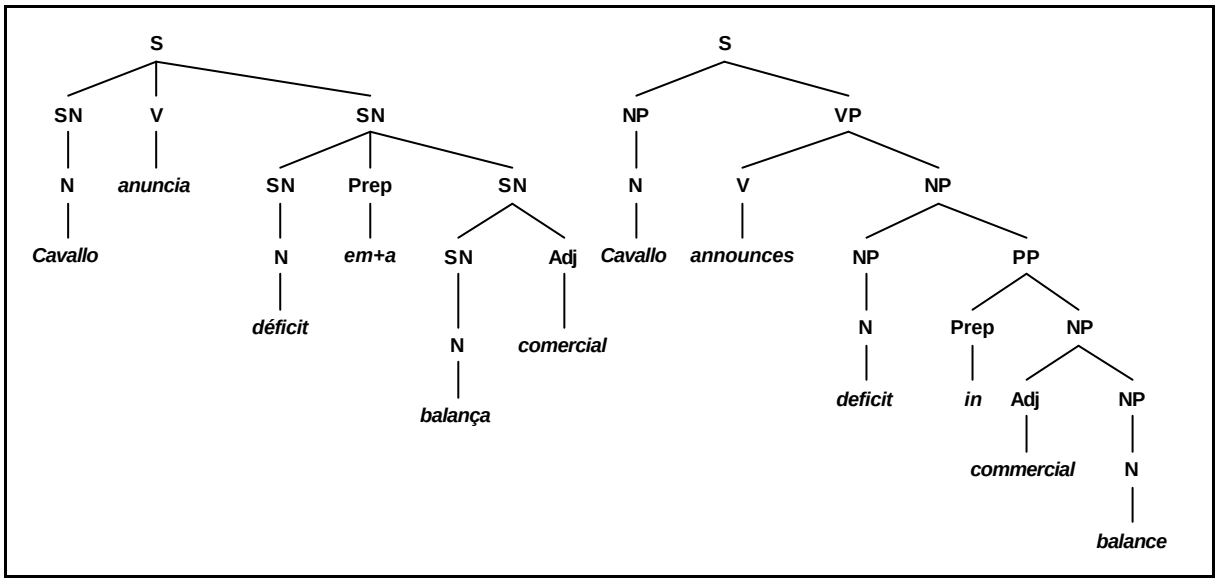


Figura 5.16 - Árvore solução <fonte, alvo> para *Cavallo anuncia déficit na balança comercial*

6. CONCLUSÃO

6.1 Utilização de STAGs para a construção de um módulo de transferência

O trabalho procura realizar um levantamento de divergências estruturais que ocorrem durante um processo de tradução entre as línguas portuguesa e inglesa, e investiga a aplicabilidade do formalismo STAGs para resolução dos problemas encontrados.

Devido à propriedade de domínio de localidade estendido, LTAGs permitem definir correspondências regulares entre estruturas complexas sem a necessidade de representações intermediárias. O mapeamento entre árvores de derivação de uma LF para uma língua-alvo utilizando o formalismo STAGs permite realizar estas correspondências de forma bastante direta. Isto permite combinar unidades com estruturas internas bastante diferentes, como foi visto nos capítulos 4 e 5.

Além disto, o fato de utilizar uma gramática lexicalizada permite capturar aspectos específicos de cada língua.

Vários níveis de transferência foram utilizados: transferência de árvores não lexicalizadas; transferência léxica e transferência de características (traços semânticos).

A transferência de árvores não lexicalizadas determina a correspondência entre as árvores elementares das duas línguas, e as ligações entre os nodos destas árvores. Restrições estruturais e semânticas podem ser definidas a este nível, modelando o conjunto de divergências sintáticas.

O segundo nível, de transferência léxica, relaciona as entradas nos dicionários das duas línguas, e é onde são modeladas tanto as divergências léxicas quanto as divergências léxico-semânticas.

Finalmente, a transferência de atributos é realizada nos dois níveis anteriores, explicitando as correspondências para cada par de itens âncora relevantes.

6.2 Aplicabilidade do formalismo STAGs

A princípio, podem ser identificados, na tradução, dois objetivos essenciais: a aquisição e a disseminação de informação. O primeiro consiste basicamente na reunião de conhecimento, enquanto o segundo relaciona-se à exportação de tecnologia. Um exemplo de disseminação de informação de caráter comercial inclui literatura de propaganda e venda, instrução de operação de produtos e dados sobre procedimentos e serviços, além de literatura técnica e acadêmica. Muitas destas aplicações também possuem a finalidade de aquisição de conhecimento, por exemplo, correspondência diária, notícias comerciais ou econômicas (Araújo 93).

Nesse sentido, acreditamos, através dos resultados obtidos, que STAGs fornecem um modelo teórico interessante para a construção do módulo de transferêncua para sistemas de tradução automática para aquisição de conhecimento.

6.3 Validação

O experimento realizado demonstrou que a grande maioria dos casos observados no corpus foram tratados de forma correta pelo formalismo STAGs. Desta forma, mostrou-se que é possível a construção do módulo de transferência de um sistema de tradução entre as línguas portuguesa e inglesa utilizando como modelo teórico este formalismo. A figura 6.1 apresenta uma relação entre as divergências encontradas na literatura e as classes de divergências trabalhadas no protótipo.

Classes de divergências encontradas na bibliografia	Número de ocorrências encontradas no corpus
Léxica conceitual & vazio léxico	32
Ordem dos constituintes	87
Associação de preposições	76
Omissão do sujeito	-
Conflacional	10
Estrutural	-
Promocional	-
Léxica (expressões)	9

Figura 6.1 - Análise comparativa das divergências

O trabalho desenvolvido por Zorzo (Zorzo 93) apresenta uma experiência de utilização de Gramáticas Transformacionais com atributos para a tradução de linguagens artificiais e sugere sua aplicabilidade para tradução da língua natural. Zorzo identifica quatro tipos de transformações que podem ocorrer a partir de uma árvore de derivação: adição, remoção, permutação e substituição de símbolos.

A figura 6.2 demonstra que as transformações observadas por Zorzo são implementadas por STAGs. O problema vislumbrado é que as operações apenas sintáticas são aplicáveis a somente 21,02% dos casos observados no corpus, ou seja, as regras não são aplicáveis isoladamente, sem as informações adquiridas dos itens léxicos utilizados.

Desta forma, pode-se dizer que o formalismo é adequado por mapear as transformações estruturais existentes, mas a generalização das classes de divergências observadas exigiria a construção de um corpus que listasse exaustivamente os casos de uma determinada classe, o que provavelmente não seria produtivo, tampouco viável. Assim, um novo corpus não iria validar, ou mesmo invalidar, o modelo apresentado, mas apenas permitiria dar maior confiabilidade aos casos modelados.

Alguns estudos da área da Lingüística Computacional afirmam que a língua não permite a engenharia reversa, ou seja, um sistema de tradução bidirecional não é factível, pois não é possível generalizar as regras utilizadas para os casos observados. A generalização só é possível pela inclusão exaustiva de sentenças no corpus. Isto só será possível a partir da disponibilidade de programas que realizem a marcação e o alinhamento de corpus bilíngües automaticamente.

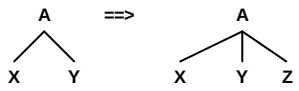
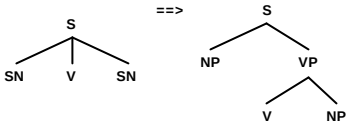
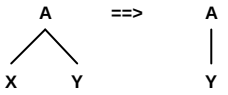
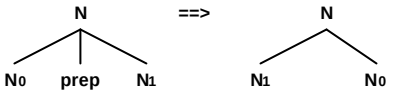
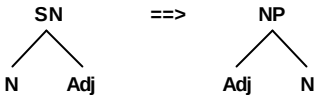
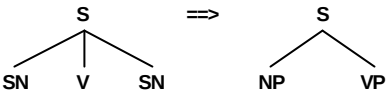
Transformação apresentada em (Zorzo 93)	Caso modelado no experimento realizado
1. Adicão 	
2. Remoção 	
3. Permutação 	
4. Substituição 	

Figura 6.2 - Modelagem das transformações apresentadas em (Zorzo 93)

6.4 Sugestão de trabalhos futuros

O trabalho desenvolvido é apenas o primeiro passo em direção à construção de um sistema de tradução automática de qualidade. A continuidade do trabalho deverá prever:

- implementação de um módulo semântico mais efetivo, já que o modelo de traços morfo-sintáticos utilizado não é suficiente para tratar casos de polissemia, encontrados com freqüência no processamento da língua natural;
- incorporação de um módulo de aquisição de conhecimento lingüístico que permita automatizar a definição de novas classes de divergências. A implementação deste módulo é complexa e exige, no mínimo, a

> disponibilidade de extensos corpora bilíngües e marcados, o que não existe[16] atualmente para o par de línguas português-inglês.

Finalmente, é necessária a construção de um módulo de geração que receba as árvores geradas pelo módulo de transferência e realize a geração do texto na língua-alvo.

O enfoque utilizado neste trabalho é o de uma visão basicamente computacional da TA, onde a tradução é realizada a partir de regras estruturais sobre uma representação gramatical das línguas envolvidas. Segundo Leffa e Filgueiras, porém, a tradução, sem a representação de mundo para se apoiar, é uma tarefa difícil, provavelmente impossível (Leffa 95, Filgueiras 94).

[16] Esta afirmação é baseada no levantamento de (Dias 96), realizado no âmbito de seu plano de doutorado, que prevê a construção de um sistema de alinhamento para corpora bilíngues português-inglês.

ANEXO - CORPUS UTILIZADO

A seguir é listado o corpus trabalhado, comentado com as classes de divergências observadas em cada caso[17], conforme figura A.1. A figura A.2 apresenta uma estatística do número de ocorrência de cada uma das classes.

Cod	***Classe de divergência***	***Dicionário utilizado***
OCons	Ordem dos constituintes	
	adjetivo - Adj	Estrutural
	locução adjetiva - Ladj	Bilíngüe
	adjunto adverbial - AAdv	Bilíngüe
DLex	Divergência léxica	
	Conceitual - Conc	Bilíngüe
	vazio léxico língua-fonte - VL-F	Bilíngüe
	vazio léxico língua-alvo - VL-A	Bilíngüe
	expressão - Expr	Bilíngüe
SPrep	Seleção de preposições	Bilíngüe
Confl	Conflacional	Bilíngüe

Figura A.1 - Classes de divergências observadas

Cod	Ocorrências observadas	
	Num	%
OCons		
Adj	45	21,02
Ladj	35	16,35
Aadv	7	3,27
Total	87	40,65
Dlex		
Conc	4	1,86
VL-F	3	1,40
VL-A	25	11,68
Expr	9	4,25
Total	41	19,15
SPrep	76	41,99
Confl	10	3,89

Figura A.2 - Classes de divergências: estatística

[17] Nas sentenças apresentadas, conjuntos de palavras delimitadas pelos símbolos '<' e '>' indicam que o conjunto foi tratado como uma unidade sintática. Isto ocorre com os nomes próprios (por exemplo: <Celso Furtado> e numerais (por exemplo: <US$ 4 bilhões>).

<001> <celso furtado> apoia aumento das aliquotas de importacao
<celso furtado> supports increase of aliquotas for imports
Diverg.: SPrep(2)

<002> inflacao de marco na <russia> cai para <8,9%>
march inflation falls to <8.9%> in <Russia>
Diverg.: SPrep(2), OCons(AAdv, LAdj),

<003> <japao> anuncia plano de abertura economica
<japan> anounces plan to open economy
Diverg.: SPrep

<004> <fhc> anuncia obras da hidrovia araguaia-tocantins
<fhc> anounces commencement of araguaia-tocantins waterway construction
Diverg.: OCons(LAdj), DLex(Expr)

<005> banco central liquida <af> administradora de consorcios
central bank liquidates <af> administration of 'purchase consortiums'
Diverg.: OCons(Adj), SPrep, DLex(VL-A)

<006> bolsa de <nova york> fecha em leve queda
<new york> stock exchange closes in slight fall
Diverg.: OCons(LAdj), SPrep, DLex(VL-A)

<007> peso <mexicano> fecha em <6,26> por dolar
<mexican> peso closes at <6.26> to dollar
Diverg.: OCons(Adj), SPrep(2)

<008> bolsa de <nova york> fecha em baixa de <11 pontos>
<new york> stock exchange closes at <11 points> below par
Diverg.: DLex(VL-A, Expr), OCons(LAdj), SPrep

<009> bolsa de <nova york> opera em baixa de <12 pontos>
<new york> stock exchange operates at <12 points> below par
Diverg.: DLex(VL-A, Expr), OCons(LAdj), SPrep

<010> <lampreia> condena embargo economico a <cuba>
<lampreia> condemns economic embargo to <cuba>
Diverg.: SPrep, OCons(Adj)

<011> fluxo cambial tem saldo negativo de <us$ 3,2 bi> este mes
this month's cash flow shows negative balance of <US$ 3.2 bilion>
Diverg.: OCons(Adj, Adj, AAdv), SPrep, DLex(Conc)
Forma gerada: cash flow shows negative balance of <US$ 3.2 bilion> this month

<012> <garoto> investe <us$ 2 milhoes> em propaganda de pascoa
<garoto> invests <US$ 2 million> in Easter advertising
Diverg.: SPrep, OCons(LAdj)

<013> banco central decreta liquidacao do consorcio <transamericana>
central bank decrees liquidation of <transamericana> consortium
Diverg.: OCons(Adj, Adj), SPrep

<014> banco <abn-ambro> tem lucro de <113 milhoes> em <1994>
<abn-ambro> bank announces profits of <113 million> in <1994>
Diverg.: OCons(Adj), DLex(Conc), SPrep(2)

<015> banco central faz hoje leilao de <r$ 1 bi> em nbc cambial
central bank to auction <R$ 1 billion> bonds today
Diverg.: OCons(Adj, AAdv), Confl(VL-A), DLex(VL-A)

<016> congresso <argentino> aprova pacote fiscal
<argentine> congress approves fiscal package
Diverg.: OCons(Adj, Adj)

<017> <iene> forte derruba bolsa de toquio em <311 pontos>
strong <yen> lowers tokyo stock exchange by <311 points>
Diverg.: OCons(Adj, LAdj), DLex(VL-A), SPrep

<018> dolar comercial abre em alta a <r$ 0,89>
commercial dollar opens at increase of <R$ 0,89>
Diverg.: OCons(Adj), SPrep(2)

<019> bolsa de <toquio> fecha com baixa de <105 pontos>
<tokyo> stock exchange closes with drop of <105 points>
Diverg.: DLex(VL-A), OCons(LAdj), SPrep(2)

<020> banco central vendeu nbc ontem a <16,04%>
central bank sold bonds yesterday at <16.04%>
Diverg.: OCons(Adj), SPrep

<022> economia do <japao> cresce <0,6%> em <1994>
<japanese> economy grew <0.6%> in <1994>
Diverg.: OCons(LAdj), SPrep

<023> <cavallo> anuncia reducao do defict na balanca comercial
<cavallo> announces reduction of deficit in commercial balance
Diverg.: SPrep(2), OCons(Adj)

<024> funcionarios da receita pedem demissao na <argentina>
inland revenue employees request resignation in <argentina>
Diverg.: DLex(VL-A), OCons(LAdj), SPrep

<025> bolsa de <nova york> opera em queda de <6,22 pontos>
<new york> stock exchange operates at fall of <6.22 points>
Diverg.: OCons(LAdj), SPrep(2), DLex(VL-A)

<026> <telebras> tem lucro de <r$ 943 milhoes> em <1994>
<telebras> makes profit of <R$ 943,00 million> in <1994>
Diverg.: DLex(Conc), SPrep(2)

<027> <jose serra> chega a associacao comercial no rio
<jose serra> arrives at commercial association in rio
Diverg.: SPrep(2), OCons(Adj)

<028> indice merval opera em baixa de <4,6%>
merval operates at <4.6%> below par
Diverg.: SPrep(2), DLex(Expr)

<029> bolsa do <mexico> opera em alta de <0,34%>
<mexican> stock exchange operates at increase of <0.34%>
Diverg.: OCons(LAdj), DLex(VL-A), SPrep(2)

<030> <cavalo> preve inflacao de <4,5%> este ano
<cavallo> predicts inflation at <4.5%> this year
Diverg.: SPrep

<031> <peso> fecha em alta a <6,98> por dolar
<peso> closes at increase of <6.98> to dollar
Diverg.: SPrep

<032> <bovespa> e <bvrj> discutem uniformizacao de procedimentos

<bovespa> and <bvrj> discuss regularization of procedures
Diverg.: SPrep

<033> reuniao na camara discute votacao de emendas
voting of amendments discussed in council chamber
Diverg.: Sprep, DLex(VL-A)
Forma gerada: council chamber discuss voting of amendments

<034> grupo <santana> fatura <us$ 50 milhoes> este ano
<santana> group make <US$ 50 million> this year
Diverg.: OCons(LAdj)

<035> banco <icatu> compra corretora em <sao paulo>
<icatu> bank buys firm of brokers in <sao paulo>
Diverg.: OCons(LAdj), DLex(VL-A), SPrep

<036> juros abrem com oscilacao moderada
interests open with moderate fluctuation
Diverg.: SPrep, OCons(Adj)

<037> <ceval> anuncia compra de <agroliane sa>
<ceval> announces purchases of <agroliane sa>
Diverg.: SPrep

<038> secretario da agricultura reage contra a votacao da <tr>
agrucultural secretary reacts against <tr> voting
Diverg.: OCons(LAdj, LAdj), SPrep

<039> <kissinger> participa de debate sobre economia no <cone sul>
<kissinger> participates in debate over economy of <cone sul>
Diverg.: SPrep(2)

<040> <argentina> aprova protecao a depositos bancarios
<argentina> approves protection for bank deposits
Diverg.: SPrep, OCons(Adj)

<041> <chile> acumula superavit comercial de <us$ 692 milhoes> em <1995>
<chile> accumulates commercial surplus of <US$ 692 million> in <1995>
Diverg.: OCons(Adj), SPrep(2)

<042> receita prorroga prazo de entrega do imposto de renda
inland revenue extends dead line for submission of tax
Diverg.: SPrep, DLex(VL-F, Expr), OCons

<043> empresa <russa> importa acucar do <brasil>
<russian> company imports <brasilian> sugar
Diverg.: OCons(Adj, LAdj)

<044> presidente da camara repudia interferencia no parlamento
council president repudiates interference in parliament
Diverg.: OCons(LAdj), SPrep

<045> empreiteiras reclamam falta de recursos para as rodovias
contract construction companies complain of lack of resources for roads
Diverg.: DLex(VL-A), SPrep(2)

<046> <malan> participa da reuniao de comercio exterior
<malan> participates in export trade meeting
Diverg.: OCons(Adj, LAdj), SPrep

<047> grupo <gerdau> compra mais uma siderurgica <canadense>

<gerdau> group buys another <canadian> ironworks
Diverg.: OCons(Adj, LAdj), DLex(Expr)

<048> <bndes> aprova financiamento para producao de leite
<bndes> approves financial aid for milk production
Diverg.: DLex(VL-A), SPrep, OCons(LAdj)

<049> <acm> critica poder judiciario na tribuna do senado
<acm> criticizes judicial power within senate tribune
Diverg.: OCons(Adj, LAdj), SPrep

<050> metalurgicos do <rio> fazem assembleia amanha
<rio> metal workers to meet tomorrow
Diverg.: DLex(VL-A), Confl, OCons(LAdj)

<051> parlamentares de <sao paulo> discutem o caso <banespa>
<sao paulo> members of parliament discuss <banespa> case
Diverg.: OCons(LAdj, LAdj), DLex(VL-A)

<052> senado aprova convocacao de <arida>
senate approves <arida> convocation
Diverg.: OCons(LAdj), Sprep

<053> titulo da divida <brasileira> sobe com pacote <argentino>
<brazilian> debt bonds rises with <argentine> package
Diverg.: OCons(Adj, LAdj, Adj), SPrep

<054> <serra> defende regulamentacao do mercado financeiro
<serra> defends regulation of finacial market
Diverg.: SPrep, OCons(Adj)

<055> <serra> diz que governo vai cobrar resultados de estatais
<serra> says "government to demand presentation of state company's results"
Diverg.: Confl, DLex(VL-A), OCons(Poss)
Forma gerada: <serra> says that government demands results of state companies

<056> presidente da camara acelera tramites da reforma
council president speeds up paths to reform
Diverg.: OCons(LAdj), SPrep

<057> <arida> adia discussao com deputados sobre <banespa>
<arida> postpones talks with deputies over <banespa>
Diverg.: SPrep(2)

<058> <cavallo> descarta congelamento de depositos bancarios
<cavallo> dismisses freezing of bank deposits
Diverg.: SPrep, OCons(Adj)

<059> <sarney> promete a <malan> ajudar na aprovacao do ajuste fiscal
<sarney> promisses <malan> help in approval of settlement
Diverg.: SPrep(2), DLex(VL-F)

<060> <fhc> fara discurso duro contra manifestantes
<fhc> will make tough speech against demonstrators
Diverg.: OCons(Adj), SPrep

<061> <fhc> faz critica indireta ao senador <dutra>
<fhc> indirectly criticizes senator <dutra>

Diverg.: Confl, OCons(Adj)

<062> banco central contesta posicao dos consorcios de montadoras
central bank contests position of engine fitter consortiums
Diverg.: OCons(Adj), SPrep, DLex(VL-A)

<063> vice-premier <russo> e destituido do comite de reformas
<russian> vice-premier dimissed from reform committee
Diverg.: OCons(Adj, LAdj), Confl, SPrep

<064> presidente mundial da <ford> chega ao <brasil> domingo
<ford> world president to arrive in <brazil> on sunday
Diverg.: OCons(Adj, LAdj), SPrep

<065> governo de <sao paulo> incentiva uso de porto paulista
<sao paulo> governor to give incentives for use of port from <sao paulo>
Diverg.: OCons(LAdj), DLex(VL-A), SPrep

<066> <febraban> desenvolve proposta sobre derivativos
<febraban> develop proposal over derivatives
Diverg.: SPrep

<067> <argentina> cria fundo para gerenciar emprestimos
<argentina> creates funds to manage loans
Diverg.: SPrep

<068> <alkimar moura> nega demissao de interventores do <banespa>
<alkimar moura> denies dismissal of government mediators to <banespa>
Diverg.: SPrep(2), DLex(VL-A)

<069> <chile> defende controles sobre capital externo
<chile> defends control of external capital
Diverg.: SPrep, OCons(Adj)

<070> <itamar> e aprovado para embaixada de <portugal>
<itamar> approved to <portuguese> embassador
Diverg.: Sprep
Forma gerada: <itamar> approved to embassy in portugual

<071> <fhc> e ministros discutem hoje a reforma tributaria
<fhc> and ministers to discuss tax payers reform today
Diverg.: OCons(Adj, AAdv), DLex(VL-A), Spep

<072> <serasa> divulga diagnostico sobre empresas na terca-feira
<serasa> to publicise analysis of companies on tuesday
Diverg.: SPrep(2)

<073> <mario covas> recepciona presidente da irlanda
<mario covas> receives irish president
Diverg.: OCons(LAdj)

<074> governo comeca a vender imoveis para aumentar a receita
government begins selling property to increase assets
Diverg.: SPrep(2)

<075> <eua> e <japao> retomam negociacao sobre veiculos
<usa> and <japan> return to the negotiation table over vehicles
Diverg.: DLex(Expr), SPrep(2)

<076> governo aumenta fiscalizacao em postos de gasolina
government steps up control in petrol stations

Diverg.: SPrep, OCons(LAdj)

<077> <peru> privatiza amanha o terceiro maior banco do pais
<peru> to privatise country's third largest bank tomorrow
Diverg.: OCons(AAdv, Poss)

<078> parlamentares discutem nova proposta para o salario minimo
members of parliament discuss new proposals over minimum salary
Diverg.: DLex(VL-A), SPrep, OCons(Adj)

<079> <fhc> quer investimento privado em petroleo e comunicacoes
<fhc> calls for private investiment in petrolium and comunications
Diverg.: OCons(Adj), SPrep

<080> bolsas <latinas> sustentam volume acima de <650 milhoes>
<latin america> stock exchange sustains volume above <650 millions>
Diverg.: OCons(Adj), DLex(VL-A), SPrep

<081> <malan> participa hoje de palestra em <new orleans>
<malan> to take part in lecture in <new orleans> today
Diverg.: SPrep, DLex(Expr), OCons(AAdv)

<082> <tradings> estimam deficit comercial de <US$ 0,5 bilhoes> <em 1996>
<tradings> estimate commercial deficit of <US$ 0.5 billion> <in 1996>
Diverg.: OCons(Adj), SPrep(2)

<083> <bndes> conclui estudo para criar titulo mobiliario
<bndes> concludes research to create property bonds
Diverg.: OCons(Adj), SPrep

<084> <lapreia> discute setor automobilistico em <sao paulo>
<lapreia> to discuss automobile sector in <sao paulo>
Diverg.: OCons(Adj), SPrep

<085> <serra> diz que leilao da <light> esta mantido para <dia 21>
<serra> announces auction of <light> to be held <21 st.>
Diverg.: SPrep(2)
Forma gerada: <serra> announces that auction of <light> is held to <21 st.>

<086> aeronautica confirma queda de aviao na <bahia>
air force confirms aeroplane crash in <bahia>
Diverg.: OCons, SPrep

<087> direcao do <pfl> quer manter alianca com <psdb> em <sao paulo>
<pfl> management wish to maintain alliance with <psdb> in <sao paulo>
Diverg.: OCons(LAdj), Sprep(2)

<088> malan almoca com empresarios britanicos na embaixada
malan lunches with british industrialists at embassy
Diverg.: SPrep(2), OCons(Adj)

<089> novo presidente da <light> devera ser um brasileiro
new president of <light> should be brasilian
Diverg.: SPrep(2)

<090> governo pede financiamento para construcao de rodovia
government asks for finance to construction of highway
Diverg.: SPrep(2)

BIBLIOGRAFIA

(Aarts 91) AARTS, E. Uniform Recognition for Acyclic Context-sensitive Grammars is NP-Complete. In: Computing Science in the Netherlands, 8, 1991, Amsterdan. **Proccedings...** Amsterdan, 1991.

(Aarts 92) AARTS, E. Recognition for Acyclic-sensitive Grammars is NP-Complete. In: International Conference on Computational Linguistics, 14, 1995, Nantes, France. **Proceedings...** Nantes: ACL, 1992. p.1157-1161.

(Abeillé 88) ABEILLÉ, A. Parsing French with Tree Adjoining Grammars, some linguistic accounts. In: International Conference on Computational Linguistics, 12, 1988, Budapest, Hungary. **Proceedings**... Budapest: ACL, 1988. p.7-12.

(Abeillé et al. 90) ABEILLÉ, A. et al. Using Lexicalized TAGs for Machine Translation. In: International Conference on Computational Linguistics, 13, 1990, Helsinki. **Proceedings**... Helsinki: ACL, 1990. p.1-6.

(Abeillé 92a) ABEILLÉ, A. Synchronous Tags and French Pronominal Clitics. In: International Conference on Computational Linguistics, 14, 1992, Nantes, France. **Proceedings**... Nantes: ACL, 1992. p.60-66.

(Abeillé et al. 92b) ABEILLÉ, A. et al. Using Lexicalized TAGs for Machine Translation. In: International Conference on Computational Linguistics, 13, **Proceedings...** Helsinki:ACL, 1990.

(Agnäs 95) AGNÄS, M.S. et. al. **Spoken Language Report: First Year Report**. SRI International, Cambridge, UK, SRI Techinical Report CRC-043, 1995.

(Agustini 95a) AGUSTINI, A. **Estudo Inicial sobre o Processamento da Linguagem Natural.** Porto Alegre, 1995. 62f. Trabalho Individual (Mestrado em Informática) - Instituto de Informática, PUCRS, 1995.

(Agustini 95b) AGUSTINI, A. **Estudo Inicial sobre a Tradução Automática de Textos Escritos em Linguagem Natural.** Porto Alegre, 1995. 65f. Trabalho Individual (Mestrado em Informática) - Instituto de Informática, PUCRS, 1995.

(Agustini & Strube de Lima 97a) AGUSTINI, A. & STRUBE DE LIMA, V.L. Tratamento de Divergências Estruturais na Tradução Automática. I Encontro Nacional de Inteligência Artificial. In: Congresso da Sociedade Brasileira de Computação, 17, 1997. **Anais...**, Brasília:SBC, 1997.

(Agustini & Strube de Lima 97b) AGUSTINI, A. & STRUBE DE LIMA, V.L. **Divergences syntaxiques et traduction automatique**. Cahiers de l'Institut de linguistique de Louvain (CILL), Université Catholique de Louvain, Louvain-la-Neuve, 1997 (aceito para publicação).

(Aho 69) AHO, A. Indexed Grammars. In: IEEE Meeting on Switching and Automata Theory, 8, 1969. **Proceedings...** 1969.

(Aho, Sethi & Ullmann 86) AHO, A.V.; SETHI, R. & ULLMANN, J.D. **Compilers: Principles, Techniques and Tools**. Reading, MA: Addison-Wesley, 1986.

(Allen 88) ALLEN, J. **Natural language understanding**. Menlo Park, CA: Benjamin/Cummings, 1988.

(Araújo 93) ARAÚJO, L.A. **Por que os computadores não são capazes de traduzir? Uma resposta a partir de uma concepção pós-estruturalista de tradução.** Campinas, 1993. Dissertação (Mestrado em Lingüística Aplicada), Instituto de Estudos da Linguagem, UNICAMP, 1993.

(Arnold 86) ARNOLD, D.J. EUROTRA: an European Perspective to MT. **Proceedings of the IEEE**, v. 74, 1986.

(Arnold et al. 94) ARNOLD, D.J. et al. **Machine Translation: an Introductory Guide**. London: Blackwells-NCC, 1994.

(Aubert 94) AUBERT, F.H. **As (In)Fidelidades da Tradução: Servidões e autonomia do tradutor**. Campinas: Editora Unicamp, 1994.

(Barros 91) BARROS, E.M. **Gramática da Língua Portuguesa**, 2. ed., São Paulo: Atlas, 1991.

(Beardon, Lumsden & Holmes 91) BEARDON, C.; LUMSDEN, D. & HOLMES, G**. Natural Language and Computational Linguistics**, England: Ellis-Horwood, 1991.

(Beaven 92) BEAVEN, J.L. Shake-and-Bake Machine Translation. In: International Conference on Computational Linguistics, 14, 1992, Nantes, France. **Proceedings**... Nantes: ACL, 1992.

(Becker et al. 94) BECKER, T. et al. **A Lexicalized TAG for English**. University of Pennsylvania, Technical Report, 1994.

(Bigolin & Castilho 93) BIGOLIN, N. & CASTILHO, J.M. Ferramenta de Auxílio para Tradução de Linguagens de Especificação no Desenvolvimento de Sistemas de Banco de Dados. In: Simpósio Brasileiro de Banco de Dados, 8, 1993, Campina Grande, PB. **Anais**... Campina Grande: SBC, 1993.

(Boitet, Ghillayme & Quezel-Ambrunaz 82) BOITET, C.; GUILLAUME, P. & QUEZEL-AMBRUNAZ, M. Implementation and conversational environment of ARIANE 78.4, an integrated system for automated translation and human revision. In: International Conference on Computational Linguistics, 9, 1982, Amsterdan. **Proceedings...** Amsterdan: ACL, 1982.

(Brew 92) BREW, C. Letting the Cat out of the Bag: Generation for Shake-and-Bake MT. In: International Conference on Computational Linguistics, 14, 1992, Nantes **Proceedings**... Nantes: ACL, 1992.

(Brill 93) BRILL. E. **A Corpus-Based Approach to Language Learning.** University of Pennsylvania, Tese de Doutorado, 1993. 151f.

(Brown et al. 90) BROWN, P.F. et al. A Statistical Approach to Machine Translation. **Computational Linguistics**, v. 16, 1990.

(Catford 65) CATFORD, J.C. **Uma Teoria Lingüística da Tradução**. São Paulo: Cultrix, 1965.

(Cohen 90) COHEN, D.I. **Introduction to Computer Theory**. New York: Wiley & Sons, 1990.

(Coulthard 91) COULTHARD, M. A tradução e seus problemas. In: **Tradução: teoria e prática**. M. Coulthard & C. R. Claudas-Coulthard (orgs.), Florianópolis: Editora da UFSC, 1991. p.1-15.

(Coulon & Kayser 92) COULON, D. & KAYSER, D. **Informática e Linguagem Natural: Uma Visão Geral dos Métodos de Interpretação de Textos Escritos**. Rio de Janeiro: IBICT-Senai, 1992. 96f.

(Dias 94) DIAS, M.C.P. **O Léxico em Sistemas de Análise e Geração Automática de Textos em Língua Portuguesa**. Rio de Janeiro, 1994. Tese (Doutorado em Letras), PUC/RJ, 1994.

(Dias 96) DIAS, G. **Alinhamento Automático de Textos**. UNL, Lisboa, 1996 (plano de doutoramento, manuscrito).

(Dorr 93) DORR, B.J. **Machine Translation: A View from the Lexicon**. Cambridge, UK: MIT Press, 1993.

(Dorr 94) DORR, B.J. Machine Translation Divergences: A Formal Description and Proposed Solution. **Computational Linguistics**, v. 20, n. 4, 1994. p.597-633

(Egedi & Palmer 94) EGEDI, D. & PALMER, M. Constraining Selection Across Languages Using TAGs. In: 3ème Colloque International sur les grammaires d'Arbres Adjoints, 3, 1993, Paris. **Proceedings...** Paris: Université Paris 7, Rapport Technique TALANA-RT-94-01, 1993. p.28-31.

(Eynde 93) EYNDE, F.V. Machine translation and linguistic motivation. In: **Linguistic Issues in Machine Translation**. London: Printer Publishers, 1993.

(Faraco & Moura 91) FARACO, C. & MOURA, F.M. **Gramática**. 7. ed., São Paulo: Ática, 1991.

(Filgueiras 94) FILGUEIRAS, M. A Successful Case of Computer Aided Translation. In: Conference on Applied Natural Language Processing, 4, 1994 Stuttgart, Germany. **Proceedings**... Stuttgart: ACL, 1994.

(Freitas & Lopes 93) FREITAS, S. & LOPES, J. Um Sistema de Representação do Discurso Utilizando DRT e a Teoria do Foco. In: Simpósio Brasileiro de Inteligência Artificial, 10, 1993, Porto Alegre. **Anais**... Porto Alegre: SBC, 1993.

(Garey & Johnson 79) GAREY, M.R. & JOHNSON, D.S. **Computers and intractability: a guide to the theory of NP Completeness**. New York: W. H. Freeman, 1979.

(Grishman 92) GRISHMAN, R. Computational Linguistics: An Introduction**. Studies in Natural Language Processing**. Cambridge: Cambridge University Press, 1992.

(Harbusch & Poller 93) HARBUSCH, K. & POLLER, P. Structural Rewriting with Synchronous Systems. In: 3ème Colloque International sur les grammaires d'Arbres Adjoints, 3, 1993, Paris. **Proceedings...** Paris: Université Paris 7, Rapport Technique TALANA-RT-94-01, 1993. p.41-44.

(Hovy 93) HOVY, E. How MT Works. **BYTE**, v.18, n.1, Jan. 1993. p.167-185.

(Hudson 82) HUDSON, R. **Word Grammar**. Oxford: Blackwell, 1982.

(Hutchins & Somers 92) HUTCHINS, W.J. & SOMERS, H.L. **An Introduction to Machine Translation**. Great Britain: Academic Press, 1992.

(Isabelle & Borbeau 85) ISABELLE, P. & BOURBEAU, L. TAUM-AVIATION: its technical features and some experimental results. **Computational Linguistics**, v. 11, 1985. p.18-27.

(Joshi, Levy & Takahaschi 75) JOSHI, A.; LEVY, L. & TAKAHASHI, M. Tree Adjunct Grammars**. Journal of the Computer and System Sciences**, v.10, n.1, New York: Academic Press, 1975.

(Joshi 85) JOSHI, A.K. Tree-Adjoining Grammars: How much context-sensitivity is required to provide reasonable descriptions?**.** In: **Natural Language Parsing**, Dowty, Kartunnen, Zwick (eds.). Cambridge University Press, 1995. p.206-250.

(Joshi 92) JOSHI, A.K. Tree Adjoining Grammars and Lexicalized Grammars. In: M. Nivat and A. Podelski (eds.) **The Automata and Languages**, Philadelphia, PA, 1992.

(Joshi 94) JOSHI, A.K. Parsing Techniques. In: COLE, R.A; et al. (eds.) **Survey of the State of the Art in Human Language Technology**. Philadelphia, PA: University of Pennsylvania, 1994.

(Joshi & Srinivas 95) JOSHI, A.K. & SRINIVAS, B. **Integration of Structural and Statistical Information: Role of Complexity of Description of Primitives**. 1995 (manuscrito).

(Katoh & Aizawa 94) KATOH, N. & AIZAWA, T. Machine translation of sentences with fixed expressions. In: Conference on Applied Natural Language Processing, 4, 1994, Stuttgart. **Proceedings**... Stuttgart: ACL, 1994.

(Kinoshita et al. 94) KINOSHITA, S. et al. Improvement in Customizability Using Translation Templates. In: International Conference on Computational Linguistics, 15, 1994, Kyoto, Japan. **Proceedings**... Kyoto: ACL, 1994. p.25-31.

(Kipper 94) KIPPER, K. Porto Alegre: **Uma Experiência de Utilização do Formalismo de Gramáticas de Adjunção de Árvores para a Língua Portuguesa.** Porto Alegre, 1994. 103f. Dissertação (Mestrado em Ciência da Computação), CPGCC-UFRGS, 1994.

(Kipper & Strube de Lima 94) KIPPER, K. & STRUBE DE LIMA, V.L. Parsing Portuguese : a syntactical analyzer using TAG formalism. In: 3ème Colloque International sur les grammaires d'Arbres Adjoints, (TAG+3), 3, 1993, Paris. **Proceedings...** Paris: Université Paris 7, Rapport Technique TALANA-RT-94-01, 1993. p.65-68.

(Kipper 95) KIPPER, K. **Machine Translation and Synchronous TAGs**. Philadelphia, PA: University of Pennsylvania, Technical Report, CIS630, 1994.

(Kroch & Joshi 85) KROCH, A.S. & JOSHI, A. **Linguistic Relevance of Tree Adjoining Grammars**. Philadelphia, PA: University of Pennsylvania, Technical Report MS-CIS-85-18, 1985.

(Kroch & Joshi 87) KROCH, A.S. & JOSHI, A.K. Analyzing Extraposition in a Tree Adjoining Grammar, **Syntax and Semantics**, v.20, 1987. p.107-149.

(Kumano & Hirakawa 94) KUMANO, A. & HIRAKAWA, H. Building an MT dictionary from parallel texts based on linguistic and statistical information. In: International Conference on Computational Linguistics, 15, 1994, Kyoto, Japan. **Proceedings**... Kyoto: ACL, 1994. p.76-81.

(Ladmiral 79) LADMIRAL, J.R. **Traduzir: Teoremas para a Tradução**. Lisboa: Publicações Europa-América, 1979.

(Leffa 95) LEFFA, V.J. Resolução da Ambigüidade Lexical na Tradução Automática de Textos: Um Estudo Exploratório**.** In: I CELSUL, 1, 1995, Florianópolis. **Proceedings**... Florianópolis: Editora da UFSC, nov. 1995.

(Lopes, Marques & Roccio 93) LOPES, J.G.P.; MARQUES, N.M.C. & ROCIO, V.J.R. POLARIS: **A Real Life Lexicon**. Portugal: Uninova-CRIA, 1993.

(Luft 87) LUFT, C.P. **Dicionário Prático de Regência Verbal**. São Paulo: Ática, 1987.

(Luft 89) LUFT, C.P. **Dicionário Prático de Regência Nominal**. São Paulo: Ática, 1989.

(Luft 91) LUFT, C.P. **Novo Manual de Português**. São Paulo: Globo, 1991.

(Luz Filho 93) LUZ Filho, S. Representação Semântica de Atitudes Proposicionais através da Teoria dos Atos da Fala. In: Simpósio Brasileiro de Inteligência Artificial, 10, 1993, Porto Alegre. **Anais**... Porto Alegre: SBC, 1993.

(Mattos 90) MATTOS, G.A. **A Língua Portuguesa no Projeto Eurotra.** Letras de Hoje, v. 25, Porto Alegre, 1990. p.57-73.

(Menezes 96) MENEZES, C.E.D. **Construção Automática de Tradutores de Linguagens Naturais**. São Paulo, 1996. Relatório Técnico (Deptartamento de Engenharia Elétrica), USP, 1996.

(Nyberg 94) NYBERG, E.H. et al. Evaluation metrics for knowledge-based machine translation. In: International Conference on Computational Linguistics, 15, 1994, Kyoto, Japan. **Proceedings**... Kyoto: ACL, 1994. p.95-99.

(Nirenburg et al.92) NIRENBURG, S. et al. **Machine Translation: A Knowledge-Based Approach**. San Mateo, CA: Morgan Kaufmann Publishers, 1992.

(Parker 95) PARKER, J.B. & STAHEL, M. (eds.) **Password: English Dictionary for Speakers of Portuguese**, São Paulo: Martins Fontes, 1995.

(Pollard & Sag 87) POLLARD, C. & SAG, I. **Information-based Syntax and Semantics**. Chicago, IL: CSLI/Chicago University Press, 1987.

(Rambow & Satta 96) RAMBOW, O. & SATTA, G. Synchronous Models of Language. In: International Conference on Computational Linguistics, 16, 1996, California. **Proceedings**... California: ACL, 1996.

(Raposo 92) RAPOSO, E.P. **Teoria da Gramática: A Faculdade da Linguagem**. Lisboa: Editorial Caminho, 1992.

(Rayner, Bouillno & Carter 95) RAYNER, M.; BOUILLNO, P. & CARTER, D. **Using Corpora to Develop Limited-Domain Speech Translation Systems**. Cambridge, UK: SRI International, SRI Technical Report CRC-059, 1995.

(Rónay 81) RÓNAY, P. **A Tradução Vivida**. Rio de Janeiro: Nova Fronteira, 1981.

(Saggion & Carvalho 95) SAGGION, H. & CARVALHO, A.M.B. Análise Textual Visando a Tradução Automática. In: Congresso da Sociedade Brasileira de Computação, 15, 1995, Canela. **Anais**... Canela: SBC, 1995. p.201-212.

(Saggion 95) SAGGION, H. **Análise automática de sumários em língua portuguesa: uma aproximação ao tratamento da estrutura de um texto**. Campinas, 1995. 166f. Dissertação (Mestrado em Ciência da Computação), IMECC-UNICAMP, 1995.

(Sato 95) SATO, S. MBT2: a method for combining fragments of examples in example-based translation. **Artificial Intelligence**, n. 75, 1995. p.31-49.

(Savadovsky 88) SAVADOVSKY, P. **A Construção de Interpretadores para Linguagem Natural**. Curitiba: Edição EBAI. 1988.

(Schabes 88) SCHABES, Y. et al. Parsing Strategies with Lexicalized Grammars: Applications to Tree Adjoining Grammars. In: International Conference on Computational Linguistics, 12, 1988, Budapest, Hungary. **Proceedings**... Budapest: ACL, 1988. p.578-583.

(Schabes & Joshi 90) SCHABES Y. & JOSHI, A.K. **Parsing with Lexicalized Tree Adjoining Grammar**. Philadelphia, PA: University of Pennsylvania, Technical Report MS-CIS-90-11, 1990.

(Schabes 91) SCHABES, Y. The Valid Prefix Property and Left to Right Parsing of Tree-Adjoining Grammars. In: Second International Workshop on Parsing Technologies, **Proceedings...** Cancun, México, fev. 1991.

(Shieber & Shabes 90) SHIEBER, S.M. & SHABES, Y. Synchronous Tree-Adjoining Grammars. In: International Conference on Computational Linguistics, 13, 1990, Helsinki. **Proceedings...** Helsinki: ACL, 1990. p.253-258.

(Shieber & Shabes 91) SHIEBER, S.M. & SHABES, Y. Generation and Synchronous Tree-Adjoining Grammars. **Computational Intelligence**, 1991.

(Slocum 85) SLOCUM J. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. **Computational Linguistics**, v. 11, n. 1, mar. 1985.

(Souza e Silva & Koch 93) SOUZA E SILVA, M.C.P. & KOCH, I.V. **Lingüística Aplicada ao Português**. 5.ed., São Paulo: Cortez, 1993.

(Steedman 93) STEEDMAN, M. Categorial Grammar. **Lingua 90**. Amsterdam: North-Holland, 1993. p.221-258

(Strube de Lima & Kipper 93) STRUBE DE LIMA, V.L. & KIPPER, K. Analisador Morfológico para Tratamento de Textos em Português. In: Encontro de Processamento da Língua Portuguesa, 1, 1993, Lisboa. **Actas**..., Lisboa: Gulbenkian, 1993. p.39-44.

(Takeda 94) TAKEDA, K. Portable Knowledge Sources for Machine Translation. In: International Conference on Computational Linguistics, 15, 1994, Kyoto, Japan. **Proceedings**... Kyoto: ACL, 1994. p.85-89.

(Takeda 96) TAKEDA, K. Pattern-Based Context-Free Grammars for Machine Translation. In: International Conference on Computational Linguistics, 16, 1996, California. **Proceedings**... California: ACL, 1996. p.85-89.

(Tapanainen & Voltilainen 94) TAPANAINEN, P. & VOLTILAINEN, A. Tagging acurately - Don't guess if you know. In: Conference on Applied Natural Language Processing, 4, 1994, Stuttgart. **Proceedings**... Stuttgart: ACL, 1994.

(Trujillo 95) TRUJILLO, A. Bi-Lexical Rules for Multi-Lexeme Translation in Lexicalist MT. In: Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, 6, 1995, Leuven. **Proceedings...** Leuven, Belgium, 1995. p.48-66.

(Tucker Junior 84) TUCKER Junior, A. B. A Perspective on machine translation: Theory and Practice. **Communications of the ACM**, v. 27, n. 4, April 1984.

(Vasconcellos 91) VASCONCELOS, M. A tradução automática: a babel conquistada? In: **Tradução: teoria e prática**. M. Coulthard & C. R. Claudas-Coulthard (orgs.), Editora da UFSC, 1991.

(Vasconcellos 93) VASCONCELLOS, M. Machine Translation: translating the languages of the world on a desktop computer comes of age. **BYTE**, v.18, n.1, Jan. 1993. p.153-164.

(Vauquois & Boitet 85) VAUQUOIS, B. & BOITET, C. Automated Translation at Grenoble University. **Computational Linguistics**, v. 11, 1995.

(Vijay-Shanker 87) VIJAY-SHANKER, K. **A Study of Tree Adjoining Grammars**. Philadelphia, USA, 1987. 172f. Tese (Doutorado em Ciências da Informação) University of Pennsylvania, 1987.

(Vijay-Shanker & Joshi 88) VIJAY-SHANKER, K. & JOSHI, A.K. Feature-Structures Based Tree Adjoining Grammars. In: International Conference on Computational Linguistics, 12, 1988, Budapest, Hungary. **Proceedings**... Budapest: ACL, 1988. p.714-719.

(Vijay-Shanker 92) VIJAY-SHANKER, K. Using Descriptions of Trees in a Tree Adjoining Grammar. **Computational Linguistics**, Reading, MA: MIT Press, v.18, n.4, 1992.

(Villavicencio 95) VILLAVICENCIO A. **Avaliando um rotulador estatístico de categorias morfo-sintáticas para a língua portuguesa.** Porto Alegre, 1995. Dissertação (Mestrado em Ciênica da Computação), CPGCC- UFRGS, 1995.

(Whitelock 92) WHITELOCK, P. Shake-and-Bake Translation. In: International Conference on Computational Linguistics, 14, 1992, Nantes **Proceedings**... Nantes: ACL, 1992.

(Whitelock & Kilby 95) WHITELOCK, P., KILBY, K. **Linguistic and computational techniques in machine translation system design**. London: UCL, 1995.

(Wilks 75) WILKS, Y. An Intelligent Analyser and Understander of English. **Communications of the ACM**, v. 15, n. 5, May 1975.

(Wilks, Slator & Guthrie 96) WILKS, Y.; SLATOR, B.M. & GUTHRIE, L.M. **Electronic Words: Dictionaries, Computers and Meanings.** Cambridge, MA: MIT Press, 1996.

(Witmer 92) WITMER, D.P. **Machine Translation of a Biblical Passage**. University of Texas, Arlington, Master of Arts in Linguistics, 1992. 172f.

(Woods 73) WOODS, W.A. An experimental parsing system for transition network grammars. In: **Natural Language Processing**, R. Rustin (ed.), Algorithmics Press, New York, 1973.

(Zorzo 93) ZORZO, A.F. **Gramática Transformacional com Atributos.** Porto Alegre, 1993. 101f. Dissertação (Mestrado em Ciência da Computação), CPGCC-UFRGS, 1993.