UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

MARIA ALICE OLIVEIRA COSTA LEAL

REVISÃO BIBLIOGRÁFICA E ANÁLISE COMPARATIVA DE TÉCNICAS DE TRADUÇÃO AUTOMÁTICA

FEIRA DE SANTANA
2013

MARIA ALICE OLIVEIRA COSTA LEAL

REVISÃO BIBLIOGRÁFICA E ANÁLISE COMPARATIVA DE TÉCNICAS DE TRADUÇÃO AUTOMÁTICA

Trabalho de Conclusão de Curso apresentado ao Colegiado de Engenharia de Computação como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação da Universidade Estadual de Feira de Santana.

Orientadora: Ana Lúcia Lima Marreiros Maia

FEIRA DE SANTANA
2013

A meus pais, Ricardo e Margarida, a meu irmão, Alexandre, e a Laete, meu amor. Por alguns momentos especiais adiados indefinidamente para depois e por todo apoio e amor incondicional.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus por me mostrar que nunca devemos desistir antes de tentar, por me mostrar continuamente que tudo é possível, abençoando-me de forma inesperada e renovando minha fé.

Agradeço a minha família, principalmente meus pais, Ricardo e Margarida, meu irmão Alexandre e meus avós, por serem responsáveis pelas minhas primeiras lições e me mostrarem que o mundo é bem maior do que eu imaginava, apoiando-me em meus sonhos e me inspirando a querer voar mais alto do que imaginei que seria capaz.

Agradeço a Nino, um anjo que Deus colocou em minha vida, por todo seu amor sem medidas, por descobrir em mim uma pessoa maravilhosa que eu desconhecia, por me fazer sentir feliz e querer descobrir o mundo ao seu lado, lugar este que nenhum fardo é tão pesado que eu não consiga carregar, nada é tão difícil que não consiga aprender, mesmo momentos difíceis se transformam em coisas leves.

Agradeço a minha orientadora Ana Lúcia Maia que acompanhou minha caminhada e fez parte dela, me ajudando a superar as dificuldades deste trabalho, incentivando-me a perseverar e me apoiando de todas as formas. Agradeço ao professor Thiago Maia por apoiar a realização deste trabalho, bem como sugerir um título ao trabalho e pela sua amizade. Agradeço ao professor David de Matos do Instituto Superior Técnico (Lisboa - Portugal), que colaborou com este trabalho através de dicas importantes para que pudéssemos compreender melhor o foco da pesquisa.

Agradeço a todos colegas que compartilharam comigo suas caminhadas ao longo do curso de graduação de Engenharia de Computação da UEFS, em especial àqueles que me apoiaram e me acompanharam mais de perto: Leonardo Sampaio, Ítalo Silva, Vinícius Bittencourt, Pedro Suzart, Ian Anderson, Jairo Henrrique, André Luiz e Ronaldo Ruiz. Agradeço à equipe maravilhosa da Assessoria Especial de Informática que me acolheu e enriqueceu de forma especial minha formação, principalmente Abraão Maia e Marijalma Campos. Agradeço aos meus alunos de Introdução à Informática da Universidade Aberta à Terceira Idade, trocamos conhecimentos, experiências e momentos maravilhosos, em especial Landualdo, Elizethe, Carmenzita.

A meus amigos que, mesmo distantes fisicamente, estiveram presentes comigo todo o tempo recarregando as minhas energias, em especial Cíntia Emanoela, Silvana Argôlo, Jane, Tânia e Marla, e, em intenção, os espíritos amigos que me motivaram e me ajudaram em todos os momentos.

"When one does a thing, it appears good, otherwise one would not write it. Only later comes reflection, and one discards or accepts the thing. Time is the best censor, and patience a most excellent teacher."

(Chopin)

RESUMO

A comunicação entre os mais diversos países, culturas e línguas, é um fator fundamental para a evolução humana através da troca de conhecimento. Diante desta premissa, pode-se entender que é um desafio possibilitar a comunicação humana, com o mínimo de ambiguidades possível, através da escrita, devido às grandes diferenças linguísticas e culturais existentes. Nesse sentido, a tradução automática (TA) despontou como um dos campos de pesquisa da Ciência da Computação mais antigos e visa contribuir efetivamente para que o conhecimento possa ser transferido de forma automática sem perdas entre dois ou mais idiomas. Esta área, apesar de consolidada, no Brasil apresenta uma carência de pesquisas. Portanto, o presente trabalho se propõe a revisar e apresentar as principais técnicas de TA, apresentando um estudo comparativo destas técnicas. Além disso, são apresentados também exemplos de tradutores que utilizam três dessas técnicas.

Palavras-chave: Tradução Automática. Processamento de Língua Natural. Técnicas de Tradução Automática.

ABSTRACT

The communication among different countries, cultures and languages, is a fundamental issue in human evolution by means of knowledge exchange. Thus, one can understand that it is a challenge to make possible human communication, with a minimum of ambiguities, due to cultural and language differences. Machine translation (MT) has emerged as one of the oldest research fields in Computer Science. It aims to contribute effectively for transferring knowledge automatically without ambiguity among two or more languages. Despite the global consolidation of this area, there is a shortage of research in Brazil. Therefore, this paper aims to review and report the main MT techniques, presenting a comparative study. Moreover, it is also described translator examples using three of these techniques.

Keywords: Machine Translation. Natural Language Processing. Machine Translation Techniques.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SÍMBOLOS

$\overline{X}$	$\overline{X}$-Theory é um dos princípios da técnica PBMT.
θ-role	Papel semântico de um constituinte numa sentença em PBMT.

LISTA DE SIGLAS

MIT	*Massachusetts Institute of Technology*
ALPAC	*Automatic Language Processing Advisory Committee*
TA	Tradução Automática
NILC	Núcleo Interinstitucional de Linguística Computacional
MT	*Machine Translation*
RBMT	*Rule-Based Machine Translation*
LF	Língua Fonte
LA	Língua Alvo
KBMT	*Knowledge-Based Machine Translation*
CMU	*Carnegie Mellon University*
RDF	*Resource Description Framework*
LBMT	*Lexicon-Based Machine Translation*
LTAGs	*Lexicalized Tree Adjoining Grammars*
CBMT	*Constraint-Based Machine Translation*
LFG	*Lexical Function Grammar*
PBMT	*Principle-Based Machine Translation*
GB	*Government and Binding theory*
NP	*Nominal Phrase*
VP	*Verbal Phrase*
LCS	*Lexical Conceptual Structure*
S&BMT	*Shake & Bake Machine Translation*
SBMT	*Statistical Based Machine Translation*
EBMT	*Example-Based Machine Translation*
DBMT	*Dialogue-Based Machine Translation*
GNU	*Gnu is Not Unix*
URL	*Uniform Resource Locator*
HTML	*HyperText Markup Language*
DP	*Distributional Profiles*
TCA	*Translation Corpus Aligner*
SGML	*Standard Generalized Markup Language*
BLEU	*Bilingual Evaluation Understudy*
SPARQL	*SPARQL Protocol And RDF Query Language*

SUMÁRIO

1 INTRODUÇÃO

O crescente avanço nas tecnologias de comunicação tem permitido interligar diferentes pessoas nos mais diversos locais do mundo. Para suprir a demanda por uma comunicação rápida e compartilhamento de informações mais eficientes entre falantes de línguas distintas, faz-se necessário o uso de ferramentas de tradução automática (TA). Há um grande interesse por estas ferramentas, que envolvem a tradução de uma língua natural para outra através de um computador (SPECIA; RINO, 2002), permitindo a disseminação de informações multilíngues.

Tendo em vista a grande demanda por ferramentas de TA, a pesquisa nesta área instiga inúmeros desafios oriundos do objetivo de pesquisa inicial, a completa automatização da tradução. A TA surgiu como uma das primeiras aplicações não-numéricas da computação (SPECIA; RINO, 2002), em meados do século XX e, atualmente, impulsiona campos de pesquisa específicos da Lingüística, Ciência da Computação, Inteligência Artificial, dentre outras áreas de conhecimento.

O que se tem de concreto até hoje, normalmente não é muito valorizado pela comunidade usuária dos sistemas de tradução automática, pois, do ponto de vista não-acadêmico, para o ser humano em geral, o processo cognitivo de aprendizado de idiomas é considerado um mecanismo natural de se dominar com algum tempo de dedicação.

O crescimento da pesquisa em TA recentemente é um indicador do reconhecimento da importância de sistemas de tradução. No entanto, uma variedade de projetos de pesquisas e tipos de sistemas de tradução automática refletem o quanto esse campo ainda tem muito a ser explorado. Vale ressaltar que a essência dos problemas relacionados à TA não está principalmente relacionada à capacidade de processamento, mas a diferenças culturais na comunicação humana, seus sinais, linguagens, entendimento.

Quando se restringe o campo de atuação, o processo de se criar uma estratégia para realizar traduções torna-se mais simplificado. Dessa forma, TA tem sido destacada em aplicações específicas, como, por exemplo, manuais técnicos, contratos comerciais, propagandas, relatórios administrativos, artigos técnicos, livros de medicina etc.

Segundo Tinsley, Way e Sheridan (2010), durante as últimas décadas a Comissão Européia tem incentivado, juntamente com a Organização Mundial de Propriedade Intelectual (WIPO), projetos que estejam envolvidos em tecnologias linguísticas, mais especificamente projetos que trabalhem com técnicas de tradução automática. O objetivo é oferecer acesso a documentos técnicos, comerciais e de patentes traduzidos para a língua

materna de inventores, consumidores e pesquisadores.

Um cenário de teste apresentado por Harriehausen-Mühlbauer e Heuss (2012) mostra deficiências em alguns tradutores existentes. As ferramentas de tradução automática de propósito geral utilizadas no teste foram: o Bing Translator, o Yahoo Babel Fish, o Google Translate, uma versão *online* de demonstração do Moses e o Linguatec Personal Translator PT. Essas ferramentas foram aplicadas na tradução do inglês para o alemão da frase *"Pages by Apple is better than Word by MS"*, no contexto de textos de computação. Foi observado que todas as ferramentas citadas não obtiveram êxito, pois não levaram em conta o contexto, dessa maneira, não foram capazes de discernir quais termos deveriam ser traduzidos com conotação diferenciada.

O objetivo geral deste trabalho é realizar uma revisão bibliográfica das principais técnicas de tradução automática. Mais especificamente, destacar as características positivas e negativas dessas técnicas, a partir de critérios de comparação das mesmas e apresentar exemplos de tradutores que utilizam algumas dessas técnicas.

O presente trabalho está organizado em seis capítulos. O primeiro capítulo apresenta um breve histórico e contextualização da TA. O capítulo de Metodologia apresenta as etapas realizadas para a construção do trabalho. O capítulo Técnicas de Tradução Automática apresenta uma revisão bibliográfica das técnicas de tradução, com os seus principais fundamentos teóricos. O capítulo Análise Comparativa das Técnicas de Tradução Automática apresenta, inicialmente, um resumo das principais características das técnicas, como vantagens, desvantagens, recursos necessários para sua utilização e área em que melhor se aplica. Além disso, neste capítulo também são apresentadas uma comparação entre as técnicas e exemplos de aplicação de algumas técnicas escolhidas. Por fim, o capítulo de Considerações Finais contém alguns destaques, sugestões de trabalhos futuros e dificuldades encontradas.

2 HISTÓRICO

Este capítulo é um resumo dos principais fatos históricos apresentados por Hutchins e Somers (1992), considerados importantes para uma melhor compreensão da evolução das pesquisas sobre a Tradução Automática desde seu começo até a atualidade.

Precursores, Descartes e Leibniz (HUTCHINS; SOMERS, 1992), imaginaram a criação de dicionários baseados em códigos numéricos universais no século XVII. Foi assim que surgiu a idéia do movimento pela linguagem universal, o desejo de se ter uma língua em que não houvesse possibilidade de se compreender mais de um significado para uma mesma frase ou texto. Houve várias propostas de línguas para serem utilizadas mundialmente, a que mais ficou conhecida foi o Esperanto.

Nas primeiras traduções realizadas, no início do século XX, segundo Hutchins e Somers (1992), é possível perceber a presença constante do ser humano como parte essencial do processo de tradução. No entanto, o tradutor automático ideal que se buscava desde o início das pesquisas é um sistema com o mínimo de intervenção do ser humano. Desde então, buscou-se alcançar esse objetivo, através de pesquisas localizadas em determinados centros e troca de experiências em conferências com pesquisadores de outros grupos.

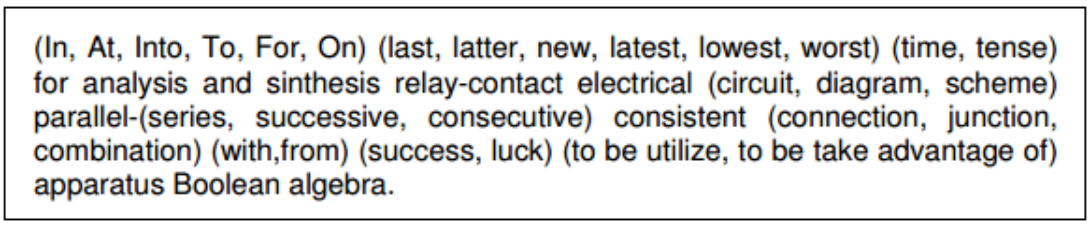
Na década de trinta (1930), o russo Petr Smirnov-Troyanskii propôs uma tradução em três estágios (HUTCHINS; SOMERS, 1992): primeiramente, um tradutor conhecedor da língua de origem da tradução se encarrega da análise da estrutura lógica das palavras, compreendendo as suas funções e formas básicas; logo após, um computador é responsável por converter essa estrutura lógica básica da língua de origem para a língua destino; no final, um editor, conhecedor da língua de destino, realizaria as adequações ao texto traduzido, para que as palavras recuperassem seu sentido mais coeso, na sua forma normalmente utilizada nas frases. Na mesma década, George Artsrouni criou um dispositivo de armazenamento que era capaz de encontrar o equivalente de uma determinada palavra em outra língua.

No início da década de 50, surgiram centros de pesquisa especializados nessa área nos Estados Unidos. Em 1952, o pesquisador Yehoshua Bar-Hillel do MIT (Massachusetts Institute of Technology) convocou uma conferência para discutir as diretrizes do futuro da pesquisa em TA com outros pesquisadores. Foi nessa conferência que, dentre outras propostas, surgiu o reconhecimento da necessidade da ajuda humana no momento anterior e posterior ao processo da TA, enquanto esta não tivesse completa autonomia para realizar de forma satisfatória a tradução. Para alguns participantes desse evento, a atitude inicial

deveria ser a demonstração técnica da viabilidade de um sistema de TA (HUTCHINS; SOMERS, 1992).

O estímulo que estava faltando, para que houvesse uma difusão em grandes proporções na pesquisa, foi uma demonstração pública de um sistema realizada em janeiro de 1954. O sistema traduziu um conjunto de sentenças do russo para o inglês, através de um vocabulário restrito de 250 palavras e somente seis regras de sintaxe. O resultado do experimento realizado com esse sistema pode ser visto na figura 1, que apresenta uma lista de todas as possibilidades de tradução para cada palavra (SILVA *et al.*, 2007). Esse sistema foi uma parceria da IBM com o pesquisador Leon Dostert da Georgetown University (HUTCHINS; SOMERS, 1992). Esse tradutor incentivou o início das pesquisas em todo mundo, mais notadamente na então União Soviética.

Figura 1 - Resultado do experimento do MIT de tradução de uma sentença do russo para o inglês.



Fonte: Silva *et al.* (2007).

Na década seguinte, houve uma diversificação nas metodologias de TA estudadas, no entanto elas podem ser caracterizadas como partes de dois grandes grupos: "força bruta" e "perfeccionista" (HUTCHINS; SOMERS, 1992). O grupo denominado como força bruta investigou a abordagem de tentativa e erro (normalmente associada à estatística). Enquanto o segundo grupo estudou mais profundamente a parte teórica como a lingüística e a lexicografia, com um foco em resoluções a longo prazo. O primeiro, sendo mais pragmático, investiu esforços em sistemas de tradução direta, enquanto o segundo contribuiu para o que se conhece atualmente como interlíngua e sistemas de transferências. A pesquisa realizada nesse período foi de grande importância para o aprimoramento da teoria da linguística, construção de dicionários automáticos e técnicas de análise sintática.

Em 1966, mais uma iniciativa surgiu, com a criação de um comitê criado pelos patrocinadores da pesquisa em TA nos Estados Unidos que divulgou um relatório sobre os resultados das pesquisas da área. O comitê ALPAC (*Automatic Language Processing Advisory Committee*) concluiu que a pesquisa nesse campo era lenta, pouco precisa e duas vezes mais cara que a tradução realizada por um tradutor humano (HUTCHINS; SOMERS, 1992). Apesar de ser largamente criticado, esse relatório influenciou os

ânimos na pesquisa da TA e foi responsável por quase uma década perdida na troca de conhecimentos e aprimoramento das metodologias.

A pesquisa foi retomada na década de 70, na qual podem ser destacados os Estados Unidos, Canadá e Europa Ocidental como países que se dedicaram em projetos como tradução de materiais técnicos, científicos, documentos administrativos e legais (HUTCHINS; SOMERS, 1992). Um grupo de pesquisadores estabelecidos em Montreal tentou, sem sucesso, a construção de um sistema em larga escala para tradução de manuais de aeronaves. No entanto, esse projeto foi reconhecido posteriormente como a base para o sistema Météo, tradutor de relatórios climáticos diários.

Em 1976, a Comissão das Comunidades Européias decidiu implantar um módulo inglês-francês de um sistema chamado Systran (HUTCHINS; SOMERS, 1992), desenvolvido por Peter Toma e utilizado durante um período como tradutor de russo-inglês para a força aérea dos Estados Unidos. Nos anos seguintes, a Comissão acrescentou ao Systran os módulos de tradução francês-inglês, inglês-italiano, inglês-alemão, além de outros pares de línguas. Em um projeto chamado Eurotra Project, a Comissão se viu impulsionada pelos últimos avanços da TA e da Linguística Computacional, dando início à pesquisa de um sistema multilíngue. Deste projeto fizeram parte diversos estudiosos de vários países da Europa.

Nos anos 80, a técnica de transferência se unificou com uma nova forma de interlíngua. Esta junção é observada nos estudos de compreensão de linguagem natural dentro da Inteligência Artificial, principalmente desenvolvidos pela instituição Carnegie Mellon University em Pittsburgh. O sustentáculo dessa atividade de pesquisa foi que a TA deve superar a análise sintática e semântica, devendo ser capaz de compreender o real contexto e informação por trás do texto. Esta abordagem apoia a utilização de representações imediatas extra-linguísticas dos significados e elementos universais. Os projetos são o sistema DLT baseado no Esperanto e o Rosetta System, que experimenta a semântica de Montague para construção da interlíngua (HUTCHINS; SOMERS, 1992).

Outras alternativas à tradução automática simples do texto surgiram, mediante o aprimoramento de tecnologias como reconhecimento da fala humana, bem como sua produção. Tais avanços deram o impulso necessário para a instalação de projetos de pesquisa nessas áreas na Inglaterra (British Telecom) e no Japão (Advanced Telecommunications Research - ATR). As técnicas de estatísticas, produto das pesquisas com a fala (reconhecimento e produção), impressionaram em sua qualidade, incentivando o interesse na aplicação dessas técnicas na TA (HUTCHINS; SOMERS, 1992). Um dos grupos que realizou pesquisas recentemente nesse campo foi o IBM Laboratories em Nova

Iorque.

O que pode ser destacado dentre as pesquisas em TA, nas décadas de 80 e 90, é o surgimento de sistemas comerciais, de acordo com Hutchins e Somers (1992). Houve uma junção entre sistemas americanos e japoneses, através de empresas de microcomputadores. No final da década de 80, foi a vez de alguns sistemas produzidos pela Siemens, Globalink, PC-Translator, Tovna e Metal System serem mesclados com outros. Muitos desses sistemas oferecem traduções simples linguísticamente, mas seu bom custo-benefício pode torná-los adequados para determinados fins. Os sistemas utilizados internamente por diversas organizações podem também ser destacados, sistemas de tradução espanhol-inglês desenvolvidos para a Organização de Saúde Panamericana (Pan-American Health Organization), sistemas projetados pela Smart Corporation para Citicorp, Ford, e o Departamento de Emprego e Imigração (Department of Employment and Immigration). Também há os projetos da Systran personalizados para algumas organizações privadas como, por exemplo, General Motors e Dornier.

Uma grande maioria dos sistemas citados necessita de uma edição posterior à tradução, apesar de que a pré-edição também é bastante popularizada em alguns sistemas. Nesses últimos, um editor humano é essencial para indicar, dentre outras coisas, o escopo de frases e os limites de cada palavra ao inserí-las. Na empresa Xerox, por exemplo, era utilizado um pré-editor da entrada do texto em inglês em um sistema da Systran composto por um vocabulário e sintaxe específico de inglês (HUTCHINS; SOMERS, 1992).

Em Oliveira *et al.* (2000), um estudo realizado com seis ferramentas de TA inglês-português e português-inglês apresentou resultados de tradução com qualidade insatisfatória. Foram utilizados excertos de textos jornalísticos da língua inglesa e portuguesa para realização dos testes, concluindo-se que menos de 50% das traduções correspondentes eram inteligíveis.

O cenário de pesquisas e resultados nessa área no Brasil foi investigado, sendo possível fazer algumas observações a respeito. Existe um grupo de pesquisa bem atuante, o NILC [1] (Núcleo Interinstitucional de Linguística Computacional) do campus da cidade de São Carlos da Universidade de São Paulo, cuja página *web* (NILC, 2011) apresenta vários projetos finalizados e em desenvolvimento. Ao longo das pesquisas do atual trabalho, pôde-se observar que o NILC tem se destacado em importantes congressos internacionais, através de publicações relevantes à área de TA. Além disso, o referido grupo faz parceria com o projeto da Linguateca (LINGUATECA, 2011), uma base de dados com diversas informações sobre o processamento computacional das duas variantes

[1] http://www.nilc.icmc.usp.br/nilc/

da Língua Portuguesa (portuguesa e brasileira).

Com o advento da Internet, novas demandas por tradução surgem e isto certamente tem impulsionado a área de pesquisa em TA. Muitos sistemas surgiram para atender esta demanda com a capacidade de fazer traduções automáticas de páginas *web*, manuais, artigos, livros, mensagens eletrônicas, conversas de bate-papos, dentre outros.

2.1 DISCUSSÕES

Através da pesquisa na literatura sobre a evolução da tradução automática (TA), foi possível sintetizar um histórico sobre os sistemas de tradução automática, desde seu surgimento. Dessa maneira, identificou-se as motivações iniciais da pesquisa, seus entraves no decorrer do tempo, os grupos de pesquisadores que se destacaram, os tipos de sistemas desenvolvidos e as metodologias utilizadas. Esse panorama da TA facilitou uma melhor compreensão sobre esta área, contribuindo para uma análise mais rica sobre as técnicas de tradução automáticas.

Pode-se estabelecer que o que impulsionou como motivação das pesquisas em TA foi o desejo por uma língua universal sem ambiguidades e por uma mecanização completa do processo da tradução automática, sem intervenção do ser humano em nenhuma das etapas. Destacam-se alguns entraves na evolução da história da TA: as limitações iniciais de não se conseguir realizar o processo de tradução, sem o auxílio do ser humano, e um relatório publicado pelo comitê ALPAC, concluindo que a pesquisa em tradução automática não compensava o esforço em custo-benefício.

Ao longo dos anos, o estudo em TA foi se disseminando em diversos grupos de pesquisa mundialmente. Pode-se destacar os Estados Unidos como um dos locais pioneiros, além do Canadá e Europa Ocidental. Posteriormente, na década de 90, podem ser apontados também a Inglaterra e o Japão.

Na década de 60, iniciaram-se as discussões sobre a formalização das metodologias do processo de tradução automática. Um estudo para a criação de uma técnica empírica, por tentativa e erro, associada à estatística, foi a base para a técnica de tradução por abordagem direta. Em outra vertente, surgiram estudos sobre linguística e lexicografia que levaram ao desenvolvimento das técnicas de abordagem indireta conhecidas como interlíngua e técnica por transferência. Nos anos 80, pode-se notar tentativas de unificação da técnica de transferência com um novo modelo de interlíngua, abrindo o caminho para a utilização de sistemas de metodologias e técnicas híbridas.

Sob outro aspecto, o histórico de TA mostrou que esta é uma área antiga de pesquisa em Computação e, no entanto, continua tão promissora quanto outras áreas de

conhecimento estudadas mais recentemente. À medida que os sistemas computacionais evoluíam, com um crescente aumento no poder de processamento dos mesmos, surgiam novas possibilidades para a TA. Inicialmente, o processamento disponível era uma barreira aos sistemas de tradução, posteriormente a comunicação humana em suas variações culturais se tornou o principal limitante no estabelecimento de metodologias e padrões eficientes ao processo da tradução automática.

3 METODOLOGIA

A metodologia aplicada ao presente trabalho consistiu de duas fases: revisão bibliográfica e análise comparativa das técnicas de Tradução Automática (TA).

A revisão bibliográfica foi iniciada com o levantamento e seleção de material bibliográfico teórico de referência geral sobre o tema do trabalho. A busca por livros publicados na área de TA foi realizada através da ferramenta da pesquisa *online Google Books* e da livraria *online Amazon*, com os termos de pesquisa "Tradução Automática", "TA", "*Machine Translation*", " MT " (*Machine Translation*), "técnicas de tradução automática" e "*machine translation techniques*".

Foram identificados os seguintes livros como referência geral do assunto de tradução automática: "*Machine Translation, Past, Present and Future*" (HUTCHINS, 1986), "*An Introduction to Machine Translation*" (HUTCHINS; SOMERS, 1992), "*Machine Translation An Introductory Guide*" (ARNOLD *et al.*, 1994), "*Speech and Language Processing*" (JURAFSKY; MARTIN, 2000), "*Readings in Machine Translation*" (NIRENBURG; SOMERS; WILKS, 2003), "*Learning Machine Translation*" (GOUTTE *et al.*, 2009), "*Machine Translation Its Scope and Limits*" (WILKS, 2009), "*Statistical Machine Translation*" (KOEHN, 2012), "*Handbook of Natural Language Processing and Machine Translation*" (OLIVE; CHRISTIANSON; MCCARY, 2011).

Para estudar e revisar o histórico da Tradução Automática e os seus conceitos iniciais, o livro "*An Introduction to Machine Translation*" (HUTCHINS; SOMERS, 1992) foi escolhido dentre os livros pesquisados, por apresentar uma introdução mais abrangente sobre o assunto. Os livros mais recentes não foram utilizados nesta etapa, como uma referência para o estudo inicial, por não apresentarem conceitos básicos sobre a área de pesquisa. Estes já tratam especificamente de uma única técnica de tradução e foram então consultados nos estudos específicos das técnicas.

Ainda na etapa de revisão bibliográfica foi necessária a escolha das técnicas de tradução a serem abordadas. Para este propósito, foram identificados artigos e grupos de pesquisa atuais e relevantes sobre as técnicas de Tradução Automática. Primeiramente, a seleção das técnicas para realização do estudo mais específico deste trabalho foi feita a partir de Dorr, Jordan e Benoit (1998) e Silva *et al.* (2007). Estes trabalhos foram utilizados como parâmetro para a escolha das técnicas, pois apresentaram um conjunto maior de técnicas de TA que os livros e outros artigos consultados. Dessa maneira, foi possível selecionar os nomes das técnicas pesquisadas: "RBMT" - "*Rule Based Machine Translation*", "KBMT" - "*Knowledge Based Machine Translation*",

"LBMT" - "*Lexicon-Based Machine Translation*", "CBMT" - "*Constraint-Based Machine Translation*", "PBMT" - "*Principle-Based Machine Translation*", "S&BMT" - "*Shake & Bake Machine Translation*", "RBMT" - "*Rule Based Machine Translation*", "SBMT" - "*Statistical Based Machine Translation*", "EBMT" - "*Example Based Machine Translation*" e "DBMT" - "*Dialogue-Based Machine Translation*".

Foram realizadas novas pesquisas, utilizando os nomes das técnicas selecionadas, com as ferramentas *online Google*, *Google Scholar* e, principalmente, na página *web Machine Translation Archive* (HUTCHINS, 2013). O repositório de artigos *Machine Translation Archive* é mantido pelo pesquisador John Hutchins, que dispõe de um número considerável de livros e artigos publicados na área de TA. É importante ressaltar como desvantagem de se utilizar este repositório que não se sabe os critérios utilizados pelo pesquisador John Hutchins para a seleção de artigos colocados no referido repositório.

Após a seleção e o estudo de alguns artigos levantados sobre as técnicas de TA escolhidas, foram sintetizadas as principais características de cada uma das técnicas estudadas: descrição do processo de tradução, vantagens, recursos necessários e limitações.

Na etapa de análise comparativa das técnicas de tradução, foram escolhidos e descritos critérios de comparação para as técnicas, com o intuito de apresentar um quadro comparativo das mesmas. Os critérios foram definidos com base nas características que mais foram documentadas nos artigos escolhidos para estudo, principalmente provenientes do repositório *Machine Translation Archive*. Os critérios utilizados na comparação foram: disponibilidade de literatura, amadurecimento da tecnologia, complexidade de implementação, custo de implementação, custo de manutenção, desempenho.

Por fim, foram escolhidas com base no quadro comparativo três técnicas que mais se destacaram por critérios específicos, que as beneficiam em determinadas aplicações. Para ilustrar tais aplicações, foram selecionados como exemplo os *frameworks* e ferramentas mais documentados baseados nestas técnicas, a partir dos artigos pré-selecionados na fase de revisão bibliográfica.

4 TÉCNICAS DE TRADUÇÃO AUTOMÁTICA

De acordo com Hutchins e Somers (1992), os seguintes passos devem ser adotados inicialmente para o projeto de um tradutor automático:

a) decidir se o tradutor será bilíngue (um par de línguas) ou multilíngue (mais de um par de línguas);

b) decidir se a técnica a ser utilizada será a direta, a de transferência ou a de linguagem intermediária;

c) decidir se o sistema terá ou não intervenções externas;

d) organizar os dados léxicos.

Os sistemas bilíngues podem ser unidirecionais, traduzir somente de uma língua para a outra língua do par, ou bi-direcionais, fazer os dois caminhos da tradução, da primeira língua para a segunda, bem como da segunda para a primeira.

Devido à dificuldade de se construir sistemas bilíngues bi-direcionais, a maioria dos sistemas tradutores bilíngues são, na verdade, dois sistemas de tradução uni-direcionais em execução na mesma máquina.

Uma das vantagens de se adotar um sistema bilíngue a um sistema multilíngue (HUTCHINS; SOMERS, 1992) é a possibilidade de se explorar similaridades de vocabulário e sintaxe entre o par de línguas analisado. Um exemplo de sistema bilíngue que se utiliza dessa estratégia é o sistema Météo, o qual aproveita semelhanças léxicas e sintáticas entre o inglês e o francês durante a tradução.

Silva *et al.* (2007) classifica as técnicas de tradução automática mais recentes em dois tipos: aquela que utiliza um conhecimento profundo, linguístico, a técnica fundamental, e aquela que utiliza um conhecimento superficial ou empírica, a técnica empírico. Diferentes técnicas podem trabalhar juntas, através de diferentes combinações, em favor de um sistema híbrido (*multi-engine*) que visa obter benefícios em aplicações de tradução automática para pares de línguas específicos.

4.1 TÉCNICA FUNDAMENTAL

As técnicas fundamentais de tradução automática são aqueles que fazem uso das restrições sintáticas, lexicais ou semânticas, sobre as línguas naturais envolvidas, caracterizando teorias linguísticas bem definidas (SILVA *et al.*, 2007). Nas seções

seguintes, são descritas algumas dessas técnicas que aplicam os diferentes tipos de conhecimento: Baseado em Regras (RBMT), Baseada em conhecimento (KBMT), Baseada em Léxico (LBMT), Baseada em Restrições (CBMT), Baseada em Princípios (PBMT) e *Shake and Bake* (S&BMT).

4.1.1 Baseada em Regras

No início das pesquisas da tradução automática, as técnicas mais utilizadas eram baseadas em regras, dirigidos linguisticamente (KAUCHAK, 2006). Hoje, a maioria das técnicas de tradução utilizam combinações baseadas em regras e empíricas. No entanto, ainda existem sistemas de tradução comercial que usam uma abordagem unicamente baseada em regras. Estes últimos são eficientes e de propósito geral. A maioria dos sistemas comerciais traduzem combinando dicionários de tradução, expressões idiomáticas, dicionários semânticos, com regras geradas por humanos.

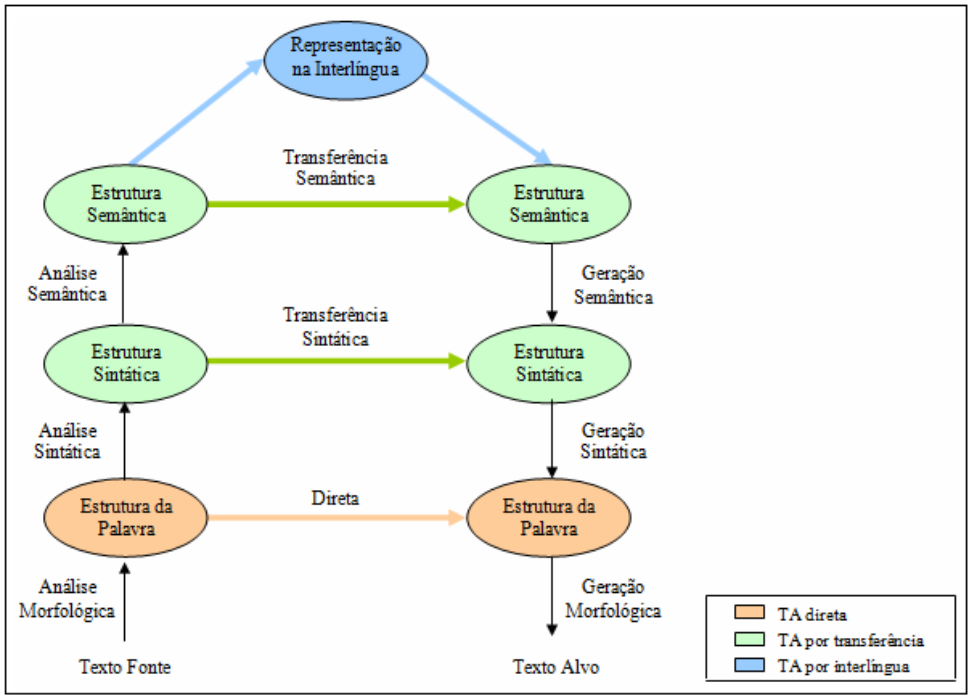
Os sistemas de TA baseada em regras, *Rule-Based Machine Translation* (RBMT), realizam a transferência do texto em Língua Fonte (LF) para Língua Alvo (LA), através de regras que representam o conhecimento em diferentes níveis linguísticos (SILVA *et al.*, 2007). Por exemplo, em uma transferência lexical, as características e restrições de itens individuais são codificadas num mecanismo de controle, por meio de regras.

Três técnicas podem ser aplicadas ao sistema de TA baseada em regras: técnica de tradução direta, por transferência e interlíngua (HARSHAWARDHAN, 2011). De acordo com Hutchins e Somers (1992), a primeira técnica de tradução criada foi a de abordagem direta. Os sistemas baseados nessa técnica foram classificados como primeira geração de sistemas de tradução automática. Posteriormente, surgiram duas técnicas de abordagem indireta: a interlíngua e a técnica de transferência. Estas, por conseqüência foram designadas como a segunda geração dos sistemas de tradução automática.

A segunda geração das técnicas de TA se caracterizou como mais sofisticada, objetivando suprir a deficiência de qualidade da tradução pela abordagem direta. A tática envolvida por essa nova geração visa analisar os textos da língua de origem em alguma forma de representação intermediária - formas equivalentes em significado com a língua de origem. Através da figura 2, pode-se observar como cada uma das técnicas citadas se organizam em relação à profundidade do conhecimento linguístico envolvido na tradução.

No topo da figura 2, estão as abordagens de interlíngua. Para traduzir de uma língua fonte para uma língua alvo, o texto da LF é primeiro traduzido em uma representação de conhecimento independente de língua. A partir dessa representação, o texto na LA é gerado. As limitações de criação da linguagem de

Figura 2 - Relação entre as técnicas direta, por transferência e interlíngua.



Fonte: Silva *et al.* (2007).

representação intermediária representaram um entrave no avanço das pesquisas dessa técnica (KAUCHAK, 2006).

Indo para a base da figura 2, estão as técnicas de transferência, nas quais ambos conhecimentos semânticos e sintáticos são extraídos do texto (KAUCHAK, 2006). Regras são aplicadas que convertem esta representação da LF em uma representação da LA. A partir dessa representação da LA, o texto final na LA é gerado.

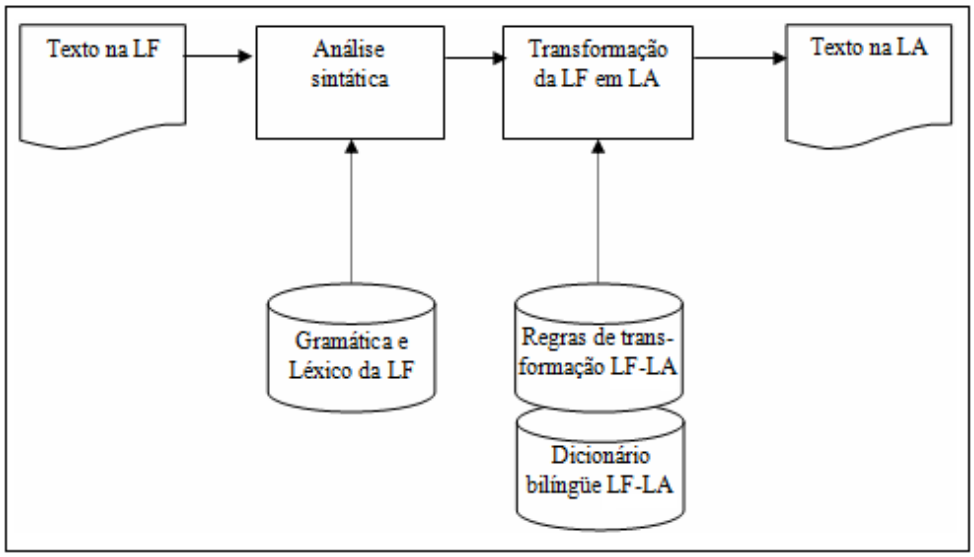
4.1.1.1 Técnica por Abordagem Direta

A técnica de tradução por abordagem direta envolve a tradução de uma língua fonte (LF) a uma língua alvo (LA) "diretamente" (figura 3), ou seja, sem nenhuma outra etapa intermediária (HUTCHINS; SOMERS, 1992). Quando do seu surgimento, a técnica direta realizava uma análise morfológica na LF, identificando as palavras do texto e alterando algumas palavras para sua forma simplificada, sem flexões de gênero ou número. A tradução era então finalizada, através de uma transcrição literal de palavra por palavra, através de um dicionário LF-LA. Inicialmente, não havia análise sintática ou semântica, no entanto, às vezes uma reordenação de adjetivos era realizada para conferir uma melhor qualidade à tradução.

Segundo Hutchins e Somers (1992), era possível obter resultados equivalentes ao se

comparar uma tradução realizada pela técnica direta e uma tradução realizada por uma pessoa sem conhecimentos profundos nas línguas envolvidas na tradução.

Figura 3 - Processo de TA pela técnica da abordagem direta.



Fonte: Silva *et al.* (2007).

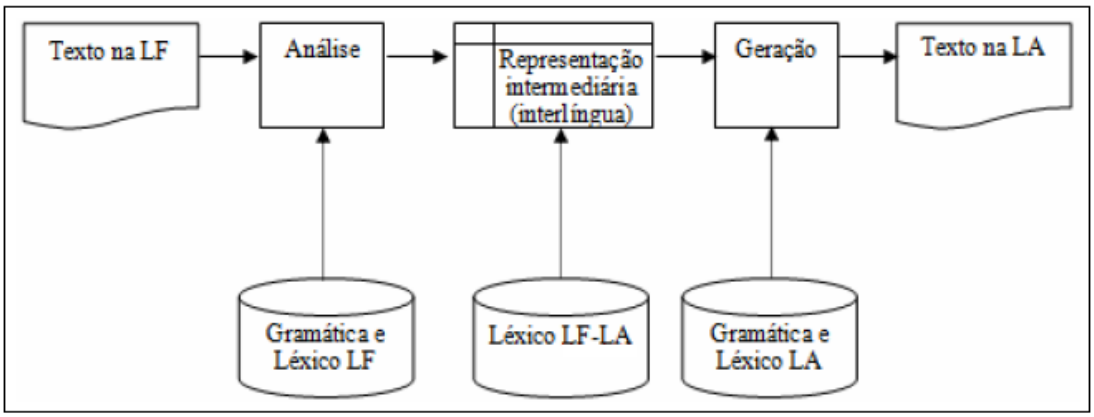
Em alguns casos, é válido utilizar esta técnica, inclusive existem sistemas comerciais que utilizam esta abordagem em conjunto com abordagens diferentes (HUTCHINS; SOMERS, 1992), apesar das restrições da técnica direta em fornecer uma tradução de qualidade. Ao combinar uma abordagem direta com uma indireta, existe uma divisão de responsabilidades. Quando as LF e LA têm muitas similaridades, utiliza-se a abordagem direta para traduzir boa parte do texto. Dessa forma, é possível concentrar esforços em áreas da gramática e sintaxe, onde as linguagens diferem mais acentuadamente, utilizando a abordagem indireta.

4.1.1.2 Técnica de Interlíngua

A técnica de Interlíngua foi a primeira técnica de abordagem indireta a ser criada (HUTCHINS; SOMERS, 1992). Sua essência é realizar a tradução através da representação da LF em uma língua intermediária e, a partir dessa última, gerar o texto traduzido na LA (figura 4). A língua intermediária conhecida como interlíngua é, ao mesmo tempo, uma projeção da LF e uma base para a LA. A interlíngua significa uma intermediação entre dois ou mais pares de línguas, de forma neutra, sem traços de nenhuma das línguas envolvidas, ela deve conter toda informação necessária para gerar a língua alvo, sem precisar fazer nenhuma análise sobre a língua fonte.

É comum a utilização da técnica indireta interlíngua em sistemas de tradução

Figura 4 - Processo de TA pela técnica de interlíngua.



Fonte: Silva *et al.* (2007).

multilíngue, que envolvem mais de um par de línguas (HUTCHINS; SOMERS, 1992). A vantagem desta técnica para aplicações dessa natureza é facilmente notada, se for analisado o baixo custo de se adicionar módulos com suporte a novas línguas a um sistema de tradução interlíngua já em uso. O acréscimo dos novos módulos não implica na modificação dos que se encontram em funcionamento, pois todos compartilham de um mesmo padrão de representação, a interlíngua. Por exemplo, um sistema de tradução português-inglês e inglês-português, feito com a técnica interlíngua, possui os módulos de geração da interlíngua a partir do português e a partir do inglês. Também existem os módulos de produção da língua alvo, a partir da interlíngua para o português e para o inglês. Caso se deseje adicionar o suporte à tradução do inglês para o francês, somente é necessário acrescentar o módulo da geração do francês a partir da interlíngua.

A maior desvantagem da técnica em questão é a dificuldade de se construir a interlíngua, de forma que viabilize uma representação neutra de transição entre qualquer língua natural (HUTCHINS; SOMERS, 1992).

4.1.1.3 Técnica por Transferência

Hutchins e Somers (1992) destacam uma outra técnica de abordagem indireta bastante utilizada, a técnica por transferência. De acordo com essa técnica, as traduções envolvem estágios temporários de conversão entre a língua fonte e a língua destino. As etapas intermediárias são consideradas modelos de transferência, através do qual o texto em uma língua será gradativamente traduzido por abstrações intermediárias dessa língua.

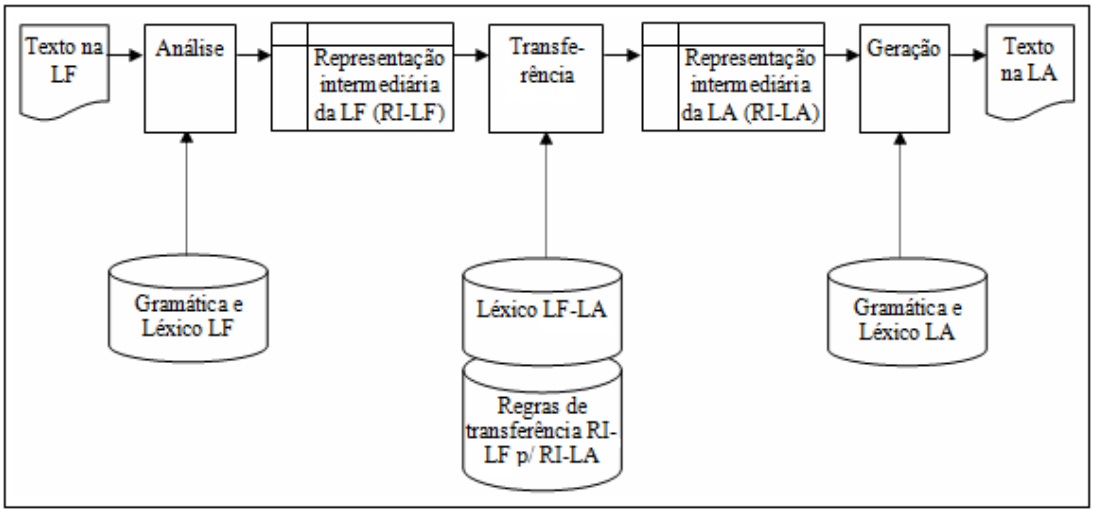
Como indicado na figura 5, a tradução pode ser observada em três estágios (HUTCHINS; SOMERS, 1992). Primeiramente, é gerada uma representação da LF com todas as informações linguísticas da mesma. Realizando-se a seguir uma análise

sobre a representação obtida, aplica-se regras pré-determinadas para converter o modelo intermediário da LF em um outro modelo intermediário com as características linguísticas da LA. Por fim, o texto na LA é produzido, através da equivalência entre a representação intermediária e a língua alvo.

A técnica de tradução indireta por transferência utiliza uma representação intermediária das línguas durante a tradução, assim como na técnica da interlíngua. No entanto, eles se diferenciam pelo fato que a representação intermediária depende das características das línguas envolvidas no processo da TA (HUTCHINS; SOMERS, 1992).

Não há um padrão de representação intermediária, como no interlíngua, e, para cada novo par de língua suportado pelo tradutor, é necessário que haja mais de dois módulos de representação intermediária (dois para partir de uma das línguas do par e gerar a outra e mais dois módulos para fazer o caminho inverso) (HUTCHINS; SOMERS, 1992).

Figura 5 - Processo de TA pela técnica indireta por transferência.



Fonte: Silva *et al.* (2007).

Apesar de ser aparentemente mais custosa em sua implementação, a técnica por transferência é mais utilizada que a técnica interlíngua (HUTCHINS; SOMERS, 1992). A justificativa principal é a de que é mais complexo projetar uma língua de representação neutra. Um outro problema, que decorre do já citado, é a dificuldade de se realizar uma análise sobre os textos em LF e LA, em um nível alto de abstração, encapsulando os detalhes culturais, linguísticos, a partir do modelo intermediário.

4.1.2 Baseada em Conhecimento

A técnica baseada em conhecimento (*Knowledge-Based Machine Translation* - KBMT) se concentra em associar conceitos mais profundos a um léxico, como informações morfológicas, sintáticas e semânticas, sem no entanto estabelecer uma relação desse conhecimento à sintaxe (DORR; JORDAN; BENOIT, 1998). Segundo Silva *et al.* (2007), o objetivo é fornecer dados para que o sistema seja capaz de manipular os conceitos e produzir novas inferências. As estruturas utilizadas para armazenar o conhecimento normalmente são as ontologias e os modelos de domínio, razão pela qual os sistemas KBMT estão diretamente influenciados pelo conhecimento de senso comum e pelas informações e características extralinguísticas.

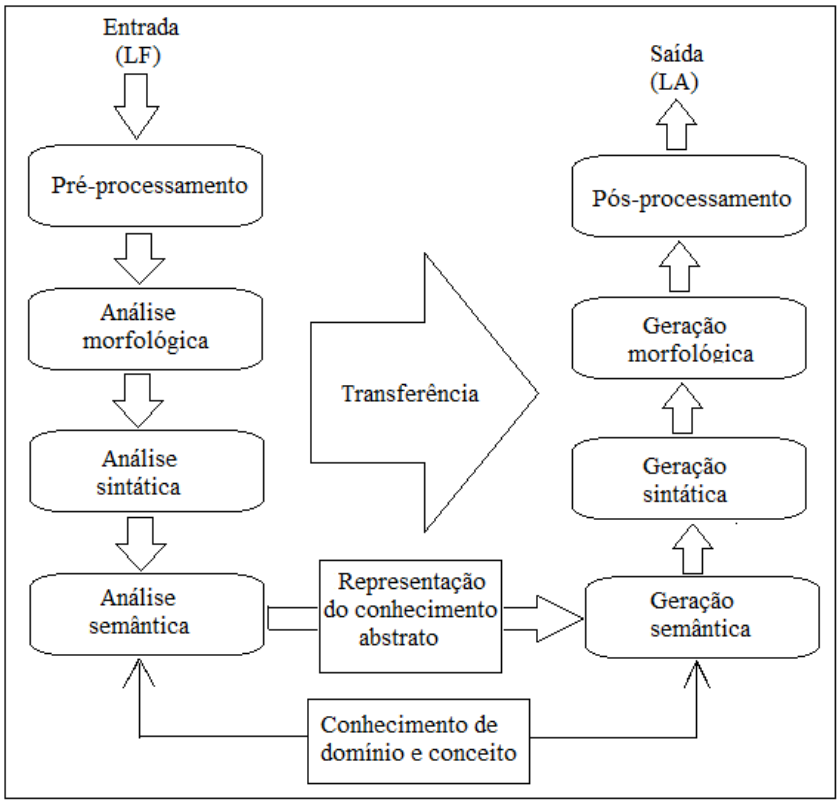
Sistemas baseados em conhecimento podem ser descritos como sistemas que utilizam bases de dados de terminologias, organizados de forma sistemática para cada domínio (VERTAN, 2005). A ontologia não é visível claramente, pois não é declarativa, ou seja, está implícita na ordem da terminologia (VERTAN, 2005). A abordagem baseada em conhecimento utiliza uma larga base de dados semântica, que representa conceitos de um determinado domínio, para realizar a transferência de um texto da língua fonte (LF) para a língua alvo (LA) (ARNOLD *et al.*, 1994). A figura 6 é um exemplo de arquitetura de um sistema KBMT.

A relação dada entre um fenômeno do mundo real (ações, eventos) e a referência deste a partir de um texto resulta do mesmo. Portanto, o mapeamento feito de um texto é consequência da visão única de mundo, do significado que cada símbolo linguístico representa para um indivíduo. Partindo do pressuposto que os símbolos e compreensões são universais a todos os falantes de uma língua, pode-se inferir que as bases de conhecimentos são como a interlíngua no papel de representação intermediária para a tradução automática (HUTCHINS; SOMERS, 1992).

É consenso que através de uma representação mais profunda do texto, será possível obter uma tradução de mais alta qualidade, dessa forma modelos de domínio podem ser as ferramentas essenciais para se desenvolver uma compreensão mais consistente do texto a ser traduzido. É nesta premissa que se baseia a abordagem baseada em conhecimento, que ganha força a partir da consideração de que, ao se obter um resultado de tradução automática mais preciso, não será necessário investir tanto na etapa de pós-edição do trabalho traduzido. Esta última tarefa tem se apresentado como muito dispendiosa em termos de tempo e custo (financeiro) (ARNOLD *et al.*, 1994).

Para uma melhor demonstração dessa abordagem, Arnold *et al.* (1994) citou alguns protótipos de TA desenvolvidos pelo *Center for Automatic Translation* (Centro para

Figura 6 - Arquitetura de um sistema KBMT.



Fonte: Adaptada de Vertan (2005).

Tradução Automática) na *Carnegie Mellon University* (CMU) em Pittsburgh, na década de 80. Os protótipos eram baseados em conhecimento no contexto de manuais de instrução de computadores pessoais do inglês para o japonês e vice-versa. Os seguintes módulos compunham os sistemas desenvolvidos:

a) uma ontologia de conceitos;

b) análise lexical e gramatical para inglês e japonês;

c) geração lexical e gramatical para inglês e japonês;

d) mapeamento de regras entre sintaxe do inglês/japonês e da interlíngua.

Segundo Arnold *et al.* (1994), foram especificados 1500 conceitos em detalhes num pequeno universo de aproximadamente 900 palavras que estavam diretamente relacionadas com a interação entre os computadores pessoais e seus usuários. Na ontologia, as informações armazenadas seguem o preceito da orientação a objetos, em que objetos são substantivos e eventos são verbos que representam ações. Através da ontologia,

os conceitos são descritos em uma linguagem de representação estrutural, em que se encontram interligados hierarquicamente.

O modelo de domínio projetado consiste nos conceitos importantes do domínio em questão, mas não possibilita extrapolar na inferência de novas informações, baseadas no conhecimento que já se encontra descrito na ontologia. Um grande papel do modelo é dar suporte a correções de problemas de ambiguidade do texto, a partir de restrições sobre o contexto de cada evento e a que papel está associado cada objeto (ARNOLD *et al.*, 1994).

Hutchins e Somers (1992) afirmam que a tarefa de desenvolvimento de bases de conhecimento é restrita a contextos e aplicações específicos. Apesar de haver tentativas de se realizar a derivação de uma linguagem de interlíngua a partir dos modelos de domínio para geração de outra língua, esta técnica não tem apresentado resultados favoráveis. Entretanto, estas bases de conhecimento são utilizadas com mais sucesso na desambiguação de textos.

De acordo com Vertan (2005), os sistemas de TA necessitam de algumas camadas básicas de conhecimento, como por exemplo conhecimento conceitual, conhecimento do mundo e conhecimento de situação. O primeiro tipo citado normalmente está associado com os conhecimentos teóricos da academia. Por exemplo, no contexto de animais, mas especificamente cachorro, seria a classificação do mesmo em espécie, família, etc. O conhecimento do mundo pode ser a associação e inferências que podem ser feitas a partir do conhecimento conceitual e de constatações feitas com os fatos e fenômenos do dia a dia, por exemplo observar que uma espécie de cachorro normalmente se apresenta em uma determinada cor. O último conhecimento citado envolve uma relação entre informações sobre situações temporárias e os outros conhecimentos, exemplo "o cachorro do José está na sala de estar". Se não há uma exigência grande de tratamento semântico, o sistema de tradução automática pode guardar a camada de conhecimento no léxico, através de características semânticas, papéis, restrições, dentre outros.

Os desafios de sistemas baseados em conhecimento são: o grande custo de se construir as bases de conhecimento; a dificuldade de se definir a abrangência da base, o nível de aprofundamento; e a escolha das ferramentas utilizadas na construção da base de conhecimento, a linguagem de representação e suas propriedades (VERTAN, 2005).

Por outro lado, segundo Vertan (2005), existe uma transparência para o criador ou mantenedor da base de conhecimento no sentido de facilmente identificar onde está e o que é determinado conceito, apesar de não ser possível prever o que pode ser derivado das regras de inferência. As bases de conhecimento semânticas são declarativas e, dessa forma, têm a vantagem de serem mantidas de forma independente dos sistemas que a utilizam,

sendo facilmente substituídas, caso seja preciso, bem como poderem ser utilizadas em outros tipos de sistemas: recuperação de informação, processamento de língua natural, dentre outros.

Em 2001, foi publicado um artigo sobre a *web* semântica (BERNERS-LEE; HENDLER; LASSILA, 2001), uma forma diferente de representar dados com significados na *web* através do padrão RDF (*Resource Description Framework*), e desde então houve um interesse crescente no campo de pesquisa das ontologias. Além de ter algumas palavras chaves indicando o conteúdo das páginas *web*, as páginas com esse novo formato representariam um conjunto de informações manipuláveis por computadores, não só visíveis e compreensíveis ao ser humano. Vertan (2005) observa que esse padrão é similar a construção de ontologias, dessa forma pode-se considerar o uso das páginas *web* semânticas como base de conhecimento para sistemas de tradução automática, bem como para outros fins, como recuperação de informações, sumarização de textos. Dessa maneira, bases de conhecimento semântico largamente validadas sanariam os inconvenientes citados anteriormente de, por exemplo, ser necessário se construir inteiramente uma determinada base para se utilizar em um TA.

Alguns exemplos de sistemas KBMT são o Translator , o KBMT-89, o KANT e o projeto UNL (SILVA *et al.*, 2007).

4.1.3 Baseada em Léxico

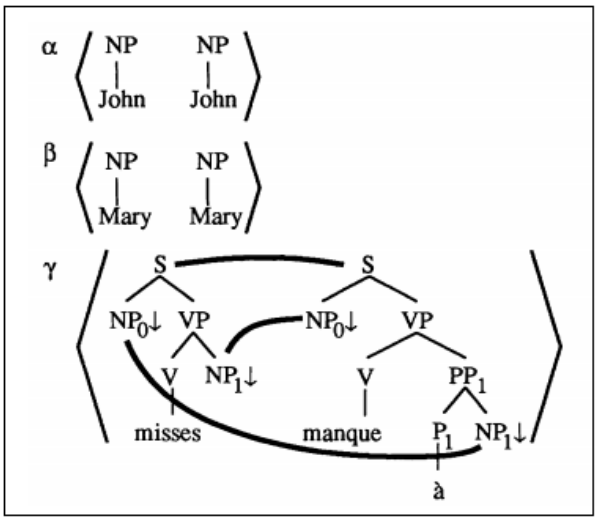
Lexicon-Based Machine Translation (LBMT), ou sistema de TA baseado em léxico, é um tipo de tradutor em que a transferência do texto em LF para LA ocorre através de regras de mapeamentos lexicais (SILVA *et al.*, 2007). Para uma melhor compreensão sobre esta abordagem, será detalhado um modelo projetado com este princípio.

Abeillé, Joshi e Schabes (1990) descrevem um exemplo de sistema baseado em léxico, em que é utilizado o formalismo das gramáticas de adjunção de árvores lexicalizadas (LTAGs - *Lexicalized Tree Adjoining Grammars*) para realizar a transferência da LF para a LA, através da correspondência de diversas unidades elementares das línguas em questão, sem necessidade de utilização de uma interlíngua ou de outro conhecimento. De forma geral, as regras de transferência ocorrem pela equivalência de nós de árvores, que significam desde representações de palavras até bloco de textos maiores. Esta abordagem aplica dependências semânticas e sintáticas nas estruturas de itens lexicais.

A tradução ocorre em três etapas, tendo como objetivo combinar as derivações de LTAG fonte e alvo através de um léxico de transferência. Na primeira etapa, são geradas árvores elementares com as características e informações do trecho de texto em LF. Na segunda etapa, a árvore de derivação em LA é criada a partir da árvore em LF, para isso

são combinadas todas as árvores elementares da derivação fonte com a de derivação alvo, através do léxico de transferência. Por fim, a árvore de derivação alvo dá origem ao texto traduzido (ABEILLÉ; JOSHI; SCHABES, 1990).

Figura 7 - Exemplo de léxicos de transferência em um sistema LBMT inglês-francês.



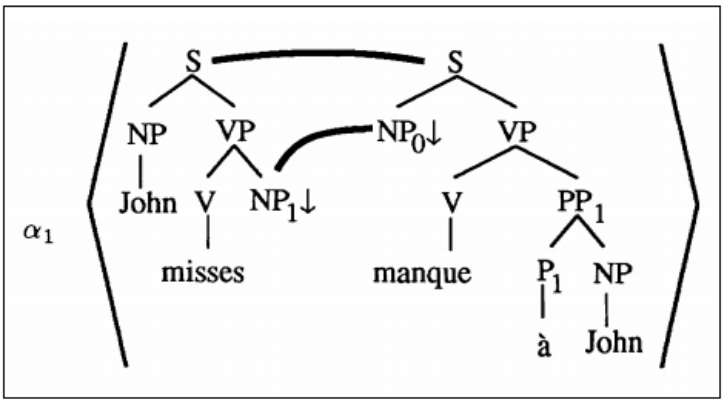
Fonte: Abeillé, Joshi e Schabes (1990).

Os pares de árvores em LF e LA, em que alguns de seus nós podem estar relacionados (figura 7), são chamados léxicos de transferência. A ação conjunta e continuada de dois ou mais léxicos de transferência, realizando operações de substituição ou adjunção em um dos pares de árvores selecionados consiste no processo de tradução da LF para a LA (ABEILLÉ; JOSHI; SCHABES, 1990).

Por exemplo, na figura 7, tem-se o léxico de transferência α, em que representa o substantivo John tanto na LF (inglês) quanto na LA (francês), e também o léxico de transferência γ que mostra a relação entre a estrutura da sentença em inglês contendo o verbo *misses* e o correspondente em francês, *manque*. Pode-se observar linhas mais reforçadas na figura 7, no léxico γ, que representam as relações dos nós das duas árvores em inglês e em francês. É através dessas conexões e da combinação de α e γ que é gerada α_1, ou seja o NP_0 de γ é substituído por John através de α (figura 8). O par de árvores derivado na figura 8 inclui em um lugar só as informações do substantivo John e da ação realizada pelo mesmo. Da mesma forma, substitui-se NP_1 de α_1 (figura 8) por Mary através de β, gerando α_2 (figura 9).

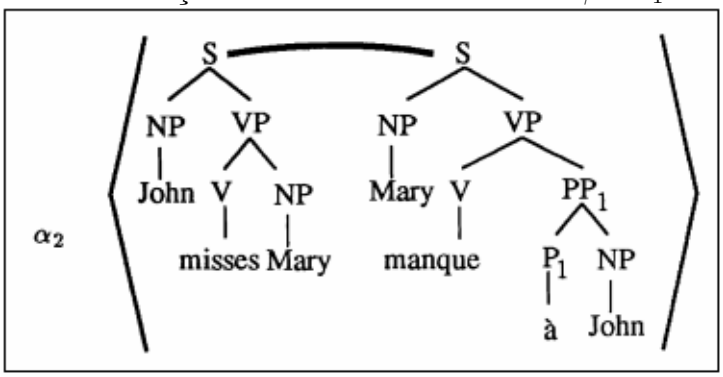
De acordo com Abeillé, Joshi e Schabes (1990), a abordagem baseada em léxico utilizando LTAGs permite uma correspondência estável entre estruturas gramaticais grandes de LF e LA. É possível que sejam observadas cuidadosamente as características particulares de cada língua, devido ao fato de que as gramáticas são lexicalizáveis.

Figura 8 - Resultado da combinação dos léxicos de transferência α e γ em um sistema LBMT.



Fonte: Abeillé, Joshi e Schabes (1990).

Figura 9 - Resultado da combinação dos léxicos de transferência β e α_1 em um sistema LBMT.



Fonte: Abeillé, Joshi e Schabes (1990).

Também é citado que alguns problemas de divergências, como de categoria, de estrutura, de tema, podem ser desconsiderados nessa abordagem da TA.

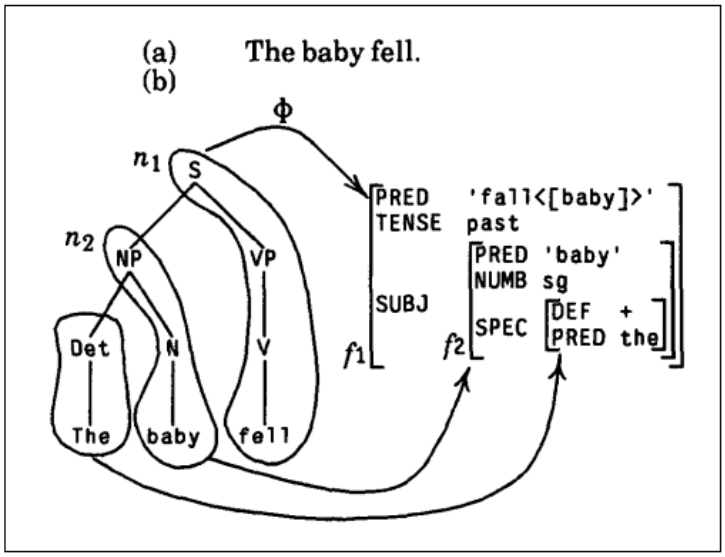
4.1.4 Baseada em Restrições

A abordagem de TA baseada em restrições (*Constraint-Based Machine Translation* - CBMT) realiza mapeamentos entre representações linguísticas, estabelecendo restrições lexicais através do princípio de gramática de função lexical (LFG - *Lexical Function Grammar*) de codescrições (KAPLAN *et al.*, 1989; SILVA *et al.*, 2007). Um grupo de funções de correspondências é especificado através de relações entre os níveis diferentes de estrutura linguística da LF e da LA.

Segundo Kaplan e Bresnan (1994 apud KAPLAN *et al.*, 1989), a gramática de função lexical é constituída por dois componentes de representação sintática que contêm informações diferenciadas do texto: o c-structure e o f-structure. O primeiro constitui um conjunto organizado das palavras e orações da sentença através do tipo formal de árvore de

estrutura de frase (*phrase structured tree*), enquanto que o segundo contém informações gramaticais através de uma função finita hierárquica (*hierarchical finite function*). A figura 10 ilustra um exemplo da frase em inglês "The baby fell" transcrita para esses dois modelos. O mais à esquerda é o c-structure e, o mais à direita, o f-structure. (KAPLAN *et al.*, 1989).

Figura 10 - Uma frase modelada em CBMT com o c-structure e o f-structure.



Fonte: Kaplan *et al.* (1989).

Na figura 10, pode-se ainda observar que existe uma relação ϕ entre a c-structure e a f-structure demonstrada através das linhas que associam uma estrutura à outra (KAPLAN *et al.*, 1989). A raiz da árvore n_1 na figura 10 é "pai" do nó n_2, podendo essa relação também ser expressa como $M(n_2) = n_1$, em que M representa a relação de filiação dos nós citados. Analisando a composição f-structure (figura 10), tem-se as informações de funções gramaticais extraídas da sentença nas equações (1) e (2). A primeira evidencia o tempo verbal "past" (passado) da sentença e a segunda uma referência ao sujeito da oração presente em f_2.

$$(f_1 TENSE) = past \tag{1}$$

$$(f_1 SUBJ) = f_2 \tag{2}$$

De acordo com Kaplan *et al.* (1989) e observando a figura 10, pode-se também constatar o nó n_1 está ligado à função f_1 através da relação ϕ, bem como o nó n_2 está conectado com a função f_2 por ϕ, conforme nas equações (3) e (4) a seguir:

$$f_1 = \phi(n_1) \tag{3}$$

e

$$f_2 = \phi(n_2). \tag{4}$$

Dessa maneira, a partir das equações (1), (2), (3) e (4), é possível derivar

$$(\phi(n_1)SUBJ) = \phi(n_2) \tag{5}$$

e, então, chegar a

$$(\phi(M(n_2))SUBJ) = \phi(n_2). \tag{6}$$

Tendo em vista a relação de filiação entre os nós da árvore (ver figura 10), o sujeito da oração presente na f-structure pode ser apontado pela a equação (6).

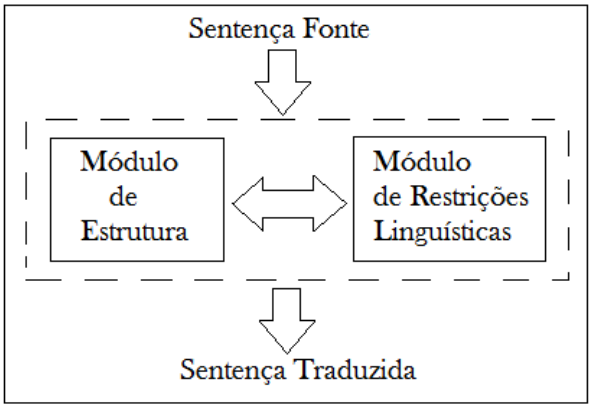
Uma das vantagens apresentada por (KAPLAN *et al.*, 1989) para o modelo LFG proposto no artigo é a de permitir tratar de forma direta características linguísticas da LF e LA, mesmo mantendo a modularidade dos blocos léxicos e gramaticais de cada uma língua. Através de Fenstad *et al.* (1987 apud KAPLAN *et al.*, 1989), é indicado que pode ser feito o encadeamento de vários grupos de estruturas como observado na figura 10 e tal relação possibilitar associações com itens semânticos.

4.1.5 Baseada em Princípios

Sistemas baseados em princípios, ou *Principle-Based Machine Translation* (PBMT), utilizam um modelo de análise baseado em parâmetros e princípios gramaticais (DORR, 1987; SILVA *et al.*, 2007). Os sistemas PBMT consistem em um conjunto de princípios, que fazem referência a fenômenos morfológicos, gramaticais e lexicais. Dessa forma, como seguem uma linha semelhante de funcionamento dos sistemas baseados em regras (RBMT), os sistemas PBMT podem ser considerados uma alternativa aos baseados em regras.

O sistema baseado em princípios tem sua arquitetura dividida em dois módulos que operam em conjunto: um módulo para a estrutura do texto e outro para as restrições linguísticas (como pode ser observado na figura 11). A essência de funcionamento desse sistema está em realizar uma alternância de controle de fluxo de execução entre os dois módulos citados até que uma estrutura pré-especificada esteja modelada por completo em uma árvore sintática instanciada (DORR, 1987).

Figura 11 - Arquitetura de um sistema PBMT.



Fonte: Adaptada de Dorr (1987).

No artigo de Dorr (1987), é apresentado um analisador sintático baseado em parâmetros e princípios gramaticais. O módulo de estrutura colabora com o módulo de restrições linguísticas, através de uma co-rotina em que o primeiro busca no segundo as restrições da teoria da Regência e Ligação (GB - *Government and Binding theory*) desenvolvida por Chomsky. Em seguida, um esqueleto sintático é definido para a sentença, para então ser excluído ou alterado segundo os princípios de GB. De acordo com Black (1999), a teoria GB define que boa parte das gramáticas das línguas existentes é composta por uma gramática universal comum a todos os idiomas. Os níveis de representação e um sistema de restrições são os dois componentes centrais desta gramática universal.

O sistema descrito se distingue de outros sistemas GB de análise sintática ou tradução, no sentido de permitir uma maior flexibilidade através de verificações "*online*" dos parâmetros dos princípios, ou seja, uma verificação e ajuste dos mesmos em tempo de processamento. Essa característica é interessante ao se trabalhar com línguas que não dispõem de uma mesma ordem de elementos na frase, como alemão e inglês (DORR, 1987).

Ao se trabalhar com o *framework* da teoria da Regência e Ligação, os princípios são modularizados de forma a generalizar as propriedades linguísticas comuns a vários idiomas. Por exemplo, normalmente a transformação para a forma passiva de uma sentença é realizada com uma regra chamada "move-α", na qual um constituinte α é realocado de posição. Esta regra é regida pelos princípios de *Trace Theory*, a ser explicada melhor adiante. Para que a regra "move-α" se mantenha aplicável de forma genérica a várias línguas, é preciso estabelecer um conjunto pequeno de regras que sejam configuradas através da mudança de parâmetros (DORR, 1987).

As teorias subjacentes de um sistema PBMT (figura 11) são expansões

correspondentes aos dois elementos definidos na figura 11: a teoria $\overline{X}$, como um subsistema do módulo de estrutura, definindo restrições sobre a ordem e posição dos constituintes de uma sentença; a teoria θ e a *Trace Theory* que estabelecem as restrições referentes ao movimento dos constituintes das frases (DORR, 1987). Esses sub-componentes fazem parte da teoria GB.

De acordo com Dorr (1987), é possível identificar duas idéias principais associadas com a teoria $\overline{X}$. Uma delas é a divisão de tipos de modelos para cada um item lexical, organizada através de um dicionário. Essa idéia pode ser melhor compreendida quando se observa o modelo do léxico *put* na frase em inglês de exemplo *"put the car in the garage"*, que possui o argumento nominal "*the car*" e o preposicional "*in the garage*". A segunda idéia relacionada à teoria $\overline{X}$ é a de que existe um léxico principal X que projeta todo o resto da estrutura da frase. O verbo *put* projeta a sentença verbal *"put the car in the garage"*, tomando-se a frase *"he put the car in the garage"* como exemplo. Nesta teoria, um dos parâmetros de variação é a ordem dos constituintes (*constituent order*), nesse parâmetro a ordem dos constituintes da língua a ser utilizada numa análise é configurada antes de começar a mesma.

Das sub-teorias do módulo de restrições linguísticas, a teoria dos papéis temáticos ou semânticos é a teoria θ. Ela assegura que somente um papel semântico (θ-role) esteja vinculado a cada um argumento de frase nominal e que cada argumento esteja somente atribuído a um *θ-role*. Um princípio de transmissão do papel semântico é utilizado para mapear *θ-roles* aos argumentos verbais na sentença através das informações específicas de um verbo no dicionário. Este princípio pode ser visto nas sentenças (7) e (8) em espanhol, em que o objeto direto do verbo *ver* é o *θ-role*. Na sentença (7), a frase nominal *el libro* é o *θ-role* atribuído referente ao verbo. Já na sentença (8), o *θ-role* está vinculado a *lo* (DORR, 1987).

$$Juan\ vio\ el\ libro. \tag{7}$$

$$Juan\ lo\ vio. \tag{8}$$

Um parâmetro de variação denominado *clitic doubling* (clítico[1] duplo) foi criado para o princípio da transmissão *θ-role*. Enquanto um clítico é um componente pronominal associado a um objeto do verbo, o clítico duplo representa um par de clítico e léxico ($< clitic, lexicalNP >$) que devem estar em concordância de gênero, número e grau. Se

[1] "Clítico é um elemento gramatical que mostra um comportamento intermediário entre um morfema e uma palavra", como por exemplo mesóclise, ênclise e próclise. Fonte: `http://pt.wikipedia.org/wiki/Cl\'itico`.

um clítico e uma frase nominal tiverem o mesmo caso, é possível, através deste parâmetro de variação, atribuir um papel semântico (*θ-role*) para o objeto nominal duplo, composto pelo clítico e pela frase (DORR, 1987).

Pode-se inferir, através de Dorr (1987), que a *Trace Theory* é outra sub-teoria GB do módulo de restrições linguísticas da arquitetura de um sistema PBMT. Esta teoria caracteriza a possibilidade de existir uma posição vazia em locais das sentenças onde *traces* são permitidos. Um *trace* é uma posição vazia natural ou gerada pelo movimento de um constituinte para outra posição da sentença. Um dos tipos de *traces* é o sujeito nulo. Dessa forma, a teoria *Trace* permite realizar uma conversão uniforme de uma sentença entre diferentes línguas, mapeando os sujeitos nulos nas línguas em que esse fenômeno é permitido, como o espanhol. Foi criado um parâmetro denominado *pro-drop* que pode informar dois dados, se permite ou não frases nominais vazias ocuparem a posição do sujeito na sentença.

$$Le\ quiere\ a\ Juan.^{2} \tag{9}$$

Um exemplo de análise sintática extraído de Dorr (1987), utilizando os princípios discutidos na presente seção, está na figura 12 que representa a análise da frase (9) em espanhol. No exemplo, é feita a expansão dos símbolos não-terminais e a finalização de ambos os símbolos terminais e não-terminais[3].

Na figura 12, em (a), o módulo de estrutura identifica que a sentença tem uma frase nominal (NP) e uma frase verbal (VP), bem como sua ordem que é determinada através do parâmetro *constituent order* verificado durante a compilação.

Nesse momento (figura 12 - (b)), o fluxo de controle é passado para o módulo de restrições que verifica o parâmetro de sujeito nulo. Para indicar que o sujeito poderá aceitar sujeito nulo ou não, no nó NP é adicionado um rótulo [+pro] (pronominal).

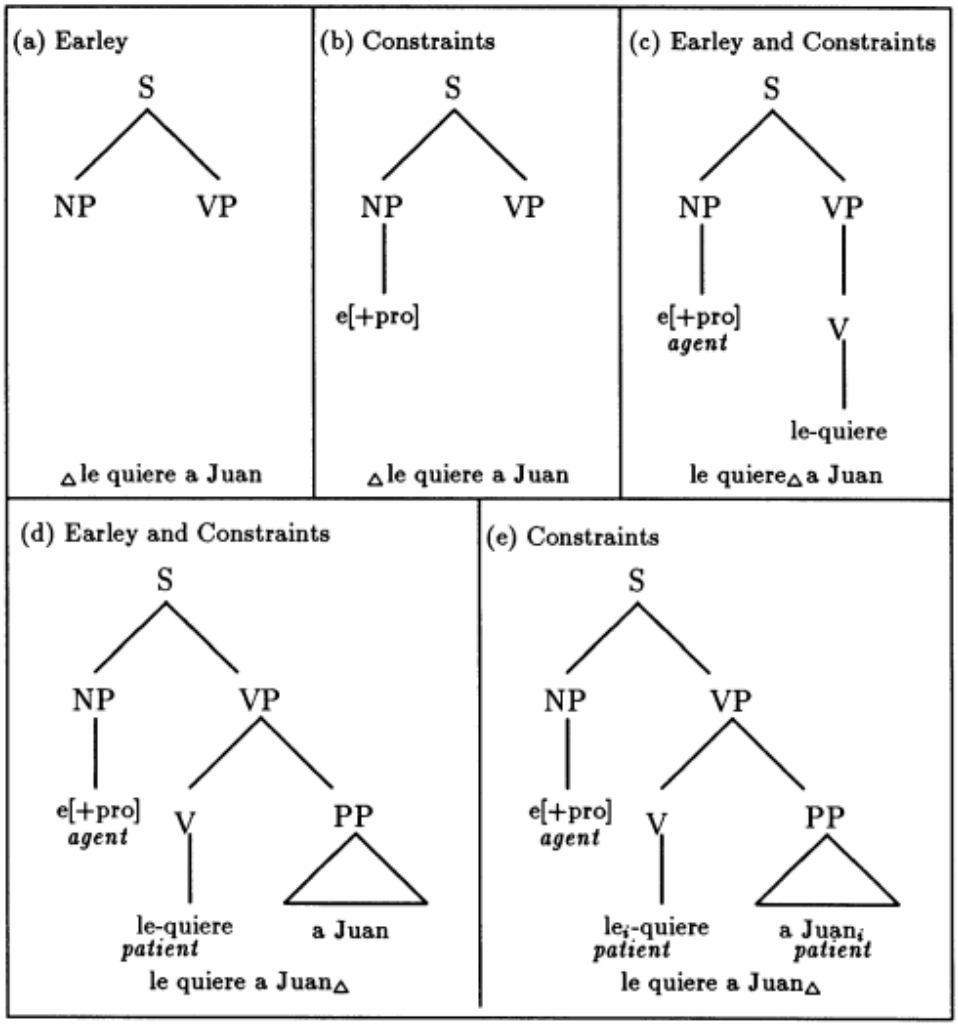
Na parte (c) da figura 12, é continuada a análise com o módulo de estrutura que lê as duas primeiras palavras e depois passa o controle para o módulo de restrições. Através da entrada do verbo *querer* do dicionário, é atribuída a informação *agent* como papel semântico ao sujeito vazio da sentença.

Em (d) na figura 12, a frase nominal *le* pode atuar como um objeto (com papel *patient*) do verbo *quiere*, através da determinação do parâmetro clítico duplo. O módulo de restrição armazena a informação sobre o clítico lido. Ao final de (d), o controle volta

[2] A tradução da sentença é: (Ela) ama João.

[3] "Símbolos terminais são símbolos que representam tokens (...). Símbolos não-terminais são aqueles que são descritos pelas regras de produção (da gramática) como combinação de símbolos terminais e não-terminais." (TAVARES, 2000)

Figura 12 - Etapas de análise sintática de um sistema PBMT.



Fonte: Dorr (1987).

ao módulo de estrutura que lê as duas últimas palavras. O próximo passo é a tentativa sem sucesso de especificar um papel para o NP *Juan* pelo módulo de restrições.

Todos os papéis associados ao verbo *querer* já foram utilizados, não sendo possível reutilizá-los nem ao menos deixar o NP sem nenhum papel, segundo o *θ-criterion*. Então, o módulo de restrições aplica um papel de *patient* ao NP, pois a regra de transmissão do (*θ-role*) é aplicável nesse caso. Dessa maneira, o NP *Juan* será co-indexado juntamente com o clítico armazenado no passo (d). Assim, é encerrada a etapa (e) (figura 12) da tradução (DORR, 1987).

O Princitran é um exemplo de sistema PBMT, baseado nos princípios sintáticos da teoria da Regência e Ligação (GB) e nos princípios LCS , elementos lexicais que carregam informações semânticas. Nesse sistema, a construção de estruturas é adiada até que as descrições satisfaçam os princípios linguísticos (DORR *et al.*, 1995).

Três características positivas podem ser destacadas dos sistemas PBMT, segundo Dorr (1987): análise uniforme independentemente da língua, tamanho de gramática reduzido e modularidade preservada. A primeira representa a capacidade de generalizar os princípios utilizados para mais de uma língua, através dos parâmetros associados aos princípios do sistema. Dessa forma, é possível extender a análise para um novo idioma, sem um trabalho complexo de adaptação da estrutura do sistema. Quanto à segunda vantagem, o sistema PBMT economiza em tempo de processamento, através da redução do tamanho da gramática em razão da existência das regras de restrições linguísticas. O esquema de "co-rotina" estabelece uma organização, dividindo as tarefas relacionadas à estrutura e às restrições linguísticas em dois grupos. A teoria GB prevê essa modularização com o fim de tornar mais simples a detecção de características da linguagem, bem como cada um dos componentes do sistema.

Uma das grandes limitações do PBMT é ser totalmente orientado à gramática. A situação de ter que processar línguas fonte e alvo estruturalmente diferentes e semanticamente similares é contornada através da inclusão dos papéis (*θ-role*). No entanto, isso deveria ser lidado de forma mais genérica (DORR, 1987).

A técnica PBMT é complementar às abordagens KBMT e EBMT, no sentido de que ela provê uma cobertura ampla para muitos fenômenos linguísticos, mas lhe falta conhecimento mais profundo sobre o domínio da tradução (SILVA *et al.*, 2007). Já os sistemas RBMT são limitados em sua cobertura linguística, devido à dificuldade de ser necessário criar muitos tipos de regras para mapeamento gramatical, lexical e semântico (DORR, 1987). Quando comparado com o RBMT, o sistema PBMT é mais flexível ao permitir uma maior extensibilidade para várias línguas e ser possível trabalhar com gramáticas de um tamanho mais reduzido que no RBMT, através da generalização linguística.

4.1.6 *Shake and Bake*

Segundo Whitelock (1992), Silva *et al.* (2007), o sistema de TA *Shake & Bake Machine Translation* (S&BMT) utiliza símbolos lexicais multidimensionais, envolvendo, portanto, uma visão lexicalista da gramática. Quando comparado com as abordagens de transferência, como a técnica de transferência e a de interlíngua, que utilizam emparelhamento de léxicos e textos, a técnica S&BMT envolve um tratamento diferenciado por não usar esse tipo de mecanismo. Ao invés de realizar operações com pares de textos, o processo de tradução ocorre através do mapeamento entre estruturas de conjuntos de léxicos de língua fonte (LF) e língua alvo (LA), chamadas de *bags* (sacolas) (TURCATO, 1995).

Através das sacolas ou multi-conjuntos de itens léxicos, a equivalência de tradução em S&BMT é realizada em duas etapas: análise do texto em LF e geração em LA. A primeira etapa constitui na análise da LF para a determinação das sacolas de léxico de um lado da equivalência da tradução. Em seguida, é feito um mapeamento através de um léxico bilíngue para selecionar as palavras do texto em língua alvo, com o objetivo de realizar o processo da combinação dos itens da sacola da LA (*shake*). O processo será concluído quando a estrutura da frase de destino estiver pronta (*bake*), conforme as restrições sintáticas da gramática da LA. As restrições semânticas necessárias são definidas a partir da relação do par de léxico bilíngue, ou seja, o compartilhamento de sinais nas duas línguas (WHITELOCK, 1992; SILVA *et al.*, 2007).

Whitelock (1992) defende que o modelo de tradução do S&BMT se baseia num conceito de gramática no qual os seus elementos mais relevantes são os léxicos. Estes carregam significado multidimensionais, com informações ortográficas, sintáticas, semânticas, dentre outras, e através da combinação de significado de vários itens lexicais se torna possível constituir a semântica de uma sentença.

Um exemplo de como funciona o processo de conversão entre um texto de LF em inglês e outro de LA em francês, pode ser acompanhado em (10), (11), (12), (13). Em (10), podem ser vistas as equivalências, como por exemplo dos léxicos Mary e Marie, na qual o termo "cite" está indicando que estas palavras possuem diferentes formas e igual valor semântico. As sentenças (11) e (12) são geradas nesse modelo tanto na tradução de ida inglês (11) - francês (12), quanto na de volta (12) - (11). As frases (11) e (12) foram produzidas da única forma que era possível dentro das restrições impostas pela gramática e associações de léxicos das sacolas em (13) (WHITELOCK, 1992).

$$\begin{gathered} X_E = X_F \\ < X_E \ \ cite > = Mary \\ < X_F \ \ cite > = Marie \\ < X_E \ \ sem \ \ index > = < X_F \ \ sem \ \ index > \end{gathered} \tag{10}$$

$$Mary \ \ loves \ \ Frances. \tag{11}$$

$$Marie \ \ aime \ \ Françoise. \tag{12}$$

$$\{Mary, Frances, love\_v, pres\} = \{Françoise, pres, aimer, Marie\} \tag{13}$$

De acordo com Silva *et al.* (2007), esta técnica é capaz de resolver casos complexos como troca de núcleo de uma sentença, bem como trabalhar em conjunto com outras abordagens. Nos últimos anos, tem sido estudada a possibilidade e desenvolvidas pesquisas nessa área, através de técnicas híbridas, como no trabalho de Carl, Rascu e Schmidt (2005), em que é construído um sistema de paráfrase utilizando a técnica RBMT e a S&BMT.

De acordo com Silva *et al.* (2007), uma das principais vantagens dessa técnica é a de ser utilizado com grande eficácia em pares de idiomas com grandes diferenças estruturais, pois ela utiliza o contraste das características das línguas como fator que a beneficia e não como fator restritivo. O fator negativo do *Shake & Bake* é a de ser um problema NP-completo[4], que vem sido contornado com tentativas de melhorar o desempenho dos algoritmos para os casos médios (WHITELOCK, 1992).

4.2 TÉCNICA EMPÍRICA

O avanço do poder de processamento dos sistemas computacionais juntamente com a crescimento da disponibilidade de dicionários legíveis ao computador, *corpora* paralela e *corpus* monolíngue impulsionaram a abordagem da técnica empírica, orientada a dados (DORR; JORDAN; BENOIT, 1998; KAUCHAK, 2006). Uma *corpora* paralela é um par de textos em duas línguas distintas que são a tradução equivalente um do outro, já o *corpus* monolíngue é um conjunto de textos de uma única língua. Nesse tipo de técnica, o mínimo de conhecimento linguístico é inserido manualmente no sistema de TA. Essas informações normalmente serão adicionadas automaticamente através de técnicas experimentais estatísticas e orientada a dados. Também, nestas técnicas existe um desafio que precisa ser resolvido, o de fazer com que os dados das extensas *corpora* e *corpus* caibam na memória, de forma a ser possível realizar o seu processamento eficientemente. Nas seções seguintes, são descritas as técnicas: Baseada em Estatística (SBMT), Baseada em Exemplos (EBMT) e Baseada em Diálogos (DBMT).

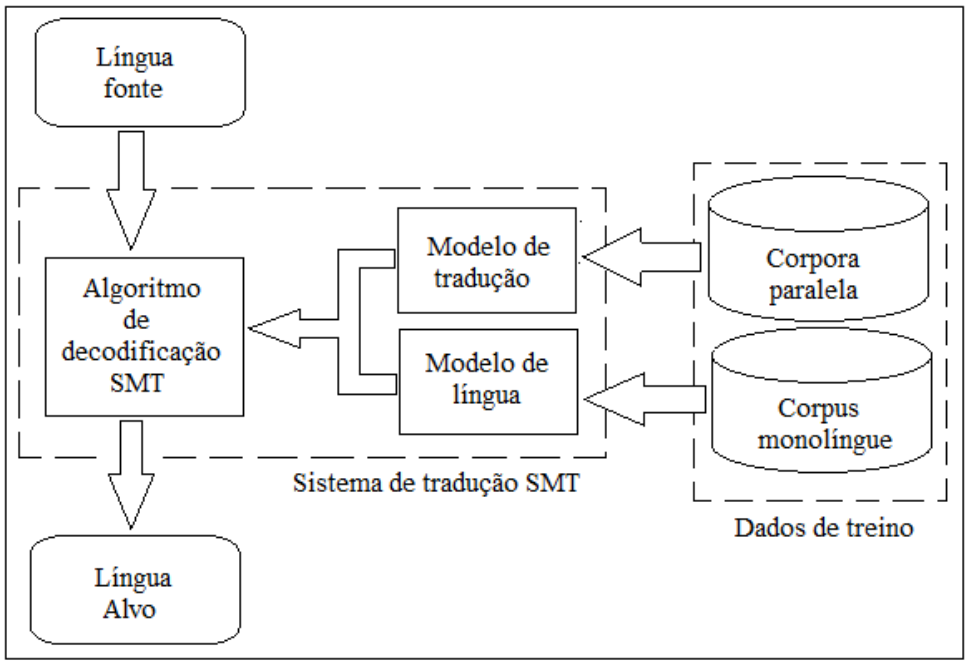
4.2.1 Baseada em Estatística

Nos projetos de tradução automática, a técnica *Statistical Based Machine Translation* (SBMT) tem ganhado bastante espaço no meio acadêmico e também na

[4]Definição de NP-Completo: "Um problema X é chamado NP-Completo se X pertence a NP, e X é NP-Árduo (NP-Difícil)" (LIMA, 2001). Definição de NP-Árduo (NP-Difícil): "Um problema X é chamado um problema NP-Árduo se todo problema NP é redutível polinomialmente a X" (LIMA, 2001). Definição de NP: "A classe de problemas que possui algoritmos não-determinísticos, cujo passo de reconhecimento pode ser realizado por um algoritmo polinomial do tamanho da entrada, é chamada de NP" (LIMA, 2001).

indústria (BRANDT; TYERS, 2011). Essa técnica é caracterizada por uma *corpora* paralela que fornece material em que será gerado e treinado o modelo da tradução para aquele par de línguas específico, a língua fonte (LF) e a língua alvo (LA). Também é necessário o *corpus* monolíngue para que seja criado o modelo da língua alvo da tradução.

Figura 13 - Arquitetura de um sistema SBMT.



Fonte: Adaptada de NTT (2007).

A arquitetura de um sistema de tradução baseado na técnica estatística pode ser observada na figura 13. Nesta é possível observar que o sistema é composto pela base de dados para o treino da ferramenta (*corpora* paralela e *corpus* monolíngue) e pelo módulo de processamento que utilizará os dados do treino sob a forma de dois modelos estatísticos, o modelo de tradução e o modelo da língua, para realizar a tradução da LF para LA.

De acordo com Silva *et al.* (2007), no SBMT alguns dados estatísticos são obtidos de forma automática da *corpora*. Os exemplos a seguir ilustram quais tipos de dados podem ser extraídos:

a) probabilidade de haver trechos de texto de LF no texto em LA;

b) probabilidade de um léxico da LF ser convertido em LA para mais de um léxico;

c) probabilidade de tradução de um léxico da LF em outro léxico da LA;

d) probabilidade de posicionamento na LA do léxico original da LF, análise feita quando há uma mudança de posição entre os léxicos correspondentes.

Dorr, Jordan e Benoit (1998) indica que a noção central por trás do SBMT pode ser observada na equação (14). Uma variação da regra de Bayes estabelece que a probabilidade de uma sentença alvo A ser a tradução de uma sentença fonte F é proporcional ao produto de outras duas probabilidades. Uma delas ($P(A)$) representa a sentença A no seu fator de fluência, que significa o nível de obediência às suas regras gramaticais. E a outra ($P(F|A)$) probabilidade é a garantia da volta segura da tradução da língua alvo para a língua fonte (fator fidelidade).

$$P(A|F) \sim P(A) * P(F|A) \tag{14}$$

De acordo com Brandt e Tyers (2011), na tradução estatística existe uma grande vantagem que consiste em não ser necessário dispor de especialistas nas línguas em questão, durante o projeto da ferramenta. O processo de tradução será baseado na maior probabilidade de que uma determinada palavra tem para se posicionar em uma determinada vizinhança no texto. Também, pode-se destacar o benefício do baixo custo de desenvolvimento para um sistema de tradução baseado em estatística, pois como sua essência está apenas na *corpora* utilizada para treinar o modelo, não é necessário implementar regras gramaticais complexas e que atendam várias particularidades das línguas.

Pluto é um exemplo de uma ferramenta de tradução *online* de patentes que utiliza a metodologia de tradução baseada em estatística (TINSLEY; WAY; SHERIDAN, 2010). O mecanismo de tradução automática é oferecido através de serviços *web*. A estrutura de Pluto inclui à ferramenta de tradução, uma memória de traduções realizadas e um mecanismo de busca de patentes.

Para a tradução nesta ferramenta, é utilizado um *framework* denominado MaTrEx (*Machine Translation Using Examples*), em que a *corpora* do sistema é treinada para produzir as traduções mais próximas do ideal, baseando-se em exemplos de texto equivalentes com os mesmos pares de línguas utilizados. O framework MaTrEx é compatível com técnicas de tradução híbridas que tenham como foco os dados. Essa flexibilidade pode ser encarada como vantagem, pois ao se trabalhar com uma língua diferente, devido à particularidade da mesma, pode ser interessante utilizar uma outra técnica de tradução (TINSLEY; WAY; SHERIDAN, 2010).

Segundo Tinsley, Way e Sheridan (2010), as seguintes abordagens podem ser utilizadas através da arquitetura híbrida citada: baseada em exemplo, estatística baseada em frases e abordagens hierárquicas para tradução. O MaTrEx pode ser utilizado em conjunto com outras ferramentas, como o Moses e o Giza++. Recentemente, o

Pluto ficou na liderança da lista das melhores ferramentas para tradução dos pares de língua inglês-francês, inglês-espanhol e inglês-chinês no *Workshop on Statistical Machine Translation* em 2009 (WMT-09) (TINSLEY; WAY; SHERIDAN, 2010).

A obrigatoriedade de ter disponível uma *corpora* para a aplicação da SBMT, é um dos grandes problemas desta técnica. Pelo fato da construção do modelo de treinamento se basear nos dados contidos na *corpora*, o sistema de tradução implementado dependerá bastante do domínio dos textos utilizados como *corpora*. Assim, normalmente a ferramenta de TA estatística pode ser projetada para ser especialista em traduções de determinados contextos. Além destas características, caso se deseje melhorar a qualidade final da tradução, pode-se manipular os parâmetros referentes às línguas fonte e alvo nos modelos probabilísticos. Uma forma de unir os benefícios de diferentes técnicas é utilizando-as em conjunto, através das técnicas híbridas (SILVA *et al.*, 2007).

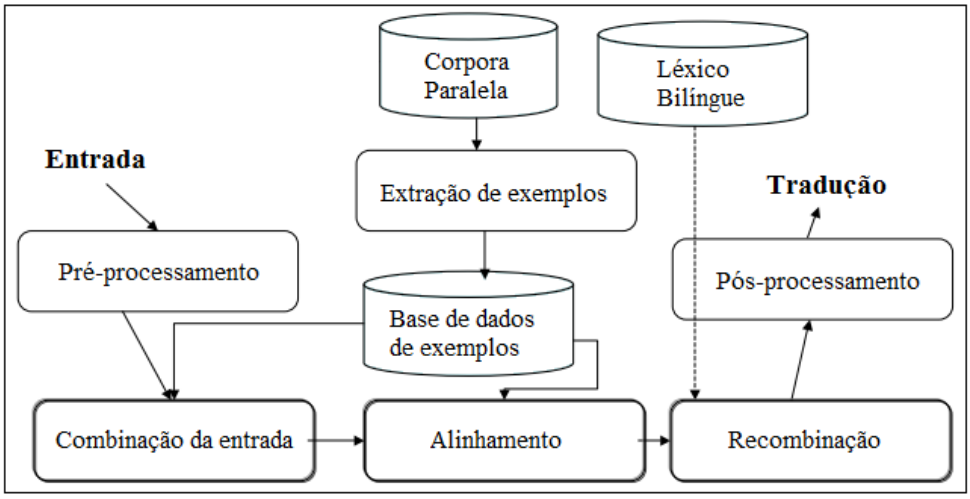
4.2.2 Baseada em Exemplos

A técnica *Example-Based Machine Translation* (EBMT), de acordo com Hutchins (2005), pode ser entendida como uma técnica semelhante à SBMT. Ambas utilizam uma *corpora* de textos paralelos, mas a EBMT realiza a tradução através de exemplos de traduções similares encontradas na base de dados, diferentemente da SBMT que seleciona a tradução de maior probabilidade na *corpora* para a mesma sequência de palavras da língua fonte.

A tradução é realizada a partir da extração de trechos de textos da língua fonte (LF) através de um *corpus* e a criação de textos na língua alvo (LA) com o mesmo significado. Para a criação dos trechos de texto em LA, primeiramente uma busca na base de dados é feita para se encontrar um texto em LF que seja equivalente ao texto que se deseja efetuar a tradução, o texto de entrada. Então, a *corpora* paralela de trechos de textos alinhados permite que o trecho equivalente em LA seja determinado. Em uma última etapa, é feita a recombinação de strings desse trecho em LA para que seja obtida a melhor tradução, respeitando-se as regras da LA (HUTCHINS, 2005). A figura 14 ilustra a arquitetura de um sistema baseado em exemplos.

Hutchins (2005) afirma que não parece existir claramente um modelo básico da EBMT, mas que existe uma variedade de técnicas extraídas de outras abordagens que são aplicadas neste modelo de tradução baseado em exemplos, como algumas técnicas de memória de tradução, de RBMT e SBMT. Uma classificação em duas frentes principais podem ser extraídas dentre os modelos de tradução EBMT, a tradução EBMT baseada em *templates* que faz uso de verificações da *corpora* paralela em tempo de "execução", e a tradução EBMT baseada em dados pré-compilados, como árvores de derivação, que faz

Figura 14 - Arquitetura de um sistema EBMT.



Fonte: Adaptada de Vertan (2004).

síntese do conhecimento da *corpora* previamente em tempo de "compilação". Esses dois tipos de EBMT se diferenciam pelo quanto o processo principal da tradução se aproxima da abordagem baseada em regras (RBMT) e pelo quanto se aproxima da abordagem estatística. A abordagem baseada em regras tem um prévio conhecimento das estruturas de formação da LF e LA, através de estruturas como as árvores de derivação, já a abordagem estatística avalia em tempo de execução quais palavras serão combinadas para formar o texto em língua alvo que possui uma maior probabilidade de ser a tradução do texto em LF.

Silva *et al.* (2007) descreve que a técnica EBMT faz uma combinação entre o texto a ser traduzido com exemplos de traduções armazenados através de uma *corpora* paralela. O conceito de EBMT sustentado pela comunidade de pesquisa atualmente, segundo Hutchins (2005), é que esta técnica tem seu núcleo de processamento baseado em analogia, ou seja, um sistema de tradução EBMT analisa o texto de entrada para que seja encontrado um texto similar existente na base de dados. Nas abordagens baseadas em regras e baseadas em estatística, a busca é por uma equivalência exata de palavras e textos a serem traduzidos, enquanto que no EBMT, a busca é por similaridade entre duas strings. Para determinar a similaridade entre dois trechos são utilizadas enciclopédias e estruturas de representação onde a extração pode ser feita a partir de equivalências parciais.

Segundo Silva *et al.* (2007), a determinação da distância semântica entre as palavras é o grau de proximidade que determinará qual a melhor escolha para a tradução de um texto específico na EBMT. Esse processo consiste na avaliação de dicionários e ontologias

que provêm estruturas de hierarquias de conceitos.

Um dos argumentos favoráveis ao uso da técnica EBMT é a sua capacidade de fácil expansão pela inserção de novos exemplos à base de dados (HUTCHINS, 2005). No entanto, em casos de ferramentas EBMT mais complexas, a expansão não é tão fácil, por ser necessário realizar uma análise substancial de dados, modificação da estrutura de análise sintática, inserção de parâmetros, dentre outros. Em alguns casos, a expansão é vista como uma diminuição da qualidade da *corpora*, pois podem se tratar de exemplos de tradução pouco comuns, que impedirão a reusabilidade da *corpora* e conduzirão a uma tradução errônea.

De acordo com Hutchins (2005), a complexidade da tradução EBMT pode ser reduzida em aplicações com linguagens restritas, como textos técnicos. A restrição do domínio definirá um contexto específico na qual a base de dados será especialmente preparada, o que significará redução na variação de linguagem utilizada e suportada pelo sistema. Por outro lado, segundo Silva *et al.* (2007), o tratamento de um número grande de divergências sintáticas e semânticas não será sinônimo de um aumento de desempenho do sistema EBMT, pois isso resultará em uma maior *corpora*, o que requererá um maior tempo de processamento de buscas e comparação de strings.

4.2.3 Baseada em Diálogo

A técnica que se baseia em um formato de tradução através de diálogos com o usuário é denominada de *Dialogue-Based Machine Translation* (DBMT). O objetivo das interações com o usuário é desambiguar o texto, a partir de informações específicas de contexto e cultura, partindo do pressuposto que o usuário é normalmente o autor do texto a ser traduzido (DORR; JORDAN; BENOIT, 1998).

De acordo com Dorr, Jordan e Benoit (1998), a interação com o usuário poderá ocorrer em dois momentos distintos. Poderá ocorrer durante o processo de tradução, em que a interação consistirá de um mecanismo de desambiguação *online* guiada pelo usuário, ou poderá ocorrer em um momento anterior à tradução, em que a interação com o usuário revisará o texto para um formato que o tradutor suporte.

A abordagem DBMT surgiu como uma alternativa à técnica baseada em conhecimento e à baseada em princípios linguísticos, quando uma base de ontologias e especialistas nos pares de idiomas não estão disponíveis ou estão fora do orçamento para o projeto de um tradutor DBMT (BOITET; BLANCHON, 1995).

O projeto LIDIA é um exemplo de ferramenta que utiliza a técnica de tradução automática DBMT, com o objetivo de realizar a tradução de um texto para diversos outros idiomas, com o auxílio de seu autor com conhecimento apenas do idioma fonte e

do contexto do texto (BLANCHON, 1994). Neste projeto, o diálogo com o usuário está associado aos *story cards* e *treatment cards*.

Um *story card* é um conjunto de duas ou mais estórias que contém uma sentença ambígua. Um exemplo de um *story card* pode ser observado nas sentenças (15) e (16) em inglês[5].

Left Story: "*From China, the captain has bring back a vase. This vase is English*". (15)

Right Story: "*The captain has bring back a Chinese vase. His boat is soiled*". (16)

Cada *story card* tem um *treatment card*, no qual o usuário irá responder algumas perguntas com o objetivo de desambiguar e compor o contexto da sentença. Quando o sistema detecta que uma sentença deve ser desambiguada, o usuário é notificado e deverá responder uma pergunta objetiva, selecionando o contexto mais apropriado ao texto. O sistema LIDIA exibe todas as tarefas que o usuário deverá realizar até finalizar a sua revisão do texto para que o mesmo possa ser finalmente traduzido (BLANCHON, 1994). Um exemplo da tela de interação com o usuário para desambiguação do termo em francês *capitaine*[6] no sistema LIDIA, pode ser visto na figura 15.

A técnica DBMT, bem como o baseado em conhecimento (KBMT) e em exemplos (EBMT), tem uma aplicação mais eficiente em domínios restritos, mas não tão restrito quanto em KBMT e EBMT (DORR; JORDAN; BENOIT, 1998; SILVA *et al.*, 2007; BOITET; BLANCHON, 1995). Quando é utilizado um contexto mais abrangente, uma maior dificuldade é enfrentada ao se processar buscas e realizar o armazenamento de uma maior quantidade de informações.

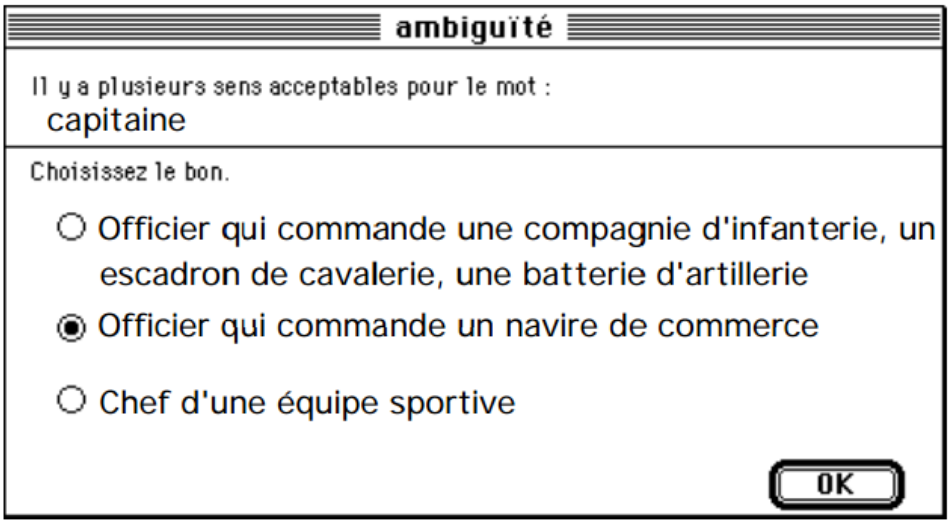
4.3 TÉCNICAS HÍBRIDAS

Para Silva *et al.* (2007), muitas técnicas apresentam dificuldades para manipular alguns parâmetros do processo de TA, especialmente os empíricos. Diferentemente dos sistemas baseados em exemplos, os sistemas de tradução baseados em estatística não conseguem manter dependências conceituais de longa distância em trechos de tradução complexos. Para evitar as dificuldades enfrentadas individualmente por algumas técnicas,

[5] As traduções das sentenças: Estória esquerda - "Da China, o capitão trouxe de volta um vaso. O vaso é inglês"; Estória direita - "O capitão trouxe de volta um vaso chinês. Seu barco está sujo".

[6] Tradução do Google Translate: capitão. Fonte: `http://translate.google.com`.

Figura 15 - Tela de desambiguação do sistema LIDIA - um sistema DBMT.



Fonte: Blanchon (1994).

pode-se combiná-las em um só sistema.

Normalmente, um sistema de tradução com técnica híbrida é modelado com uma técnica empírica e outra linguística. A proporção sobre a predominância de cada uma das técnicas no sistema projetado é determinada pela especificação do resultado desejado. Caso se deseje uma maior robustez e cobertura de dados, há um foco maior na técnica estatística. Para o caso de se desejar cobrir mais detalhes específicos das línguas em questão, a técnica linguística é priorizada na maioria das regras de conversão do texto (SILVA *et al.*, 2007).

5 ANÁLISE COMPARATIVA DAS TÉCNICAS DE TRADUÇÃO AUTOMÁTICA

As técnicas estudadas apresentam algumas características, como vantagens, desvantagens, recursos necessários para sua utilização e área em que melhor se aplica. Serão destacados esses pontos num resumo geral das técnicas estudadas. Em seguida, será apresentado uma sistematização de comparação entre as técnicas. Por fim, alguns exemplos de aplicação de três técnicas selecionadas serão descritos.

5.1 RESUMO DAS CARACTERÍSTICAS DAS TÉCNICAS

A técnica baseada em regras (RBMT) tem seu processo de tradução baseado na transferência do texto em LF (Língua Fonte) para LA (Língua Alvo), através de regras que representam o conhecimento em diferentes níveis linguísticos (técnicas direta, por transferência e interlíngua) (HARSHAWARDHAN, 2011). Esta técnica tem como vantagem um bom desempenho e qualidade da tradução, capacidade de ser aplicada em contextos de propósito geral e em casos de línguas sem recursos de processamento linguístico (por exemplo, *corpora* paralela) como o português-Br. As limitações associadas a esta técnica são o fato de ser necessário criar manualmente as regras gramaticais do sistema, o que resulta em uma dificuldade de manutenção quando for preciso realizar a atualização das regras. Os recursos envolvidos na criação de um tradutor automático com RBMT são as próprias regras de transferência lexicais, gramaticais e semânticas.

A técnica baseada em conhecimento (KBMT) se concentra em associar conceitos mais profundos a um léxico através de modelos de domínio ou ontologias, como informações morfológicas, sintáticas e semânticas, sem no entanto estabelecer uma relação desse conhecimento à sintaxe (DORR; JORDAN; BENOIT, 1998). As vantagens associadas a esta técnica são uma tradução de mais qualidade ao se encarar a ontologia como uma interlíngua; a fácil manutenção, por ter a base de dados da tradução representada por uma ontologia, sendo facilmente atualizada; a modularidade, por ser possível substituir a base, sem muitas alterações no sistema. As limitações do KBMT são o alto custo envolvido na criação das bases de conhecimento, a restrição a um domínio específico a que ficam sujeitas e a dificuldade de se delimitar um domínio (abrangência e aprofundamento). Os recursos necessários à utilização desta técnica são uma ontologia de conceitos no contexto desejado de tradução e regras de mapeamento lexical e gramatical.

A técnica baseada em léxico (LBMT) combina árvores de derivações LTAG (*Lexicalized Tree Adjoining Grammars*) fonte e alvo através de um léxico de transferência

(ABEILLÉ; JOSHI; SCHABES, 1990). A vantagem que pode ser destacada é uma correspondência estável entre grandes estruturas gramaticais de LF e LA. A dificuldade de utilização dessa técnica é criar as árvores LTAGs fonte e alvo manualmente, bem como as regras de mapeamento dos léxicos entre os nós dessa estrutura e, por consequência a sua manutenção quando houver a necessidade de modificá-las. Os recursos necessários para aplicação dessa técnica são as LTAGs e as regras de mapeamento lexicais, gramaticais e semânticas entre as LTAGs fonte e alvo.

A técnica baseada em restrições (CBMT) realiza o mapeamento entre representações linguísticas, através da gramática de função lexical de codescrições (LFG) (KAPLAN *et al.*, 1989). Suas vantagens são tratar diretamente as características da língua fonte e da língua alvo e possibilitar o encadeamento de várias estruturas de LFG para alcançar associações semânticas entre itens de LF e LA. A sua principal limitação é ser preciso criar manualmente as gramáticas de função lexical, com as regras de mapeamento entre uma LFG fonte a alvo. Os recursos essenciais para utilização do CBMT são as LFGs, constituídas pelo *c-structure* (árvore que representa a estrutura da sentença) e *f-structure* (contém informações gramaticais das palavras da sentença).

A técnica baseada em princípios (PBMT) utiliza um modelo de análise baseado em um conjunto de princípios que fazem referência a fenômenos morfológicos, lexicais e gramaticais (SILVA *et al.*, 2007). Suas vantagens incluem uma análise uniforme independente da língua; tamanho da gramática reduzido, devido à possibilidade de generalização das regras gramaticais de mais de uma língua; modularidade preservada. Suas limitações são a necessidade de se criar manualmente as regras de restrições linguísticas para composição dos princípios utilizados nessa técnica. Os recursos requeridos para sua aplicação são a criação de um módulo de análise de estrutura e outro de restrições linguísticas, de acordo com a Teoria da Regência e Ligação - Chomsky.

A técnica S&BMT (*Shake & Bake Machine Translation*) realiza o mapeamento entre estruturas de conjuntos não-ordenados de léxicos de língua fonte e língua alvo, as sacolas (*bags*) (TURCATO, 1995). Seus pontos positivos são a capacidade de lidar com casos complexos, como a troca de núcleo de uma sentença, e ser eficaz nos casos de idiomas com características bem diferentes. Suas desvantagens incluem ser um problema NP-completo e ser preciso criar manualmente as regras de mapeamento lexical e gramatical. Os recursos necessários para sua aplicação são um léxico bilíngue (língua fonte - língua alvo) e um conjunto de regras de restrições gramaticais para formação da sentença de LA.

A técnica baseada em estatística (SBMT) obtém de forma automática dados

estatísticos de um par de textos que são a tradução equivalente um do outro, a *corpora* paralela, para realizar a tradução (SILVA *et al.*, 2007). Suas vantagens são a existência de uma literatura vasta de artigos e livros de referência, possuir um baixo custo e rápido desenvolvimento, o que a faz se destacar dentre as outras técnicas. Seus pontos negativos são a restrição do tradutor ao domínio da *corpora* utilizada, a dificuldade em lidar com características específicas de um idioma, suas particularidades. O recurso essencial para sua aplicação é uma *corpora* paralela extensa para que seja possível a realização do treinamento do tradutor para uma grande variedade de exemplos de sentenças nos idiomas escolhidos.

A técnica baseada em exemplos (EBMT) realiza a tradução através da combinação e análise de pequenos trechos de textos de LF, extraídos de uma *corpora* paralela, similares ao texto que se deseja traduzir, para geração de textos em LA (HUTCHINS, 2005). Seus pontos positivos são a grande aplicabilidade em situações com linguagens de contexto restrito e, para sistemas com bases de dados pequenas, a facilidade de adição de novos exemplos. Os pontos negativos ao se trabalhar com esta técnica é que se tem um limiar bem frágil de até que ponto deve-se expandir a *corpora*, pois em muitos casos obtém-se uma base de dados com exemplos de expressões incomuns, o que deixa a *corpora* com pouca reusabilidade. O recurso utilizado na construção de um sistema EBMT é uma *corpora* paralela composta de exemplos variados no contexto de tradução específico da ferramenta.

A técnica baseada em diálogo (DBMT) traduz um texto tendo como base informações passadas por um usuário através de diálogos com o sistema (DORR; JORDAN; BENOIT, 1998). Sua maior vantagem é a possibilidade de ser aplicado em situações nas quais não estão disponíveis uma base de ontologias, um especialista linguístico, ou ambos se encontram fora do orçamento do projeto. Sua desvantagem está mais associada ao fato de que seu processo principal de execução depende da intervenção de um usuário, fornecendo dados sobre contexto e desambiguação, atividades que remetem aos primeiros sistemas de tradução auxiliados por humanos e não totalmente automáticos. Os recursos necessários à implementação desse sistema são regras de mapeamento da língua fonte para a língua alvo e plataforma de interação com o usuário para desambiguação.

5.2 SISTEMATIZAÇÃO DA ANÁLISE

O quadro 2 apresenta um resumo comparativo das técnicas apresentadas criado a partir dos critérios de disponibilidade de literatura, amadurecimento da tecnologia,

complexidade de implementação, tempo de implementação, custo de implementação, manutenabilidade e aplicabilidade no foco específico desse trabalho, a tradução automática de textos técnicos. Os critérios foram escolhidos empiricamente, através das características mais documentadas na literatura consultada.

Os critérios de disponibilidade de literatura e amadurecimento da tecnologia foram definidos com base no quadro 1, que mostra a distribuição da quantidade de artigos por técnica e por período. Este quadro foi construído a partir das informações contidas no repositório *online Machine Translation Archive* (HUTCHINS, 2013) mantido pelo pesquisador John Hutchins.

O critério de disponibilidade de literatura teve como base a informação de quantidade disponível de artigos publicados sobre cada técnica, através do quadro 1. Dessa forma, foram criados de três graus de disponibilidade de literatura para as técnicas no quadro 2: baixa, para até 100 artigos; média, de 101 a 500 artigos; e alta, acima de 500 artigos.

O critério de amadurecimento da tecnologia foi analisado com base no período de maior concentração da quantidade de artigos publicados de cada técnica, também com base no quadro 1. Assim, foi possível definir que a técnica classificada como de atual amadurecimento tem sua concentração de artigos se der do ano 2001 a 2013; de médio amadurecimento, se a concentração for de 1991 a 2000; e de antigo amadurecimento, se a concentração de artigos for anterior a 1990.

O critério de complexidade de implementação foi analisado com base na complexidade estimada de se desenvolver as estruturas mínimas necessárias à aplicação de cada uma das técnicas. Para a necessidade de criação manual das regras de mapeamento lexicais, gramaticais e semânticas, foi atribuída uma complexidade de implementação alta; para a necessidade de criação de estruturas híbridas de mapeamento lexical, gramatical e semântica, juntamente com outras estruturas como as sacolas do *Shake & Bake*, o *c-structure* e o *f-structure* da técnica baseada em restrições (CBMT), foi atribuída uma complexidade média; para a necessidade de apenas implementar tradutores com bases de dados e frameworks já disponíveis, como o baseado em estatística, conhecimento e exemplos, foi atribuída uma complexidade baixa de implementação.

O critério custo de implementação foi criado com base no tempo de implementação e na necessidade de contratação de especialistas de linguística nos idiomas envolvidos. Para a técnica em que houver a necessidade de contratação de um especialista e for necessário criar todas as regras lexicais, sintáticas e semânticas manualmente, o custo de implementação será atribuído como alto; para a técnica em que não houver a necessidade

Quadro 1 - Distribuição de artigos por técnicas e por período.

Ano	RBMT	KBMT	LBMT	CBMT	PBMT	S&BMT	SBMT	EBMT	DBMT
1951 - 1960	43	0	0	0	0	0	0	0	0
1961 - 1970	29	0	0	0	0	0	0	0	2
1971 - 1980	20	0	0	0	0	0	0	0	9
1981 - 1990	169	21	2	1	2	0	4	4	30
1991 - 2000	189	57	12	2	1	3	69	109	47
2001 - 2010	217	3	1	4	0	0	1068	178	57
2011 - 2013	68	2	4	2	0	0	432	16	20
Total	735	83	19	9	3	3	1573	307	165

Fonte: Próprio Autor.

de contratação de um especialista, mas que houver a necessidade de criação manual de algumas regras mais simples de mapeamento, o custo de implementação será atribuído como médio; para as outras técnicas em que não seja preciso criar regras de restrições e mapeamento, o custo será atribuído como baixo.

O critério de manutenção foi estabelecido para destacar as técnicas que oferecem uma maior dificuldade de atualização da ferramenta. Para as técnicas em que é necessário realizar atualização de forma totalmente manual, é atribuído um custo de manutenção alto; para os que dispõem de partes modularizadas e partes de atualização manual, é atribuído um custo de manutenção médio; para os que dispõem de um bom grau de modularidade ou flexibilidade, é atribuído um custo de manutenção baixo.

O critério de desempenho foi criado com o intuito de destacar como cada técnica equilibra o tempo de processamento com a satisfatoriedade dos resultados. Foi atribuído um baixo desempenho para as técnicas que possuem alguma interação com o usuário e uma complexidade de execução alta; foi atribuído um médio desempenho para técnicas que necessitam realizar buscas em largas bases de dados, apesar do processo ser totalmente automatizado; e foi atribuído um desempenho alto às técnicas que são personalizadas e desenvolvidas sob medidas com regras de mapeamento específicas.

Analisando o quadro 2 com base nos critérios de disponibilidade de literatura, desempenho e amadurecimento da tecnologia, a técnica que se destaca é a técnica baseada em regras (RBMT). No entanto, se forem observados os critérios de disponibilidade de literatura em conjunto com complexidade de implementação, custo de implementação e manutenção, as técnicas que mais se destacam são, primeiramente, a baseada em estatística (SBMT), seguida por baseada em exemplos (EBMT) e baseada em conhecimento (KBMT). Dessa forma, foram selecionadas para ilustração de aplicação as técnicas RBMT, KBMT e SBMT. A técnica EBMT não foi selecionada por apresentar forte similaridade com a técnica SBMT.

5.3 EXEMPLOS DE TRADUTORES

Foram selecionadas como exemplo três tradutores para ilustrar cada uma das três técnicas escolhidas. Os tradutores selecionados são o Moses que utiliza a técnica baseada em estatística (SBMT), o Apertium que utiliza a técnica baseada em regras (RBMT) e o Jena que utiliza a técnica baseada em conhecimento (KBMT). A escolha das três ferramentas foi feita com base na literatura e documentação disponíveis.

Quadro 2 - Resumo comparativo das técnica de Tradução Automática estudadas.

Critérios de Comparação	**RBMT**	**KBMT**	**LBMT**	**CBMT**	**PBMT**	**S&BMT**	**SBMT**	**EBMT**	**DBMT**
Disponibilidade de literatura	média	baixa	baixa	baixa	baixa	baixa	alta	média	média
Amadurecimento da Tecnologia	atual	médio	médio	atual	antigo	médio	atual	atual	médio
Complexidade de implementação	alta	média	alta	alta	média	média	baixa	baixa	média
Custo de implementação	alto	baixo	alto	alto	médio	médio	baixo	baixo	médio
Custo de manutenção	alto	baixo	alto	alto	médio	médio	baixo	baixo	médio
Desempenho	alto	médio	alto	alto	alto	baixo	médio	médio	baixo

Fonte: Próprio Autor.

5.3.1 Exemplo de Tradutor Baseado em Estatística

Nesta seção, será apresentado um exemplo de como projetar um sistema que utilize a técnica baseada em estatística. Inicialmente, será descrito um passo-a-passo da construção de uma *corpora* paralela de textos técnicos com base no trabalho de Aziz e Specia (2011) e, posteriormente, será visto o processo de construção de um tradutor utilizando a ferramenta Moses, tendo como referência a *translation task* (tarefa de tradução) do evento *Workshop on Statistical Machine Translation*[1].

5.3.1.1 *Corpora* Paralela Técnica Inglês-Português-Br

Pode-se iniciar a tarefa de confecção de uma *corpora* paralela, a partir da busca de dados disponíveis publicamente na *web*. Esses dados precisam ser devidamente alinhados em nível de palavras, frases e documentos para que possam então ser utilizados como recurso de dicionários de tradução e técnicas de tradução automática estatística (SBMT).

A metodologia utilizada por Aziz e Specia (2011) para a construção da *corpora* é dividida nas etapas de rastreamento de dados, alinhamento de documento e de sentença. A base de dados de textos técnicos utilizada para se construir a *corpora* paralela inglês-português-Br de textos técnicos foi a revista científica brasileira "Pesquisa FAPESP *Online*" (AZIZ; SPECIA, 2011)

Para a fase de rastreamento de dados, o GNU wget[2] e um template de URL podem ser utilizados para fazer o *download* de páginas HTML do site da revista (AZIZ; SPECIA, 2011). Existe uma página de *index* com um *link* para cada um dos artigos de um determinado tema, dessa forma primeiramente é escaneado o *index* e recuperado seus *links* para, então, ser feito o *download* dos artigos propriamente ditos. No entanto, não há uma correspondência direta entre os identificadores de artigos, logo técnicas de alinhamento de documento baseadas em conteúdo precisam ser aplicadas para estabelecer uma relação entre os artigos traduzidos.

De acordo com Aziz e Specia (2011), o alinhamento de documentos pode ser feito segundo a técnica explicitada nos passos:

a) os pares de artigos inglês e português são capturados, todas as letras dos textos são convertidos para letras minúsculas e o texto é dividido em tokens;

b) as palavras originais de cada documento em português são traduzidas para suas n-melhores traduções de acordo com um dicionário bilíngue;

[1] `http://www.statmt.org/wmt11/`
[2] http://www.gnu.org/software/wget/

c) é criado um vetor com cada palavra pertencente ao documento em questão, bem como sua frequência de ocorrência no mesmo, dessa maneira o documento passa a ser representado por um perfil distribuído DP (*Distributional Profiles*);

d) calcula-se a similaridade do cosseno dos DPs para cada par de documentos em análise;

e) Por fim, é encontrado o melhor documento equivalente em português para um dado documento em inglês, quando o valor da similaridade é dado acima de um determinado valor estabelecido.

Para o alinhamento automático das sentenças, pode ser utilizada a técnica TCA (*Translation Corpus Aligner*) (AZIZ; SPECIA, 2011). O alinhamento das sentenças é feito entre documentos já alinhados. No entanto, nos casos em que não se estabelecer uma relação de uma sentença da língua fonte para uma sentença da língua alvo, será necessário aplicar algumas técnicas de correção. Neste caso, podem ser aplicados dois tipos de heurística:

a) h1, esta heurística é aplicada em casos em que há uma exclusão de um termo seguida por uma inserção de outro, ou o contrário, durante o processo de alinhamento dos textos para tradução. De acordo com h1, como os trechos alinhados sempre deveriam consistir em uma relação de 1-1, esta situação é consequência de uma decisão incorreta do alinhador automático;

b) h2, esta outra heurística é aplicada nos casos em que o alinhador automático divide uma sentença ao meio ou toma alguma decisão incorretamente, ao tentar alinhá-los. Estas decisões podem ter sido desencadeadas por uma vírgula ou ponto continuativo, em que a depender da língua representam trechos de textos que deveriam estar juntos ou separados no alinhamento.

O resultado esperado dos procedimentos indicados é a obtenção de uma *corpora* paralela técnica, que poderá ser aplicada em um tradutor estatístico descrito na próxima seção. De acordo com os experimentos realizados por Aziz e Specia (2011), através da base da revista científica brasileira "Pesquisa FAPESP *Online*" podem ser obtidas 150.000 sentenças alinhadas. Diante da ausência de *corporas* paralelas técnicas para o par inglês-português-Br, isso já pode ser considerado um avanço. No entanto, quando comparada com a *corpora* paralela *Europarl* (KOEHN, 2011a) para o par de línguas português-inglês que tem 1.960.407 sentenças alinhadas, a *corpora* obtida é pequena.

Quanto maior for a *corpora* aplicada a um tradutor estatístico, mais satisfatório poderá ser o resultado obtido com as traduções, principalmente se a *corpora* tiver uma dimensão semelhante ou maior que a da *Europarl*.

5.3.1.2 Tradutor Estatístico com a Ferramenta *Moses*

Após obter a *corpora* paralela, pode-se passar para a etapa de desenvolvimento do tradutor estatístico. Primeiramente, deve ser preparado o ambiente de desenvolvimento com a instalação das ferramentas necessárias: a ferramenta de tradução estatística *Moses*[3] (KOEHN, 2013); e a biblioteca SRILM[4] para criação e aplicação do modelo de linguagem estatística (SRI, 2011); a biblioteca GIZA++[5] que realiza o treinamento do modelo de tradução estatístico da *corpora* bilíngue (GIZA-PP, 2013). As ferramentas citadas são *opensource*, o que significa uma redução de custo no orçamento do tradutor, pois não é necessário pagar por uma licença para utilização das mesmas.

De acordo com Koehn (2011a), o processo de construção de um tradutor estatístico básico com a ferramenta *Moses* é dividido nas etapas: preparação dos dados, construção do modelo da linguagem, treinamento do modelo, otimização, execução do conjunto de testes e avaliação. Na figura 16, pode-se observar a aplicação dos *scripts* de tokenização (*tokenizer*), conversão para minúscula (*lowercase*), SRILM e *Moses*, no processo de tradução estatística.

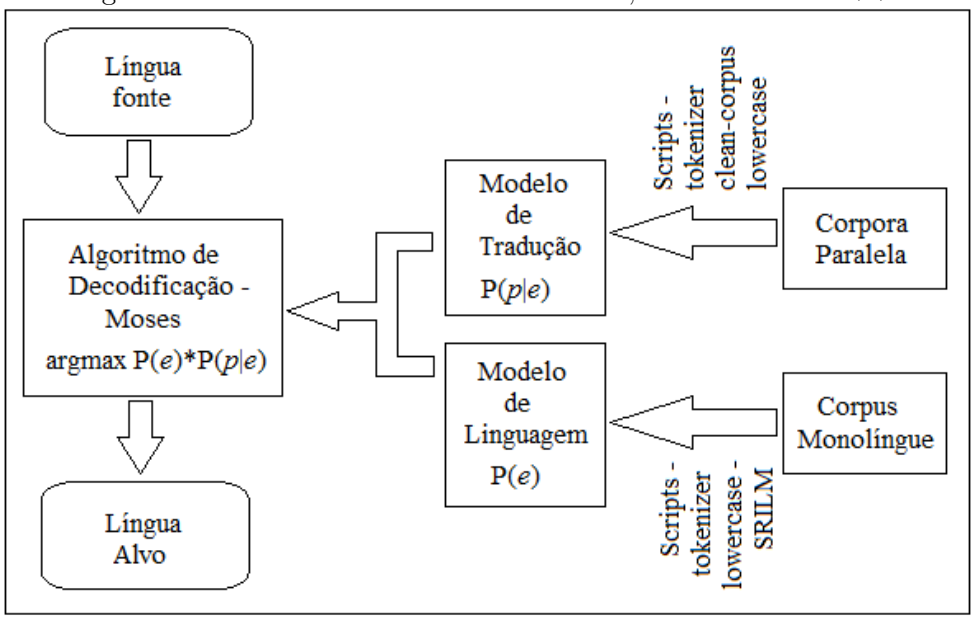
Na etapa de preparação dos dados, são utilizados *scripts*[6] para realizar a tokenização da *corpora* (*tokenizer.perl*), separação do texto em frases (*tokens*), conversão de todas as letras para o formato minúsculo (*lowercase.perl*) e exclusão de frases acima de um tamanho especificado (*clean-corpus-n.perl*) (KOEHN, 2011a).

Antes da etapa de construção do modelo da linguagem, são utilizados os mesmos *scripts* da etapa anterior para tokenizar e converter os dados monolíngues da *corpora* para letras minúsculas. Então, é utilizada a ferramenta SRILM para criar o modelo da linguagem que será utilizado no treinamento do tradutor estatístico (KOEHN, 2011a). O modelo da linguagem é responsável por atribuir probabilidades às sequências de n palavras (frase) que ocorrerem na base de dados monolíngue da *corpora*. A ferramenta SRILM permite escolher o tamanho da sequência de *n-grams* (sequência de n palavras) que será utilizada para construir o modelo.

O treinamento do modelo de linguagem criado na etapa anterior é feito através do *Moses*, utilizando os dados pré-processados da *corpora* paralela na primeira etapa

[3] `http://www.statmt.org/moses/`
[4] `http://www.speech.sri.com/projects/srilm/`
[5] `https://code.google.com/p/giza-pp/`
[6] `http://www.statmt.org/wmt09/scripts.tgz`

Figura 16 - Tradutor estatístico com *Moses*, *SRILM* e *GIZA++*.



Fonte: Adaptada de NTT (2007).

(KOEHN, 2011a).

Na etapa de otimização, são tokenizados e convertidos para minúsculo conjuntos de dados específicos de entrada e de referência de tradução do par de idiomas que se está trabalhando. Depois é executado o *script* de refinamento com o objetivo de chegar a uma taxa de erro mínima (KOEHN, 2011a).

Na etapa de testes, primeiramente são tokenizados e convertidos para minúsculo os textos de testes que se deseja traduzir, bem como a sua tradução de referência. Por fim, é executada a decodificação do texto em língua fonte (LF) para a língua alvo (LA) com o *Moses* (KOEHN, 2011a).

A última etapa da construção de um tradutor estatístico é a de avaliação dos resultados obtidos com os testes, para verificar o nível de acertos do tradutor. Nesta, o arquivo resultante da etapa de testes será reformatado com as iniciais maiúsculas e destokenizado, além de convertido para o formato de marcação de documento SGML (*Standard Generalized Markup Language*) para então ser comparado com o conjunto de testes de referência. A pontuação de acertos se dará através do sistema NIST BLEU (*Bilingual Evaluation Understudy*) de avaliação (KOEHN, 2011a). De acordo com Papineni, Roukos S. Ward e Zhu (2002), a idéia central do sistema Bleu é atribuir uma melhor pontuação para as traduções automáticas que mais se aproximem de uma tradução humana.

5.3.2 Exemplo de Tradutor Baseado em Regras

Nesta seção será apontado um dos caminhos possíveis para o projeto de um tradutor automático baseado em regras, utilizando a técnica por transferência, através da ferramenta Apertium.

5.3.2.1 Tradutor Baseado em Regras com a Ferramenta Apertium

As seguintes etapas podem ser seguidas para o projeto de um tradutor automático baseado em regras: a definição do par de idiomas a ser utilizado, levantamento de recursos linguísticos dentro do domínio desejado (dicionários monolíngues e bilíngues), configuração e personalização do Apertium, execução de testes e avaliação dos resultados obtidos com o padrão BLEU.

Uma das ferramentas que se destacam na construção de tradutores com a técnica baseada em regras (RBMT) é o Apertium[7] (FRANÇOIS; RIBICZEY; RAMÍREZ-SÁNCHEZ, 2010). Este motor de tradução surgiu em 2005 e foi inicialmente voltado para pares de idiomas linguísticamente próximos. No entanto, o Apertium tem sido melhorado ao longo dos anos para, dentre outras coisas, ampliar os idiomas suportados, abranger pares de idiomas não tão similares e permitir trabalhar com vocabulário de domínio específico, além de existirem trabalhos em que esta ferramenta é utilizada em conjunto com outras (BRANDT; TYERS, 2011).

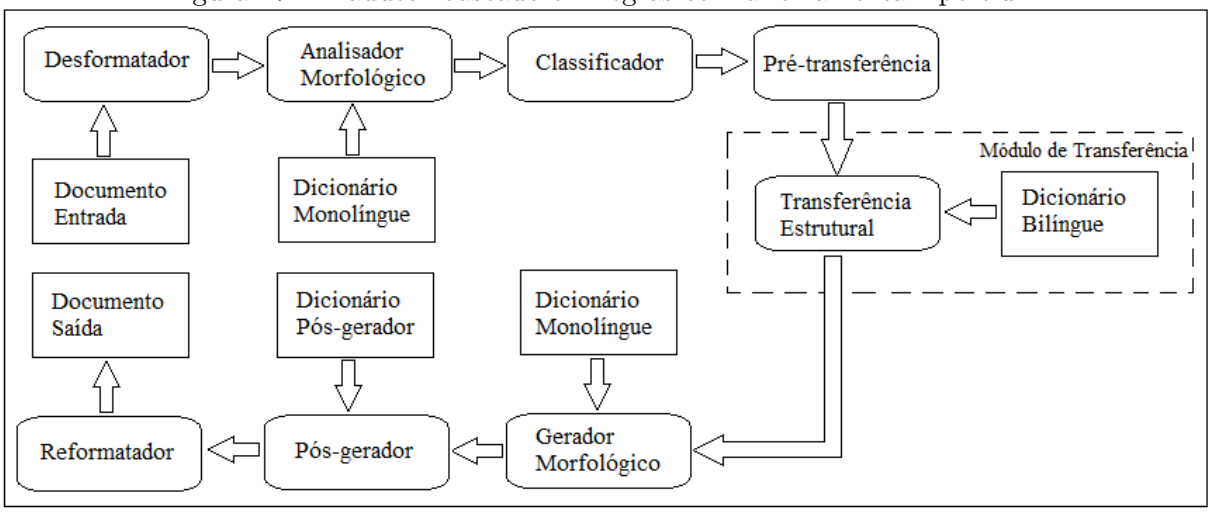
A definição dos pares de idiomas e dos recursos linguísticos deve ser feita anteriormente à configuração da ferramenta. Para se aplicar o Apertium a um contexto de domínio restrito como o de textos técnicos, é necessário utilizar dicionários que abranjam o domínio desejado. Neste caso, pode-se realizar a personalização dos dicionários monolíngues de inglês e português-Br e bilíngue inglês-português-Br disponíveis atualmente na biblioteca do Apertium, inserindo e adequando os vocábulos para o domínio a ser utilizado, por exemplo (FRANÇOIS; RIBICZEY; RAMÍREZ-SÁNCHEZ, 2010). No presente estudo de caso, esta etapa significa a inserção e personalização dos dicionários para abranger adequadamente o vocabulário do domínio dos textos técnicos inglês-português-Br.

Na figura 17, pode-se observar a arquitetura da ferramenta, com seus principais módulos. A idéia central do Apertium pode ser compreendida como uma sequência de traduções simples em diversos formatos através de módulos independentes (FRANÇOIS; RIBICZEY; RAMÍREZ-SÁNCHEZ, 2010). Os módulos que compõem esta ferramenta são: os módulos correspondentes à formatação, os módulos de processamento léxico, o

[7] `http://sourceforge.net/projects/apertium/`

módulo de desambiguação léxica e o módulo de transferência estrutural.

Figura 17 - Tradutor baseado em regras com a ferramenta Apertium.



Fonte: Adaptada de François, Ribiczey e Ramírez-Sánchez (2010).

Os módulos formatador e desformatador realizam a separação entre o texto e sua formatação, o primeiro armazena a formatação e passa o texto pronto para ser processado na próxima etapa, já o segundo coloca de volta a formatação no texto para exibí-lo ao usuário (FRANÇOIS; RIBICZEY; RAMÍREZ-SÁNCHEZ, 2010).

Os módulos do processamento léxico são compostos pelo analisador morfológico, pelo módulo de transferência, pelo gerador morfológico e pelo pós-gerador. O analisador morfológico identifica a forma lexical utilizada em cada palavra do texto de entrada. O módulo de transferência converte grupos de palavras ou uma palavra do idioma de origem para o idioma de destino. O módulo do gerador morfológico imprime a forma lexical correta ao trecho de texto ou palavra traduzidos no módulo de transferência. Por fim, o módulo pós-gerador faz algumas correções ortográficas, como contrações (FRANÇOIS; RIBICZEY; RAMÍREZ-SÁNCHEZ, 2010).

De acordo com François, Ribiczey e Ramírez-Sánchez (2010), o módulo de desambiguação atua no módulo de transferência, quando o analisador léxico fornece mais de uma forma lexical possível para uma palavra. Cabe ao módulo de desambiguação determinar qual a forma lexical estatisticamente mais provável para que o gerador morfológico finalize a geração do léxico na língua alvo.

O módulo de transferência estrutural fica responsável, quando necessário, por realizar algumas adequações estruturais entre a LF e a LA. São alguns exemplos dessas modificações mudanças de tempos verbais, de preposições, de concordâncias, de ordenações, dentre outros (FRANÇOIS; RIBICZEY; RAMÍREZ-SÁNCHEZ, 2010).

Como sugerido em François, Ribiczey e Ramírez-Sánchez (2010), alguns recursos podem ser acoplados ao Apertium para melhorar o nível de qualidade de tradução, como por exemplo, textos técnicos em inglês e sua tradução em português-Br (*corpora* paralela) podem ser usados como memórias de tradução para influenciar nas regras do módulo de transferência e imprimir a característica mais apropriada ao estilo do texto a ser traduzido. Também podem ser feitas listas de significados para *n-grams* (grupos de n palavras seguidas em um texto), para corrigir problemas em que juntas essas palavras teriam um significado diferenciado de quando estivessem isoladas.

5.3.3 Exemplo de Tradutor Baseado em Conhecimento

O terceiro exemplo de ferramenta de tradução a ser apresentado utiliza a técnica baseada em conhecimento. Será utilizada a ferramenta Jena para ilustrar o uso da técnica KBMT.

5.3.3.1 Tradutor Baseado em Conhecimento Semântico com a Ferramenta Jena

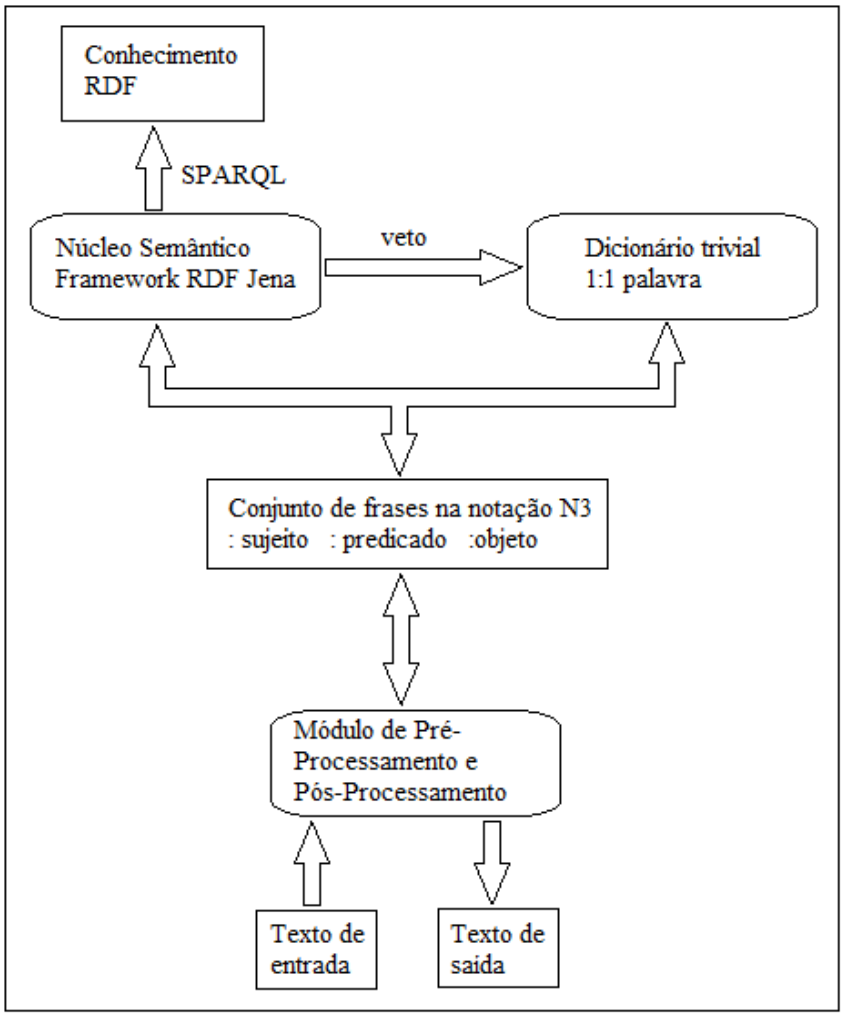
Nesta seção, será apresentada um exemplo de como projetar uma ferramenta de tradução automática inglês-português-Br baseada em conhecimento da *web* semântica através da ferramenta Jena, tendo como base o trabalho desenvolvido por Harriehausen-Mühlbauer e Heuss (2012). A ferramenta Jena[8] é um *framework* que pode ser utilizado para realizar este processamento do núcleo semântico do tradutor. Ela é voltada para aplicações de processamento da *web* semântica, através de manipulações de arquivos no padrão RDF e suporte à linguagem de solicitações SPARQL (APACHE, 2013).

Os recursos necessários ao projeto são uma base de domínio no padrão RDF (*Resource Description File* - padrão de descrição de conhecimento na *web* semântica), um dicionário bilíngue trivial inglês-português-Br, um núcleo de processamento semântico e um módulo de pré-processamento e pós-processamento (figura 18).

O módulo de pré-processamento realiza a conversão do texto de entrada em inglês para um conjunto de frases composta por três partes, de acordo com a notação N3[9] do RDF: sujeito, predicado e objeto. Esta notação é utilizada por possuir a vantagem de ser mais compreensível ao ser humano que outras notações formais, permitindo ao desenvolvedor do sistema realizar algum acompanhamento sobre o atual estado de análise do tradutor. O módulo de pós-processamento realiza a tarefa inversa da do pré-processamento, convertendo do formato N3 para o texto em português-Br. O trabalho

[8] `http://jena.apache.org/`
[9] `http://www.w3.org/DesignIssues/Notation3.html`

Figura 18 - Tradutor baseado em conhecimento com a ferramenta Jena.



Fonte: Adaptada de Harriehausen-Mühlbauer e Heuss (2012).

realizado por este último módulo é de somente unir as três partes da frase, sujeito, predicado e objeto (HARRIEHAUSEN-MÜHLBAUER; HEUSS, 2012).

O dicionário trivial será responsável pelas traduções palavra a palavra, similarmente à técnica da tradução direta. Não há verificações de contexto, nem restrições lexicais, gramaticais ou semânticas, neste módulo somente é feita a substituição de palavras em inglês por palavras em português-Br (HARRIEHAUSEN-MÜHLBAUER; HEUSS, 2012).

Uma vez feita a tradução "crua" pelo módulo do dicionário, o núcleo semântico realizará a tradução contextualizada de acordo com a base RDF. Segundo Harriehausen-Mühlbauer e Heuss (2012), o núcleo semântico deve procurar uma tradução mais bem qualificada que a do dicionário, através de consultas com a linguagem SPARQL à base de conhecimento semântico RDF.

Os resultados obtidos por Harriehausen-Mühlbauer e Heuss (2012), na ferramenta similar à proposta, mostram que os experimentos com *web* semântica na criação de tradutores automáticos baseados em conhecimentos ainda precisam amadurecer um pouco mais. Para obter um resultado satisfatório a nível de profundidade do conhecimento especificado, a base RDF deverá ser criada de forma eficiente. A especificação de um campo chamado palavra-chave (*trigger word*) em um conceito da base RDF é utilizado com o intuito de disparar o carregamento de conceitos relacionados. Também é utilizado o campo rótulo multilíngue (*multi-language-label*) que serve para armazenar a palavra correspondente em inglês e em português-Br do conceito em questão. Quanto mais complexa a rede for projetada, mais consultas SPARQL serão necessárias para obter uma informação desejada e uma maior quantidade de processamento é requerida para o sistema. Um outro problema é a incapacidade do sistema determinar a melhor tradução que pode ser extraída da base, quando houver mais de uma possibilidade.

A maior vantagem de utilizar essa abordagem está em solucionar problemas de caráter sintático e semântico que muitas vezes sequer são verificados em outras abordagens. Então, um benefício incremental seria obtido da combinação da abordagem proposta desta ferramenta baseada em conhecimento semântico juntamente com a abordagem baseada em estatística. Neste caso, a abordagem *n-grams* estatística poderia resolver o problema da determinação da melhor tradução, enquanto a baseada em conhecimento filtraria as traduções relevantes dentro do contexto restrito especificado.

5.4 DISCUSSÕES

A partir do estudo das técnicas de tradução automática levantadas, foi possível estabelecer um perfil de utilização de cada uma delas, determinar suas características marcantes, recursos necessários para sua aplicação, suas limitações, desempenho e custo. Tais informações foram detalhadas ao longo da explicação das técnicas e, posteriormente, na seção de análise comparativa foram sintetizadas e apresentadas graficamente através de um quadro comparativo construído com base em critérios descritos na mesma seção.

Ao longo do estudo sobre as técnicas, foi possível observar que a técnica de tradução em destaque atualmente na TA é a estatística (SBMT), uma evolução da abordagem de tradução direta. Tal decorre de sua grande vantagem de não ser necessário ter disponível um profissional da área de linguística durante o desenvolvimento da ferramenta, nem ser necessário criar manualmente as regras de mapeamento lexical, sintática ou semântica. Essas vantagens citadas ainda resultam nas vantagens de rápido desenvolvimento e baixo custo para criação de um tradutor com essa técnica (BRANDT; TYERS, 2011).

Se por um lado, vários projetos como o Pluto (BRANDT; TYERS, 2011) vêm obtendo bons resultados nas traduções de textos técnicos utilizando a técnica de tradução estatística, em algumas situações não é viável sua aplicação. Por exemplo, existem línguas (português-Br e islandês, como exemplo) que não dispõem de recursos essenciais, como uma *corpora* paralela extensa (inglês-português-Br e inglês-islandês, respectivamente), para seu processamento com a técnica citada. Nesses casos, normalmente, é empregada a técnica de tradução baseada em regras, juntamente com alguma outra abordagem, em que os extensos textos da *corpora* paralela não são requeridos.

Segundo Frankenberg-garcia e Santos (2002), não existe uma *corpora* paralela pública de textos técnicos inglês-português-Br em razão de normalmente não ser realizada a tradução de materiais técnicos, artigos, patentes, pois os profissionais que necessitam acessar ou publicar tal material normalmente já possuem a habilidade de ler e escrever diretamente em inglês. Dessa maneira, assim como o islandês, atualmente a língua brasileira se configura como um idioma com poucos recursos para o êxito da implementação de uma ferramenta baseada em tradução estatística.

Existem estudo iniciais no sentido de criar uma *corpora* paralela inglês-português de textos técnicos extraídos de uma revista científica brasileira, a "Pesquisa FAPESP *Online*", como é o caso do experimento apresentado em Aziz e Specia (2011). Como explicado, os resultados não são animadores, quando comparados com uma base de referência na área da tradução automática, a *Europarl* (KOEHN, 2011a), por exemplo. Ainda sem resultados perfeitos, o par inglês-português da *Europarl* possui 1.960.407 sentenças alinhadas, contra as 150.000 sentenças alinhadas que podem ser obtidas do experimento realizado por Aziz e Specia (2011).

No estudo apresentado por Brandt e Tyers (2011), é apresentado um sistema de tradução automática islandês-inglês que utiliza a técnica de tradução baseada em regras, como forma de contornar a ausência de *corpora* paralela do par de línguas em questão. No trabalho, foi utilizado o *framework* Apertium que implementa esta técnica, com uma modificação no módulo de processamento do islandês. O objetivo do trabalho desenvolvido foi descobrir se um sistema de tradução por transferência simples utilizando os módulos de processamento de texto do islandês (IceNLP) teria uma qualidade de tradução superior ao framework Apertium meramente, sem qualquer alteração.

Brandt e Tyers (2011) informam que as taxas de erro obtidas com a ferramenta desenvolvida são altas quando comparadas com as de outras ferramentas de tradução utilizadas publicamente (Google Translate e Tungutorg). Segundo a análise de erros de Brandt e Tyers (2011), uma melhor qualidade da tradução poderia ter sido obtida caso

somente o componente novo fosse utilizado em conjunto com o componente de análise morfológica já presente originalmente no Apertium.

No exemplo de tradutor estatístico apresentado, pode-se observar que já é presente na literatura materiais bem específicos com os passos a serem seguidos para a construção de tradutores com esta técnica, através ferramentas *opensource*, como o *Moses* que viabilizam o desenvolvimento rápido da mesma. No entanto, a construção de um tradutor estatístico técnico com resultados em destaque atualmente ainda não é possível, devido à inexistência de uma *corpora* paralela de textos técnicos inglês-português-Br. Dessa forma, é apresentada uma sugestão de como confeccionar uma *corpora*, com base em estudos preliminares realizados por Aziz e Specia (2011).

Com relação ao exemplo de tradutor baseado em regras apresentado, é possível notar que os materiais presentes na literatura não são tão específicos quanto aos referentes à técnica estatística. No caso da técnica RBMT, cada projeto desenvolvido constitui um projeto feito sob medida, por cada regra de mapeamento lexical e sintático determinados, para os pares de línguas e tipo de linguagem do contexto utilizado. No entanto, pôde-se identificar a ferramenta *Apertium* (FRANÇOIS; RIBICZEY; RAMÍREZ-SÁNCHEZ, 2010) voltada para desenvolvimento de tradutores baseados em regras que tem como objetivo automatizar o processo da tradução principal desta técnica, deixando a cargo do desenvolvedor a personalização e configuração dos dicionários bilíngues, monolíngues e módulos de desambiguação utilizados. Para experimentações simples, é possível utilizar dicionários em versões de desenvolvimento disponíveis na biblioteca do *Apertium* para os pares de língua inglês-português-Br. No entanto, mais especificamente, para realizar testes no domínio de textos técnicos é preciso modificar a base de dicionários existente.

O exemplo de tradutor apresentado utiliza a técnica baseado em conhecimento (KBMT). Neste caso, é sugerida a utilização da ferramenta *Jena* (HARRIEHAUSEN-MÜHLBAUER; HEUSS, 2012), que realiza o fluxo de processamento central de um tradutor KBMT. Para a criação de um tradutor técnico inglês-português-Br ainda seria necessário desenvolver ou buscar uma base de conhecimentos RDF da *web* semântica para utilização como base no sistema. De acordo com Harriehausen-Mühlbauer e Heuss (2012), a personalização de uma base RDF e processamento de uma base do mesmo tipo com relações bem complexas acarretam em um custo de desenvolvimento e desempenho grandes, respectivamente. Uma sugestão, caso venha a ser decidida a utilização desta técnica, é o uso da mesma é conjunto com outra, por exemplo da estatística (SBMT) que poderia suprir a carência da indecidibilidade da KBMT pela melhor tradução pesquisada na ontologia.

Por isso, conclui-se que, apesar de viáveis, os exemplos de tradutores apresentados como sugestão para a tradução inglês-português-Br não são de aplicação imediata. Tais técnicas exigem a execução de algumas tarefas custosas anteriores ao desenvolvimento do tradutor referido, como desenvolvimento da *corpora* paralela, adequação dos dicionários ou desenvolvimento e personalização da base de conhecimentos semânticos RDF, por vezes, necessitando de especialistas linguísticos.

6 CONSIDERAÇÕES FINAIS

O presente trabalho apresenta uma revisão bibliográfica sobre o campo de estudo da tradução automática (TA), identificando referências norteadoras da pesquisa sobre o assunto.

Foi apresentada a evolução das pesquisas em TA pelos acontecimentos que foram destacados no capítulo de Histórico. Em seguida, um estudo mais detalhado de cada uma das técnicas de tradução automática mais populares foi descrito, com um enfoque sobre os conceitos envolvidos em cada uma das técnicas, seus processos de tradução e princípios envolvidos. Para uma melhor compreensão das técnicas estudadas, foi feita uma análise comparativa das mesmas, utilizando suas principais características para formalizar um quadro comparativo que permitisse identificar falhas e benefícios de cada uma das técnicas. Por fim, foram apresentados exemplos de tradutores utilizam três das técnicas estudadas.

Na etapa de revisão bibliográfica das técnicas de tradução automática, pôde ser observado que as referências teóricas encontradas de algumas das técnicas são anteriores a 2000. Quando não são antigas, são especializadas em uma única técnica de tradução. Observou-se também que há uma tendência atual na publicação de artigos voltados especificamente à modelagem e implementação de projetos, em detrimento de publicações mais teóricas sobre o assunto. Imagina-se que a conjuntura encontrada se deva principalmente ao fato de que TA é uma área de pesquisa que já data de algumas décadas, o que induz a uma sensação de certo amadurecimento teórico na área e, por consequência desse fato, há uma falta de estímulo nas publicações de cunho teórico.

Quanto ao cenário de pesquisa nacional em TA, foi detectada a ausência de outros grupos de pesquisa, além do Núcleo Interinstitucional de Linguística Computacional (NILC) da USP de São Carlos, e poucas publicações na área de TA nos eventos nacionais voltados para computação. Aparentemente, esta área não é muito explorada no Brasil. Portanto, o presente trabalho apresenta uma revisão das principais técnicas de TA resultando em um estudo comparativo destas técnicas.

Diferentemente do cenário nacional, as pesquisas no âmbito internacional são abundantes em variedade de técnicas utilizadas, ferramentas novas propostas, novos pares de idiomas testados, novas combinações de técnicas híbridas experimentadas, além de propostas para interpor barreiras de particularidades linguísticas e limitações computacionais.

O trabalho apresenta um estudo inicial e diversas extensões podem ser sugeridas como continuação do presente estudo: o desenvolvimento de uma *corpora* paralela

inglês-português-Br para um domínio específico utilizando um processo semelhante ao sugerido por Aziz e Specia (2011), possivelmente para textos técnicos da área de Computação; o desenvolvimento de uma base RDF voltada para a tradução automática inglês-português-Br; o desenvolvimento de uma ferramenta de busca e geração automática de bases de dados da *web* semântica para a TA baseada em conhecimento; a implementação de um tradutor automático inglês-português-Br utilizando umas das três técnicas apresentadas como exemplo.

REFERÊNCIAS

ABEILLÉ, A.; JOSHI, A. K.; SCHABES, Y. Using lexicalized tags for machine translation. In: **Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)**. Philadelphia: [s.n.], 1990. Disponível em: <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1347&context=cis_reports>. Acesso em: 26 mar. 2013.

APACHE. **Apache Jena**. 2013. Disponível em: <http://jena.apache.org/>. Acesso em: 24 jun. 2013.

ARNOLD, D. J. *et al.* **Mahine Translation: An Introdutory Guide.** London: NCC Blackwells, 1994.

AZIZ, W.; SPECIA, L. Fully automatic compilation of portuguese-english and portuguese-spanish parallel corpora. In: **Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL-2011)**. Cuiabá: [s.n.], 2011. Disponível em: <http://clg.wlv.ac.uk/papers/AzizSpecia_STIL2011.pdf>. Acesso em: 26 mar. 2013.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. **The Semantic Web.** 2001. Disponível em: <http://www-sop.inria.fr/acacia/fabien/lecture/licence_travaux_etude2002/TheSemanticWeb/>. Acesso em: 30 mar. 2013.

BLACK, C. A. **A step-to-step introduction to the Government and Biding theory of syntax.** Summer Institute of Linguistics, 1999. Disponível em: <http://www-01.sil.org/mexico/ling/e002-introgb.pdf>. Acesso em: 09 mai. 2013.

BLANCHON, H. Perspectives of dbmt for monolingual authors on the basis of lidia-1, an implemented mock-up. In: **Proceedings of the 15th conference on Computational linguistics (Coling-94)**. Association for Computational Linguistics, 1994. Disponível em: <http://www-clips.imag.fr/geta/herve.blanchon/Pdfs/COLING94.pdf>. Acesso em: 06 jun. 2013.

BOITET, C.; BLANCHON, H. **Multilingual Dialogue-Based MT for Monolingual Authors: The LIDIA Project and a First Mockup (1994)**. 1995. 99-132 p. Disponível em: <http://www-clips.imag.fr/geta/herve.blanchon/Pdfs/MT95.pdf>. Acesso em: 06 jun. 2013.

BRANDT, M.; TYERS, F. Apertium-icenlp: A rule-based icelandic to english machine translation system. In: **Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT-2011)**. Lovaina: [s.n.], 2011. Disponível em: <http://www.ru.is/~hrafn/papers/is-en.pdf>. Acesso em: 10 jan. 2013.

CARL, M.; RASCU, E.; SCHMIDT, P. Using template-grammars for shake & bake paraphrasing. In: **Proceedings of the 10th EAMT**. Budapest: [s.n.], 2005. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.3552&rep=rep1&type=pdf>. Acesso em: 13 mai. 2013.

DORR, B. **Principle-Based Parsing for Machine Translation.** Massachusetts Institute of Technology, 1987. Disponível em: <http://dspace.mit.edu/bitstream/handle/1721.1/6462/AIM-947.pdf?sequence=2>. Acesso em: 15 abr. 2013.

DORR, B.; JORDAN, P.; BENOIT, J. **A survey of current paradigms in machine translation.** University of Maryland, 1998. Disponível em: <http://www.dtic.mil/dtic/tr/fulltext/u2/a455393.pdf>. Acesso em: 15 abr. 2013.

DORR, B. *et al.* **Efficient Parsing for Korean and English: A Parameterized Message-Passing Approach.** Computational Linguistics, v. 21, p. 255-263, 1995. Disponível em: <http://acl.ldc.upenn.edu/J/J95/J95-2005.pdf>. Acesso em: 15 abr. 2013.

FRANÇOIS, M.; RIBICZEY, P.; RAMÍREZ-SÁNCHEZ, G. **Using the Apertium Spanish-Brazilian Portuguese Machine Translation System for Localization.** St. Raphael: EAMT, 2010. Disponível em: <http://www.researchgate.net/publication/228527060_Using_the_Apertium_Spanish-Brazilian_Portuguese_machine_translation_system_for_localization/file/d912f50a5f71ca97b1.pdf>. Acesso em: 23 jun. 2013.

FRANKENBERG-GARCIA, A.; SANTOS, D. **COMPARA, um corpus paralelo de português e inglês na Web.** Florianópolis: NUT, v. 1, p. 61-79, 2002. Disponível em: <http://www.periodicos.ufsc.br/index.php/traducao/article/download/5981/5685>. Acesso em: 10 jan. 2013.

GIZA-PP. **giza-pp - GIZA++ statistical translation models toolkit.** 2013. Disponível em: <https://code.google.com/p/giza-pp/>. Acesso em: 19 jun. 2013.

GOUTTE, C. *et al.* (Ed.). **Learning Machine Translation**. Cambridge: MIT Press, 2009.

HARRIEHAUSEN-MÜHLBAUER, B.; HEUSS, T. Semantic web based machine translation. In: **Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics**. Avignon: [s.n.], 2012. p. 1–9. Disponível em: <http://www.mt-archive.info/EACL-2012-Harriehausen.pdf>. Acesso em: 24 jun. 2013.

HARSHAWARDHAN, R. **Rule Based Machine Translation System for English to Malayalam Language**. Coimbatore: Amrita School of Engineering, 2011. Disponível em: <http://nlp.amrita.edu:8080/project/mhrd/ms/Dissertation.pdf>. Acesso em: 26 mar. 2013.

HUTCHINS, W. J. **Machine Translation: Past, Present and Future.** New York: Halsted Press, 1986.

HUTCHINS, W. J. **Example Based Machine Translation - a Review and Commentary**. 2005. 197-211 p. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.1082&rep=rep1&type=pdf>. Acesso em: 06 jun. 2013.

HUTCHINS, W. J. **Machine Translation Archive.** 2013. Disponível em: <http://www.mt-archive.info/>. Acesso em: 15 jun. 2013.

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

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 1st. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000.

KAPLAN, R. M. *et al.* Translation by structural correspondences. In: **Proceedings of 4th EACL**. Manchester: [s.n.], 1989. p. 81–272. Disponível em: <http://acl.ldc.upenn.edu/E/E89/E89-1037.pdf>. Acesso em: 26 mar. 2013.

KAUCHAK, D. **Contributions to Research on Machine Translation.** Dissertation (Doctor of Philosophy) - University of California, San Diego, 2006. Disponível em: <http://www.cs.middlebury.edu/~dkauchak/thesis/kauchak06.thesis.pdf>. Acesso em: 26 mar. 2013.

KOEHN, P. **Baseline System: Moses.** 2011a. Disponível em: <http://www.statmt.org/wmt11/baseline.html>. Acesso em: 19 jun. 2013.

KOEHN, P. **Statistical Machine Translation.** 2012. Disponível em: <http://www.statmt.org>. Acesso em: 03 dez. 2012.

KOEHN, P. **Welcome to Moses!** 2013. Disponível em: <http://www.statmt.org/moses/>. Acesso em: 19 jun. 2013.

LIMA, W. E. M. **Problemas NP-Completo**. 2001. Disponível em: <http://www.ime.usp.br/~weslley/probNP.htm>. Acesso em: 29 jul. 2013.

NIRENBURG, S.; SOMERS, H.; WILKS, Y. A. (Ed.). **Readings in Machine Translation**. Cambridge: MIT Press, 2003. (Bradford Books).

NTT. **Statistical Machine Translation (SMT).** 2007. Disponível em: <http://www.ntt.co.jp/RD/OFIS/active/2007pdfe/hot/ct/06.html>. Acesso em: 17 jun. 2013.

OLIVE, J.; CHRISTIANSON, C.; MCCARY, J. (Ed.). **Handbook of Natural Language Processing and Machine Translation**. New York: Springer Verlag, 2011. Accepted for publication.

OLIVEIRA, O. N. J. *et al.* **A Critical Analysis of the Performance of English-Portuguese-English MT Systems.** São Paulo: ICMC/USP, 2000. 85-92 p. Disponível em: <http://nilc.icmc.sc.usp.br/nilc/download/criticalanalysis.zip>. Acesso em: 03 dez. 2012.

PAPINENI, K.; ROUKOS S. WARD, T.; ZHU, W. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)**. Philadelphia: [s.n.], 2002. p. 311–318. Disponível em: <http://acl.ldc.upenn.edu/P/P02/P02-1040.pdf>. Acesso em: 20 jun. 2013.

SILVA, B. C. D. *et al.* **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações.** São Paulo: Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC - USP, 2007. Disponível em: <http://www.letras.etc.br/ebralc/NILCTR0710-DiasDaSilvaEtAl.pdf>. Acesso em: 03 dez. 2012.

SPECIA, L.; RINO, L. H. M. **Introdução aos Métodos de Tradução e Paradigmas de Tradução Automática.** São Paulo: Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC - ICMC - USP, 2002. Disponível em: <http://www.nilc.icmc.usp.br/nilc/pessoas/specia/publications/TR0204-SpeciaRino.pdf>. Acesso em: 03 dez. 2012.

SRI. **SRILM - THe SRI Language Modeling Toolkit**. 2011. Disponível em: <http://www.speech.sri.com/projects/srilm/>. Acesso em: 03 dez. 2012.

TAVARES, O. L. **Análise Sintática**. 2000. Disponível em: <http://www.inf.ufes.br/~tavares/labcomp2000/anasint.html>. Acesso em: 29 jul. 2013.

TINSLEY, J.; WAY, A.; SHERIDAN, P. Pluto: Mt for online patent translation. In: **Proceedings of AMTA 2010**. Denver: [s.n.], 2010. Disponível em: <http://www.pluto-patenttranslation.eu/sites/default/files/AMTA_2010_Pluto.pdf>. Acesso em: 10 jan. 2013.

TURCATO, D. Shake-and-bake mt and morphology. In: **Proceedings of the sixth International Conference on Theoretical and Methodological Issues in Machine Translation - TMI95**. Leuven: [s.n.], 1995. p. 318–326. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.136.9627&rep=rep1&type=pdf>. Acesso em: 13 mai. 2013.

VERTAN, C. **Example-based Machine Translation**. 2004. Disponível em: <http://nats-www.informatik.uni-hamburg.de/pub/MT05/VeranstaltungsMaterial/introd_ebmt.pdf>. Acesso em: 17 jun. 2013.

VERTAN, C. **Knowledge Based Machine Translation**. 2005. Disponível em: <http://nats-www.informatik.uni-hamburg.de/pub/User/IntensiveCourseInMachineTranslation/kbmt.pdf>. Acesso em: 30 mar. 2013.

WHITELOCK, P. Shake-and-bake translation. In: **Proceedings of the Fifteenth International Conference on Computational Linguistics - Coling-92**. Nantes: [s.n.], 1992. p. 784–791. Disponível em: <http://www.mt-archive.info/Coling-1992-Whitelock.pdf>. Acesso em: 13 mai. 2013.

WILKS, Y. **Machine Translation - Its Scope and Limits.** Sheffield: Springer, 2009.