UNIVERSIDADE DE RIO VERDE
FACULDADE DE CIÊNCIA DA COMPUTAÇÃO

ESTUDO DA UTILIZAÇÃO DE TÉCNICA DE PROCESSAMENTO DE LINGUAGEM NATURAL PARA OTIMIZAÇÃO DE TRADUTORES AUTOMÁTICOS

RÊGES FARIA VINHAES
Orientador: Profº. Ms: BRUNO QUEIROZ PINTO

Monografia apresentada à Faculdade de Ciência da Computação como parte das exigências para obtenção do título de Bacharel pela Universidade de Rio Verde.

RIO VERDE -GO
2005

DEDICATÓRIA

Aos meus pais Fábio José Vinhaes e Elizete Arlene de F. Vinhaes, pelo amor, educação, compreensão e espírito de luta que me passaram em todos os momentos que precisei.

A minha filha Edianny Morais de F. Vinhaes, por compreender à minha ausência, mas que sempre esteve ao meu lado.

Ao meu irmão Alex Faria Vinhaes, pela amizade e companheirismo que juntos nos acompanhou.

A minha namorada Nábia Pereira da Silva, que esteve ao meu lado durante todos estes anos de estudo, para que eu concluísse meus estudos.

Ao meu irmão Leandro Faria Vinhaes e sua esposa Vanuza Lopes dos Santos, que tanto me apoiaram nesta luta.

Aos meus sobrinhos Kaike, Lainny Amélia e João Pedro, pelo amor, carinho e todas as perguntas que me fizeram durante o desenvolvimento deste trabalho.

AGRADECIMENTOS

A Deus, pela oportunidade de poder estar vivendo este momento.

Aos meus colegas, que passamos juntos durante todos estes anos, compartilhando momentos de alegria, tristeza e vitória, e que juntos lutamos pelo mesmo objetivo. Fica agora as lembranças da convivência que tivemos, mas que nunca se apagarão de nossas mentes.

Bruno Queiroz Pinto, meu orientador, por ter aceitado este desafio, sem medir esforços para que pudéssemos juntos concluirmos este trabalho, meu eterno agradecimento.

A profª Lúcia Specia, pela sua atenção ao me ajudar com este trabalho, sem sequer conhecermos, minha eterna gratidão.

A profª Milena, pela humildade e interesse em colaborar com a conclusão deste trabalho, minha admiração.

A todos os professores, que durante estes anos nos transmitiram conhecimento, minha eterna amizade.

RESUMO

VINHAES, R. F. **Estudo da Utilização de Técnica de Processamento de Linguagem Natural para Otimização de Tradutores Automáticos.** 2005. 57f. Monografia (Curso de Ciência da Computação) – Universidade de Rio Verde, Rio Verde. 2005[*].

A preocupação com a qualidade da tradução feita por tradutores automáticos, tem sido um grande desafio para os profissionais da área de Computação e Lingüística, isto se deve à grande complexidade neste tipo de problema e também da sua aplicação em qualquer tipo de linguagem natural, quando na tradução de texto de uma língua fonte para uma língua alvo. Os profissionais desenvolvedores destes sistemas vem se preocupando com o fator qualidade nas traduções realizadas, pois tem sido apresentado um alto grau de ineficiência destes sistemas principalmente quando se trata de traduções de palavras ambíguas, considerando a semântica do texto traduzido. Para poder entender melhor o funcionamento destes sistemas, analisamos algumas das diversas ferramentas disponíveis, para fazer um estudo sobre a eficiência nas traduções, dentre estas foram utilizados o Systran, FreeTranslation e E-Translation Server. Mesmo utilizando técnica de processamento de linguagem natural tais sistemas não apresentam a capacidade de evitar problemas de ambigüidade, sendo necessário para isto, o desenvolvimento de desambiguadores para cada tipo de problema. Entretanto, estes sistemas apenas conseguem apresentar um certo grau de eficiência, se aplicados a áreas especificas.

Palavras-chave: Processamento de Linguagem Natural, Tradução Automática, Ambigüidade, Processamento de Corpus, Semântica.

---

[*] Orientador: Profº Ms: Bruno Queiroz Pinto, Mestre em Ciência da Computação, Universidade de Rio Verde. Banca Examinadora: Profº. Ms: Miguel Raimundo de Oliveira Neto, Mestre em Ciência da Computação, Universidade de Rio Verde, Profª. Ms: Luciana Recart Cardoso, Mestre em Ciência da Computação, Universidade de Rio Verde.

ABSTRACT

VINHAES, R. F. **Study of Utilization of Technique of Natural Language Processing for Optimization of Automatics Translation.** 2005. 57s. Monograph (Course of Computer Science) – Universidade de Rio Verde. 2005[*].

The concern with the quality of translation done by automatic translation programs, has been a great challenge to the professionals of Computing and Linguistic. It happens because there is a big complexity in the subject and also in its application in text translations from a source language to a target language. The professionals that develop these systems are worried with the quality of the translations done by them, because there has been a high (degree) of failure in these programs, mainly in translations of ambiguous words, taking into consideration the semantics of the translated text. For a better understanding of the system operation it was analyzed some of the tools available, to do a study about the efficiency in the translations, among some programs, it was tested Systran, Free Translations and E-Translation Server. Even using the utilization of Technique of Natural Language Progressing, the systems don't show the capacity of avoiding the ambiguity problem, that's why it's necessary the development of unambiguous programs for each type of problem. However, those systems will only succeed if they were applied in specific areas.

Key-Words: Natural Language Processing, Automatic Translation, Ambiguity, Corpus Processing, Semantic.

---

[*] Orientador: Profº Ms: Bruno Queiroz Pinto, Mestre em Ciência da Computação, Universidade de Rio Verde. Banca Examinadora: Profº. Ms: Miguel Raimundo de Oliveira Neto, Mestre em Ciência da Computação, Universidade de Rio Verde, Profª. Ms: Luciana Recart Cardoso, Mestre em Ciência da Computação, Universidade de Rio Verde.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AC – Ambigüidade Categorial

AD – Análise do Discurso

AL – Ambigüidade Léxica

ALS – Ambigüidade Léxica de Sentido

AM – Análise Morfológica

AS – Análise Semântica

DLS – Desambiguação Léxica de Sentido

DCG – *Definite Clause Grammar*

EBMT – *Exemple-Based Machine Translation*

IA – Inteligência Artificial

KBMT – *Knowledged-Based Machine Translation*

LA – Língua Alvo

LBMT – *Language-Based Machine Translation*

LC – Lingüística Computacional

LF – Língua Fonte

LN – Linguagem Natural

PLN – Processamento de Linguagem Natural

POS taggers - *Part-of-speech taggers*

PROLOG – *Programming in Logic*

SE – Sistemas Especialistas

STA – Sistemas de Tradução Automática

TA – Tradução Automática

TH – Tradução Humana

SUMÁRIO

1 INTRODUÇÃO

Desde os primórdios da vida humana o homem vem aperfeiçoando os meios de comunicação para melhor se relacionar com os diferentes idiomas utilizados pelas civilizações. A partir da década de 90, devido à globalização, a comunicação entre os povos superou todas as barreiras, principalmente com o uso da Internet, que aproximou toda e qualquer civilização à necessidade de entender os diferentes tipos de idiomas utilizados. Devido a grande dificuldade que as pessoas apresentam durante o aprendizado de uma outra linguagem natural, houve a necessidade de construir *softwares* que fossem capazes de traduzir textos utilizando computadores.

Geralmente estes *softwares* fazem as traduções utilizando dicionários bilíngües, traduzindo palavra por palavra. Tradutores que utilizam estas técnicas não possuem eficiência na construção do texto, pois quando é feita a tradução, não são capazes de fazer análise sintática, léxica, morfológica e semântica do texto traduzido.

O objetivo deste trabalho é o estudo da utilização de Processamento de Linguagem Natural (PLN) aplicados à Tradução Automática (TA).

O PLN é uma técnica utilizada em Inteligência Artificial (IA) que possui métodos formais capazes de entender, compor texto e formular a construção da tradução levando em consideração sua semântica, aproximando-se a realidade da linguagem natural (LN). A tradução de um texto não é tão simples, pois palavras de idiomas diferentes não possuem a mesma estrutura ou significado. A pesar dos recursos de TA serem muito limitados para o computador, a técnica de PLN, facilita o melhoramento destes softwares, diminuindo erros e melhorando o sentido dos textos traduzidos.

Para que um sistema computacional interprete uma sentença em LN, é necessário manter informações morfológicas, sintáticas e semânticas, armazenadas em dicionários, juntamente com as palavras que o sistema compreende.

Verificando-se a crescente demanda por novos sistemas de tradução automática (STA) e sabendo-se que atualmente tais sistemas não apresentam o desempenho esperado, devido a sua incapacidade de tratar da semântica do texto a ser traduzido, tal problema se

mostra crítico para o desenvolvimento de um *software* que atenda as funcionalidades necessárias.

Uma vez identificados tais problemas em STA, realizou-se um estudo sobre estes sistemas para comparar as traduções retornadas, procurando entender o funcionamento destes sistemas.

Aplicou-se conceitos de Lingüística Computacional (LC), para identificar e classificar os problemas que surgirem em uma tradução, facilitando o desenvolvimento de métodos, que venham a facilitar o desenvolvimento de uma proposta para uma futura implementação do código fonte que consiga corrigir ou aproximar ao máximo a tradução de uma LN.

Este estudo visou à utilização da linguagem lógica *Prolog* para uma futura implementação. Apesar de poder utilizar outros tipos de linguagens é viável que se utilizem linguagens lógicas, devido à facilidade que estas oferecem na implementação, pois foram desenvolvidas para este propósito.

Após pesquisar os substantivos ambíguos, as sentenças foram selecionadas aleatóriamente em um *corpus* para serem inseridas em diferentes STA, foram comparados o desempenho de cada um destes sistemas. Após serem examinados, todos os resultados obtidos, desenvolveu-se um estudo, da viabilidade de correção de problemas encontrados.

2 REVISÃO DA LITERATURA

Neste capítulo serão abordados todos os tópicos necessários para estudo e desenvolvimento deste trabalho.

2.1 Inteligência Artificial

A IA é a parte da Ciência da Computação que estuda o desenvolvimento de sistemas computacionais que tenham características que podem ser associadas com a inteligência no comportamento humano, tais como: compreensão da linguagem, aprendizado, raciocínio, resolução de problemas, etc (FERNANDES, 2003). A IA também pode ser definida como o estudo das faculdades mentais através do uso de modelos computacionais (BITTENCOURT, 2004).

Em 1950, Alan Turing, propôs um teste para fornecer uma definição operacional do que seria inteligência. Para Turing, uma máquina poderia ser considerada inteligente se ela fosse capaz de responder perguntas feitas por um interrogador, e que o mesmo não saberia se as respostas estavam vindo de uma pessoa ou não (RUSSEL e NORVIG, 2004).

Uma máquina para passar neste teste teria que ter as seguintes capacidades:

- PLN: capacidade de comunicação com um idioma natural;
- Representação do conhecimento: capacidade de armazenar conhecimento de tudo o que foi aprendido;
- Raciocínio automatizado: capacidade de utilizar todo conhecimento adquirido, e tirar novas conclusões;
- Aprendizado de máquina: poder adaptar-se a novas circunstâncias (RUSSEL e NORVIG, 2004)

Os primeiros conceitos relacionados à I.A surgiram na Grécia antiga, com o surgimento da lógica. Entretanto, somente a partir da década de 40, com o desenvolvimento

da computação, a I.A começou a desenvolver-se, pois os avanços da computação permitiram simulações artificiais da inteligência através do computador.

A IA é uma ciência nova, iniciada na década de quarenta, baseada em pesquisas que estudavam o funcionamento do cérebro com o objetivo de formalizar seu comportamento. Mas somente a partir de 1956, a IA foi reconhecida como uma ciência.

A I.A fundamentou-se em várias áreas do conhecimento humano, tais como: neurociência (que fornece conhecimentos sobre o funcionamento do nosso cérebro), filosofia (estuda como é o pensamento humano), entre outras.

Entre estas e outras áreas, a lingüística é essencial para o desenvolvimento da PLN, pois além de fornecer formas de representar o conhecimento, ela propicia todos os conhecimentos necessários para o desenvolvimento de um sistema que seja capaz de processar alguma linguagem natural.

A IA pode ser subdividida basicamente em duas linhas de pesquisa para construção de sistemas inteligentes: a linha conexionista (biológica) e a linha simbólica (cognitiva) ou psicológica. A linha conexionista (biológica) trabalha com pesquisas que visam modelar a inteligência humana através dos componentes do cérebro, os neurônios e suas ligações. Esta linha de pesquisa foi formalizada em 1943 pelo neuropsicólogo McCulloch e pelo lógico Pitts, que propuseram o primeiro modelo matemático para neurônio. O modelo conexionista é que originou a área de redes neurais artificiais (BITTENCOURT, 2004).

A linha simbólica (cognitiva) ou psicológica seguiu a tradição lógica e teve como seus principais defensores McCarthy e Newell. Esta linha de pesquisa iniciou-se a partir da década de cinqüenta com a utilização de estratégias lógicas com finalidade matemática. Mas somente a partir da década de setenta, esta linha de pesquisa estabeleceu-se com o sucesso dos sistemas especialistas (SE), utilizando a manipulação simbólica como corrente paradigma para a construção de sistemas inteligentes lógicos. Algumas das principais áreas de pesquisa da IA simbólica, além de SE, temos também: aprendizagem, representação do conhecimento, robótica, linguagem natural, etc (BITTENCOURT, 2004).

2.2 Tradução Automática

A tradução automática (TA) é uma das atividades que mais utiliza o conhecimento de lingüística, visto que é necessário fazer a codificação das informações extraídas de um texto de uma língua fonte (LF) para outra na língua alvo (LA). Não é surpresa saber que foi a primeira área em que se trabalhou com PLN (SANTOS, 2005).

A TA é um dos campos importantes da IA, que vem melhorando constantemente seus recursos, utilizando PLN para obter melhor qualidade nos serviços prestados.

2.2.1 Histórico da Tradução Automática

Após a segunda guerra mundial, ingleses e americanos, desenvolveram esta aplicação computacional com intuito de obter informações dos russos. Este invento então atribuído ao inglês Booth e ao americano Warren Weaver, que criaram uma calculadora com dados suficientes para fazerem traduções de palavra por palavra, sem considerar qualquer erro de ordem sintática ou lexical (GARRÃO, [entre 2000 e 2004]).

Em 1948 o inglês Richens, introduziu informações gramaticais e sintáticas da língua russa acelerando o desenvolvimento de consultas automáticas. No inicio de 1950, Weaver propõe a exploração de termos, com intuito de diminuir problemas de ambigüidade e semântica, pois acreditava-se que os circuitos lógicos eram capazes de solucionar os elementos lógicos da linguagem. Em 1954, na Universidade de Georgetown, realizaram o primeiro teste bem sucedido, entre russo e inglês, realizado por computador. O vocabulário tinha 250 palavras e seis regras sintáticas (GARRÃO, [entre 2000 e 2004]).

A década de 50 foi considerada a década do otimismo se tratando de TA, porém na próxima década, todos os projetos praticamente ficaram parados. Mais recentemente na década de 80, muitos projetos TA voltaram a ser desenvolvidos principalmente pela criação da Comunidade Econômica Européia, a explosão da informatização, processamento de línguas naturais, baseadas em gramática de análise e de geração. Devido há estes desenvolvimentos, a IA começou a ter o domínio sobre os STA, recebendo apoio de muitos países (GARRÃO, [entre 2000 e 2004]).

Nesta época já era possível descartar a possibilidade de construir um tradutor que obtivesse ótimos resultados. Portanto, na década de 80, todas as expectativas criadas para desenvolvimento de tradutores foram descartadas, ficando apenas STA que tivessem auxilio humano. Um STA é considerado aceitável, se após a tradução ele tenha uma margem de erro inferior a 20% (GARRÃO, [entre 2000 e 2004]).

2.2.2 Os processos da Tradução Automática

A maior parte das informações utilizadas em LN, depende das informações que estão presentes na estruturação dos dados de entrada, atendendo principalmente os fatores que são

considerados relevantes. Isso significa que é necessário extrair dados sobre um determinado assunto para depois alimentar uma base de dados sobre um determinado assunto, em vez de criar manualmente um repositório de informações. Este é um dos objetivos mais antigos do PLN (SANTOS, 2005).

Os STA ou tradução por máquina é a conversão de textos de uma LF para uma LA, feita totalmente ou parcialmente por processo automático. Este é um processo que vem despertando grande interesse por pesquisadores na produção de *softwares* que melhorem o desempenho da TA.

Esse processo se mostrou útil em diversas tarefas, principalmente em:

- Tradução bruta: seu objetivo é apenas obter o significado de uma sentença avaliada, sem se preocupar com as questões gramaticais, desde que, o significado da sentença seja claro;
- Tradução de origem restrita: são traduções de textos, cujo conteúdo é severamente restrito a um número limitado de recursos. São utilizados para tradução de termos altamente regulares e que não sofrem mudanças constantes nos termos utilizados, por isso são obtidos ótimas traduções, como por exemplo, o sistema (TAUM-METEO), que traduz relatórios sobre o tempo do inglês para o francês;
- Tradução pré-editada: são traduções onde é feita uma edição prévia do conteúdo, antes de ser submetido a um STA. Esta técnica é bastante utilizada para traduzir manuais e documentos legais de empresas que vendem seus produtos para diferentes países;
- Tradução literária: são traduções onde todos os sentidos do texto original têm que ser preservados, neste caso a TA, está além do estudo da arte.

O problema é que o mundo subdivide-se em idiomas distintos. Para fazer uma tradução fluente, a máquina ou o homem, primeiramente deve ler todo o texto, entender a situação à qual é referenciada, e encontrar um texto que corresponderá à originalidade ou semelhança do idioma alvo.

A formalização da TA, é compreendida hoje pela diversificação de práticas de pesquisas e projetos em desenvolvimento, sendo que muitos deles afastaram-se do programa original de domínio. Os sistemas originais de TA foram redefinidos progressivamente como sistemas de tradução auxiliados por humanos (*human-aided machine translation*). A maioria

dos desenvolvedores redimiram a produzirem tradutores totalmente automáticos, passando a trabalhar com objetivos bem mais modestos (MARTINS, 2003).

Ao invés de produzirem tradutores equivalentes na LA, para enunciados produzidos na LF, estes sistemas, pautam-se em alternativas organizadas em quatro eixos:

- Redefinição da língua de partida: passa pelo processo de complexificação, reduzindo a um pequeno subconjunto normalizado da variedade real, que é muito poluída e de difícil tratamento computacional;
- Pela especialização do tratamento do discurso: trata de uma seleção temática, uma forma de composição e estilo, com um conjunto limitado de variações, cujo reconhecimento é necessário para o treinamento da ferramenta de tradução;
- Pelo abandono sistemático de produzir resultados definitivos: usados pelas traduções cruas, recuperando apenas movimentos mais mecânicos e menos inteligentes do processo de tradução;
- Pela redução da tradução automática, ou tradução grosseira: é um mecanismo de triagem, sinalizando que o texto contém ou não assunto de interesse, sendo melhor ser encaminhado a um tradutor humano.

Para todos estes casos, a dificuldade para os STA, é o processo de análise e interpretação do enunciado na LN. Pois a TA é extremamente sensível e dependente do conteúdo semântico da sentença (MARTINS, 2003).

As estratégias adotadas salientam a abordagem de pelo menos três estratégias que predominam: tradução baseada exclusivamente em conhecimento lingüístico (*Language-Based Machine Translation* – LBMT), ou melhor, em dicionários e gramática; a tradução baseada em conhecimento (*Knowledge-Based Machine Translation* - KBMT), utiliza dicionários, gramáticas, enciclopédias e bases de conhecimento; e tradução baseada em exemplos (*Exemple-Based Machine Translation* – EBMT), que utilizam dicionários, gramáticas e corpora (corpus) (MARTINS, 2003).

Os dois primeiros casos (LBMT e KBMT) são baseados em modelos de tradução em regras ou conhecimento lingüístico, tem menor custo e é adequado a sistemas mais genéricos. Já o (EBMT), envolve recursos dispendiosos, produz respostas mais exatas, sendo indicado para sistemas especializados (MARTINS, 2003).

A TA é a conversão de textos de uma LF para uma LA, feita totalmente ou parcialmente por processo automático. Este é um processo que vem despertando grande

interesse por pesquisadores na produção de *softwares* para que melhorem o desempenho da TA.

2.2.3 Sistemas de Tradução Automática

Diversos STA, tornaram-se produtos comerciais tais como Translator Pro, Tradunet, Globalink L&H Power Translator Pro ou de distribuição gratuita como Systran, FreeTranslation, E-Translation Server, Amikai, Alta Vista e Enterprise Translator Server, são todos sistemas considerados de tradução preliminar, pois não são capazes de fazer uma tradução refinada ou seja coerente. As traduções obtidas por estes sistemas possuem freqüentemente erros e imperfeições nos resultados obtidos (VIEIRA e STRUBE, [entre 2002 e 2004]).

Diferentes metodologias podem ser aplicadas na tradução automática, entre elas podemos citar os sistemas transferenciais, sistemas interlínguas e os sistemas diretos (memória). Os STA baseados em transferências mantém um banco de dados com regras de traduções. Se essas regras coincidem, são feitas as traduções diretamente. Para isto, estes sistemas precisam efetuar análise sintática na LF e através de suas regras de transferência sintática, podem construir sua representação sintática na (LA), (VIEIRA e STRUBE, [entre 2002 e 2004]).

Estas transferências podem ocorrer tanto no nível léxico, sintático ou semântico. Por exemplo, uma regra estritamente sintática mapeia do inglês (adjetivo substantivo) para o francês (substantivo adjetivo), ou seja, a posição dos substantivos são invertidas com os adjetivos (RUSSELL e NORVIG, 2004).

Os STA, possuem variações na forma em que analisam seus textos, como os sistemas interlínguas que trabalham com uma representação intermediária entre a LF e a LA que, em principio podem ser utilizadas na tradução de quaisquer idiomas (VIEIRA e STRUBE, [entre 2002 e 2004]).

Alguns destes sistemas analisam o texto de entrada por inteiro em uma representação de interlínguas, gerando posteriormente sentenças na língua alvo, a partir dessa representação. Isso é difícil, devido à dificuldade de compreensão da linguagem como um subproblema, que é lidar com uma interlíngua. Este processo não é seguro, porque se falhar não haverá nenhuma saída. Mas tem como vantagem em que o sistema não depende do conhecimento de duas linguagens ao mesmo tempo, significando a possibilidade de construir

sistemas de interlínguas para realizar traduções entre *n* línguas (RUSSELL e NORVIG, 2004).

Quando uma sentença é transferida diretamente para outra é chamada de método de tradução direto baseado em (memória), buscando uma correspondência direta entre as palavras, porque são baseadas na memorização de grandes conjuntos de pares de línguas. O método de transferência é considerado robusto, porque sempre existe um retorno de pelo menos algumas das palavras (RUSSELL e NORVIG, 2004).

2.2.4 Traduções diretas e indiretas

Em traduções diretas a própria LA é considerada um instrumento para análise da LF, não havendo estágio intermediário entre a LF e LA. Os vocabulários das sentenças de entrada são convertidos automaticamente para LA através de dicionários bilíngües e processamento morfológico. Quando as equivalências lexicais são geradas na LA, estes ítens lexicais são reordenados para produzirem resultados mais aceitáveis. Não havendo processamento sintático das sentenças na LF. Os resultados são bastante ruins, devido sua simplicidade (MARTINS, 2003).

Em traduções indiretas são desenvolvidas formas para representação intermediária entre a LF e LA. Esta representação é dependente das línguas envolvidas, no sentido de criar uma interface especifica ou também ser independente da LF e LA, onde ela procura organizar-se como se fosse outra língua artificial, autônoma e neutra, porém de maneira a adaptar-se a tradução automática (MARTINS, 2003).

2.2.5 Conflito paradigmático em Tradução Automática

Existem diversos conflitos que se instalam tanto do lado da TA como dos estudos da tradução, pois os processos que são utilizados para tradução de uma língua para outra, permite perceber diversos pontos de divergências.

Podemos referenciar sobre a TA e não aos estudos da tradução, de que a língua é um código onde podem ser definidos por recursos léxicos e uma gramática contendo as regras de combinação (sintaxe), referenciação (semântica) e de uso (pragmática) do vocabulário utilizado. Para validação de uma descrição lingüística que figure em um sistema de conhecimento, é necessário estabelecer relações de suas partes constituintes. Podemos admitir a existência de um contexto nulo, a composição dos enunciados lingüísticos, estabilidade das

unidades lingüísticas, etc. Sobre a importância destes pressupostos, podemos dizer que o condicionamento de que o contexto zero, revela-se a única estratégia viável para tratamento dos fatos da língua, até que seja possível representar para a máquina o conhecimento do mundo do homem (MARTINS, 2003).

Ao definir a língua como um código, a TA pode subscrever a idéia da possibilidade de elaborarmos um modelo teórico que não se comprometa com a realidade psicológica ou sociológica do falante.

Como código, a língua pode ser vista como a possibilidade de multiestratificação em níveis de análises distintos, podendo cada um ser descrito independentemente. As funções não referenciadas representam um excedente teórico e necessário para explicar uma série dos fenômenos lingüísticos, podendo expurgar as descrições mínimas da língua, privilegiando apenas as estruturas de significado. Esta opção torna nítida para as unidades de análise da TA, quando as sentenças ou frações das sentenças, ou mesmo palavras isoladas, cujo, a produção de seu significado pode operar mecanismos estritamente composicionais, os quais o significado do todo é uma função direta das partes que a compõem. Geralmente os tradutores automáticos operam em nível de sentença, não observando, senão muito eventualmente, qualquer fenômeno macrotextual, implicando que os textos são definidos como um aglomerado de sentenças, que não reconhecem a textualidade das diversas partes de um texto (MARTINS, 2003).

2.3 Processamento de Linguagem Natural

A linguagem é todo sistema do qual é possível extrair significado capazes de estabelecer a comunicação entre os diversos sistemas existentes, por exemplo, comunicação entre humanos ou não, naturais ou artificiais, verbais ou não verbais. Para todos os tipos de linguagens sempre é encontrado uma forma para produzir um significado que formaliza a comunicação entre os diversos sistemas. O PLN é uma forma artificial desenvolvida com a finalidade de aprender através da LN, artifícios que possam ser utilizados com recursos computacionais, para podermos desenvolver programas capazes de tratar os problemas relacionados à LN.

Para se desenvolver recursos capazes de tratar destes problemas, é necessário descobrir como os homens se comunicam, para modelar-se estes processos utilizando PLN. Para isso, essa possibilidade não é tão simples assim, pois ainda não há conclusões científicas definidas de como o cérebro humano trabalha o entendimento das línguas humanas. Existem

duas linhas de pesquisa que delineiam esta área, a Lingüística (enfoque da gramática gerativa transformacional) e a Psicologia Cognitiva (procurando detectar unidades temáticas como: metas, intenções, conseqüências e causas, existentes em textos para realizar o processamento).

A linguagem é considerada uma estrutura, onde os elementos de um conjunto estão devidamente ordenados, e que, a função de cada um é definida em relação aos demais (ARARIBÓIA, 1988).

O PLN nasceu junto com o computador. A partir de 1946, começou o interesse por TA, devido à 2ª Guerra Mundial, tendo como principal objetivo obter informações científicas dos soviéticos, onde vários pesquisadores começaram a aprofundar suas pesquisas utilizando o tratamento da linguagem natural informatizada (FILHO, 2004).

O PLN, visto como uma área da IA e da Lingüística, surgiu com o propósito de estudar a linguagem natural (LN), com o objetivo de usá-la como um meio de fazer a comunicação entre os homens e os computadores. Do ponto de vista da Ciência da Computação, as principais metas do PLN é construir sistemas computacionais que facilite esta comunicação de forma efetiva, via LN, através dos conhecimentos lingüísticos implementados sobre aplicações computacionais (SPECIA, 2000c).

Em 1968, com a publicação da obra de Arthur C. Clarke – 2001 Uma Odisséia no Espaço, vários programas surgiram com o intuito de serem utilizados para a compreensão da linguagem natural. Durante a década de 60, os computadores já eram capazes de interpretar algumas questões em inglês referentes a alguns assuntos, por exemplo, medicina, álgebra e relações de parentescos (OLIVEIRA, 2004).

Nesta época quatro categorias de programas tinham como meta o PLN:

- Programas, (por exemplo: BASEBAL, SAD-SAM, STUDENT e ELIZA) tinham como objetivo, utilizar uma base de dados para fazer a tradução de uma frase simples na qual muitos dos problemas da linguagem natural eram ignorados;
- Sistema como o (PROTO-SYSTHEX1) eram necessários para armazenamento de textos, para que durante uma consulta, frases e palavras pudessem ser recuperadas. Como qualquer texto armazenado podia cobrir qualquer assunto, este sistema era considerado semanticamente fraco;
- Sistemas de lógica limitada, (por exemplo: SIR, DEACON, TLC e CONVERSE) utilizavam uma base de dados para denotarem de forma mais formal uma tradução. A intenção deste sistema era fazer deduções a partir dos dados armazenados;

- Sistemas com base de conhecimento, (por exemplo: LUNAR e o SHRDLU) tinham conhecimento específico de alguns assuntos, para a compreensão de frases de entrada, com vários poderes dedutivos.

Entre os sistemas citados, destacou-se o ELIZA, criado por Joseph Weizenbaum, em 1966. Este sistema passava por um psiquiatra que conversava com o usuário, utilizando truques semânticos, onde não tinha a menor compreensão do assunto tratado (OLIVEIRA, 2004).

Apesar da sobrevivência de alguns destes projetos, a década de 60 foi um verdadeiro caos aos projetos de traduções automáticas, pois muitos tiveram todas suas verbas governamentais cortadas, porque na maioria deles, suas aplicações teóricas não correspondiam ás aplicações práticas, remanescendo apenas três projetos em 1973. É importante frisar que muitos destes esforços iniciais, que não emergiram, contribuíram para o desenvolvimento computacional, lingüística e IA, tomando como ponto de partida as falhas detectadas em programas anteriores (ALFARO, 1998).

2.3.1 Etapas do Processamento de Linguagem Natural

Esta etapa consiste em esclarecer as fases necessárias em que um sistema computacional precisa para interpretar uma sentença em LN, sendo as seguintes: análise morfológica, análise sintática, análise semântica, análise do discurso e processamento pragmático (SPECIA, 2000c).

2.3.1.1 Análise Morfológica

A análise morfológica (AM) estuda a forma de como as palavras e os grupos de palavras constituem os elementos que expressam uma língua. Todo o conhecimento sobre a estrutura da palavra é tratado especificamente pela morfologia. Podemos verificar que certas palavras não podem ser quebradas como “árvore”, mas isso pode ocorrer com as palavras como “árvores, arvorezinhas ou impossível”. Cada uma das unidades que constituem as palavras são denominadas morfemas. Cada um destes constituintes pode ser independente como em “árvore” ou dependentes como no caso do sufixo “zinhas em arvorezinhas” ou do prefixo “im em impossível” (VIEIRA e STRUBE, [entre 2002 e 2004]).

Em morfologia, além de estudar a estrutura das palavras, também estudam a classificação delas em parte do discurso (*part-of-speech*, ou POS) comentado no ítem (2.5).

O analisador morfológico tem a função de identificar em um texto, palavras ou expressões isoladas. Estas palavras são classificadas de acordo com sua categoria gramatical especificada em LN.

A AM tem por objetivo dividir o texto em *tokens*, que são elementos essenciais para identificação dos termos prefixos, sufixos e formação das raízes que compõem a língua em questão. Como resultado, teremos palavras isoladas e descritas juntamente com seus componentes. Para que o reconhecimento destes *tokens* seja possível, é necessário que estes sejam reconhecidos pelo domínio da LF. Para fazer uma busca destes *tokens* é preciso que o padrão permita gerá-lo dentro de uma estrutura de armazenamento do léxico (SPECIA, 2000c).

Se um *token* analisado não estiver dentro da estrutura léxica analisada, será retornado um erro, interrompendo o processo de compreensão da sentença. Veja no exemplo abaixo como um sistema operacional compreende uma frase em LN:

"Eu quero imprimir o arquivo.init do João"

Em AM, são realizados os seguintes procedimentos:

- Primeiro separa-se a expressão "do João" no substantivo próprio "João" na preposição "de" e no artigo "o".
- O passo seguinte é reconhecer a seqüência ".init" como um tipo de extensão de arquivo que está funcionando como adjetivo na sentença.
- Finalmente são atribuídas as categorias sintáticas a todas as palavras da sentença, pois as interpretações dos afixos podem estar dependentes de sua categoria sintática (SPECIA, 2000c).

Em uma sentença gramaticalmente válida, qualquer palavra substituída por outra do mesmo tipo, ainda será válida, por exemplo, os substantivos, pronomes, verbos, etc. Dentre os mesmos tipos de palavras, há vários grupos que caracterizam o comportamento dos vocábulos das linguagens. Desta forma, a morfologia trata as palavras de acordo com sua estrutura gramatical, flexão e classificação (OLIVEIRA, 2004).

O autômato finito é uma forma utilizada para fazer o reconhecimento da análise morfológica de uma frase. Porém, para uma grande quantidade de vocábulos, não são eficientes, daí a utilização de autômato determinístico acíclico.

É importantíssimo, o emprego de um analisador morfológico, para compreensão de uma frase ou texto, pois, para construir uma estrutura coerente que venha a condizer com a

sentença analisada, é necessário que o significado de cada palavra seja conhecido e compreendido.

2.3.1.2 Análise Sintática

Para reconhecer um objeto, é necessário verificar se a organização se seus vários componentes estão corretos. Esta verificação é chamada de análise sintática, pois em Grego significa montagem. A análise sintática pode ser definida como o estudo minucioso da montagem e do modo de colocar as várias partes de um todo em posições adequadas (ARARIBÓIA, 1988).

Para realizar o tratamento sintático das palavras em PLN, existem ferramentas especificas capazes fazer este trabalho, a principal delas é a Notação de Regras Gramaticais (NRG), disponível em *Prolog*. Estas regras especificam como escrever um conjunto de regras denominado "gramática". Este conjunto de regras, especifica como se deve realizar a análise sintática das expressões da linguagem (ARARIBÓIA, 1988).

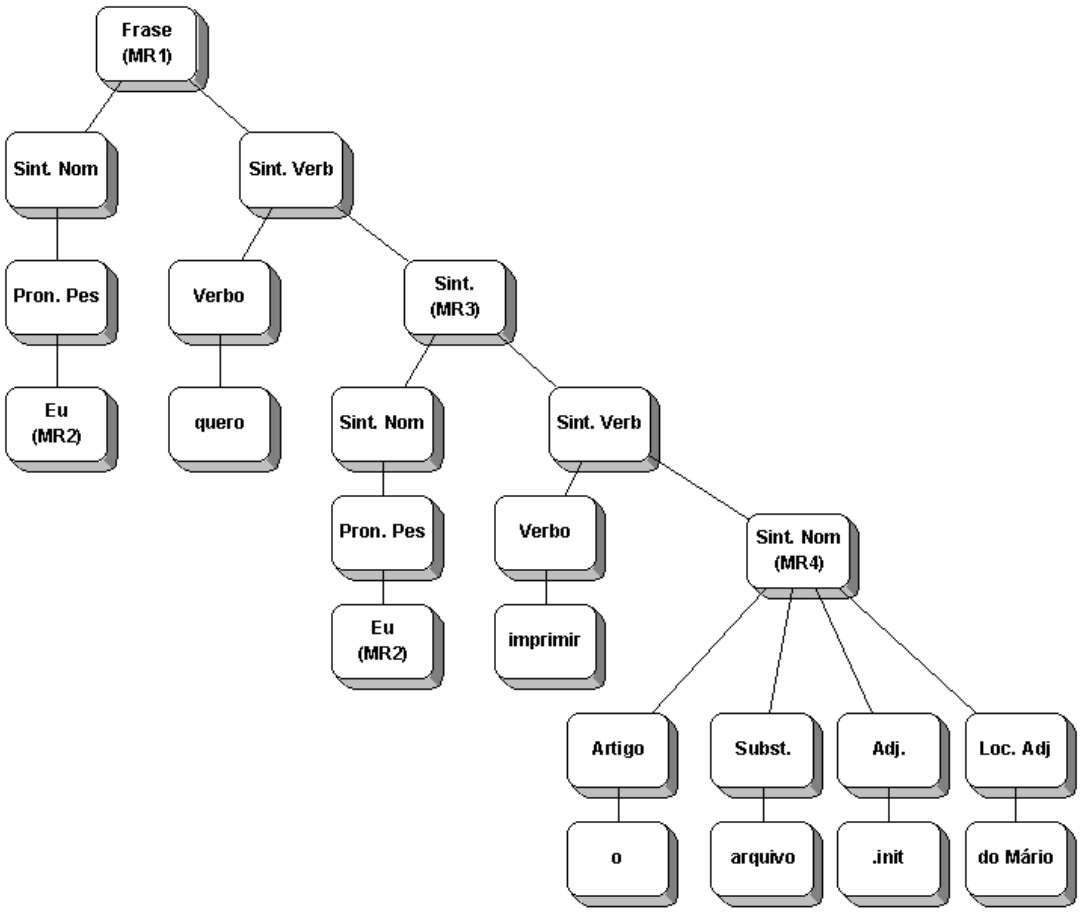
É preciso lembrar, que as palavras que constituem uma frase são classificadas seguíndo certas categorias gramaticais. Algumas destas categorias são: artigos definidos, artigos indefinidos, nomes próprios, nomes comuns, verbos, etc. (ARARIBÓIA, 1988).

O analisador sintático utiliza a gramática da LN a ser analisada e cruza com as informações do analisador morfológico, construindo uma árvore de derivação para a sentença analisada, mostrando o relacionamento das palavras entre si. Esta árvore de derivação tem como função converter a lista de palavras que formam a frase em uma estrutura hierárquica representativa de cada palavra da frase separadamente (OLIVEIRA, 2004).

Cada uma das unidades, que foram geradas, correspondem aos componentes que serão atribuídos aos significados quando realizada a análise semântica. A função da análise sintática é diminuir a quantidade de componentes que a semântica pode analisar, reduzindo a complexidade do sistema, considerando que o processamento sintático utiliza menos recurso computacional do que o processamento semântico (SPECIA, 2000c).

A representação da gramática de uma linguagem é feita através de um conjunto de regras de produção, ou seja, são regras que tem a finalidade de definir as estruturas legais de uma linguagem que será utilizada no processo de análise. Quando estas regras são analisadas pelo analisador, geralmente ocorrem duas operações: o casamento dos "componentes da sentença" com as regras da gramática e a "construção da estrutura" que é a combinação dos componentes encontrados na análise sintática (SPECIA, 2000c).

A estrutura mais simples e a "árvore sintática" onde cada nó desta árvore corresponde a uma palavra da sentença ou um símbolo não terminal da gramática, e cada nível é correspondente a uma regra gramatical, onde são feitas as verificações dos significados de cada palavra, para verificar se estas estão dispostas corretamente na sentença, e que estão atendendo às regras gramaticais da LN utilizada, veja o exemplo na FIGURA 1 (SPECIA, 2000c).



Fonte: OLIVEIRA (2004)

FIGURA 1: Representação de um modelo de árvore sintática.

Em qualquer LN é preciso levar em consideração os seguintes sintagmas: termos essenciais, termos integrantes e termos acessórios, para que a oração analisada seja entendível pelo analisador sintático (OLIVEIRA, 2004).

Segundo Oliveira (2004), um dos maiores problemas de PLN, é quando da transformação de uma frase potencialmente ambígua noutra não ambígua, a qual será utilizada por um sistema de TA.

A utilização de linguagens formais em estudos da análise sintática tem um conceituado sucesso com o uso das gramáticas regulares, gramáticas livres de contexto e gramáticas sensíveis ao contexto:

- Gramáticas regulares: são gramáticas simples e bastante reconhecidas, porém apresentam um poder de expressão bastante limitado.
- Gramáticas livres de contexto: são extremamente úteis para descrever a gramática em linguagem natural, pois são mais poderosas que as gramáticas regulares e são representáveis em linguagens com maior grau de complexidade. Porém, possuem alto grau de complexidade para representar concordância simples, como por exemplo, concordância entre verbo e o sintagma nominal. Este tipo de gramática não é suficientemente poderosa para descrever o gênero de uma linguagem natural. Utiliza-se então o *Definite Clause Grammar* (DCG), disponível em *Prolog,* para analisar sentenças.
- Gramáticas sensíveis ao contexto: todos os problemas de dependência não solucionados pelos métodos anteriores, aqui são resolvidos. Sua utilização depende do reconhecimento. Decidir se uma sentença pertence ou não a uma gramática sensível ao contexto, torna-se um problema exponencial dependendo do tamanho da sentença analisada, ou seja, sua implementação em sistemas computacionais é extremamente complexa. Muitas pesquisas propuseram utilizar uma gramática que seja intermediária à gramática livre de contexto e a sensível ao contexto. Facilitando assim a modelagem das dependências de um modelo computacional viável.

2.3.1.3 Análise Semântica

A expressão análise semântica (AS) tem origem grega e significa dissecação (estudo minucioso) de significado. Tendo como objetivo descobrir estratégia para representar o sentido das frases ou sentenças.

Em PLN, podemos imaginar a existência de uma linguagem capaz de pensar, sendo um pouco diferente do que usamos para falar e ouvir, mas com grande capacidade para sofrer inferências, recordações e outras formas de atividade mental (ARARIBÓIA, 1988).

Para que ocorra o processamento semântico é necessário definir em que linguagem está sendo proposta o mapeamento, onde está a linguagem final para a representação do

significado, ou seja, a linguagem objeto, da qual depende o que será feito com os significados depois de construídos.

Há duas famílias de linguagem-objeto utilizadas em PLN: (a) quando a LN é considerada um fenômeno por si próprio, por exemplo, um sistema que tem como finalidade ler um conjunto de sentenças e depois ter a capacidade de responder perguntas sobre estas sentenças; (b) quando a LN é usada na interface com outro programa (como em SE), a linguagem-objeto precisa ser uma entrada legal neste outro sistema (SPECIA, 2000c).

Após a definição da linguagem a ser utilizada, o processo de análise semântica irá procurar em um léxico (dicionário), para extrair o significado das palavras que constituem as estruturas repassadas pelo analisador sintático (*parser*). Após este passo, as sentenças serão processadas como um todo, para verificação das regras semânticas, de acordo com a linguagem em questão (SPECIA, 2000c).

O analisador semântico tem como função verificar o sentido da estrutura das palavras que foram reagrupadas pelo analisador sintático, junto à árvore de derivação (árvore sintática), construída com as informações do analisador morfológico e sintático. Pois existem muitos morfemas que compõem uma palavra que podem mudar o sentido da frase, como por exemplo, "mercado, hipermercado", a ambigüidade (lavar, em "lavar a casa", "lavar o carro" ou "lavar a roupa") ou diferenciar o significado do sentido ("casa", "minha casa").

Segundo Oliveira (2004), a AS é feita para dar sentido às estruturas das palavras que foram reagrupadas pelo analisador sintático, uma vez que o AM já identificou cada uma delas individualmente.

A semântica é dividida em duas formas, a semântica léxica e a semântica gramatical: a semântica léxica busca uma representação conceitual para modelar o sentido. Para elaborar esta representação pode-se decompor a semântica em unidades léxicas (primitivas ou de traços semânticos). Para isso, compreender o relacionamento entre todas as palavras de uma frase é tão importante quanto saber o significado isolado de cada uma delas.

A semântica gramatical tenta descrever o sentido lógico de uma frase. Como não há uma correspondência instantânea entre a sintaxe e a semântica, a mesma estrutura sintática pode gerar várias representações de semântica, ou seja, sentidos diferentes na mesma frase. Por exemplo, "uma professora de capoeira pernambucana", pode referir-se a uma pessoa nascida em Pernambuco, a qual ensina capoeira, ou a uma pessoa que ensina capoeira no estilo em que é praticada em Pernambuco (OLIVEIRA, 2004).

O processo de uma ou mais representações semânticas em árvores de derivação é chamado de análise semântica, podendo ser feita posteriormente ou durante a geração das árvores.

Um sistema com capacidade de interpretação semântica avançado, conseguiria solucionar a maioria dos casos de ambigüidade encontrados nestas sentenças, visto que, não há regra sintática que solucionaria estes problemas. Um sistema com tal capacidade teria um grande problema, que é o alto processamento de informações, para que pudesse obter a melhor acepção na tradução, para isso, a consulta em outras bases de dados seria indispensável se necessário, com a finalidade de encontrar a resposta satisfatória (OLIVEIRA, 2004).

2.3.1.4 Análise do Discurso

A análise de discurso em uma sentença em estudo pode depender ou não do significado da sentença que a antecede, influenciando no significado da sentença sucessora. Com isto, vê-se a necessidade de que todo o contexto de um conjunto de sentenças deve ser analisado, para que se possa compreender o sentido da sentença, principalmente quando se trata de textos ou diálogos.

Para a frase utilizada anteriormente “Eu quero imprimir o arquivo.int do João”, já se sabe do que se trata esta sentença, agora se faz necessário a integração do discurso, para especificar que indivíduos estão sento citados na sentença. Para que isso possa ser esclarecido, precisa-se de um modelo de contexto que permita descobrir qual usuário fez o pedido (aquele que digitou “Eu”) e quem é o usuário “João”. Sabendo-se, quem é João, pode definir qual é o arquivo (ou arquivos) com extensão .init, cujo o dono é João. Este tipo de reconhecimento só será possível, se o programa for capaz de compreender diversas sentenças, percorrendo uma grande base de conhecimento ou a fortes restrições de domínio do discurso, de maneira a diminuir a base de conhecimento (SPECIA, 2000c).

2.3.1.5 Processamento Pragmático

A pragmática é uma medida necessária a ser tomada para fazer a interpretação geral de toda a sentença analisada, e não mais analisar o significado de suas partes separadamente do ponto de vista gramatical e léxico. Ou seja, um significado, tal como o ser humano seja capaz de entender da mesma forma que se ele estivesse lendo ou ouvindo uma sentença, e

nela absorvesse elementos que não estão representados unicamente nas unidades e nas relações semânticas. Apesar do conteúdo dito "literal", torna-se necessário a ligação das sentenças entre si, de maneira que exista uma coerência de acordo com a situação e a condição do enunciado. Por exemplo, a sentença: "*o professor disse que duas semanas são o tempo necessário para resolver este problema"*. Para uma compreensão literal, os próprios mecanismos apresentados anteriormente, poderiam solucionar o problema até mesmo em uma tradução, mas em uma situação mais aprofundada, exigiria a que problema o professor estaria sendo referenciado, já que o problema pode ter sido a própria construção da sentença (VIEIRA e STRUBE, [2002 e 2004]).

A pragmática possui dois pontos importantes: o relacionamento entre as frases, (onde a construção de uma nova representação de um texto, depende da representação da nova frase apoiada na frase precedente), e o contexto, (depende da situação e a condição que ocorre o enunciado). À medida que novas sentenças vão sendo enunciadas, vão aumentando o universo à suas referências, unindo as já existentes à base de dados (VIEIRA e STRUBE, [2002 e 2004]).

Essa análise é importante porque uma sentença após ser reconstruída, pode não ter na sua estrutura nenhum dado de interesse de interpretação, por exemplo, nesta sentença, "*Bela camisa Fernandinho*", podem ser identificados dois sentidos distintos, vai depender do modo de interpretação.

A pragmática tem uma abrangência muito maior do que simplesmente analisar uma sentença isolada, ela passa a analisar todo o discurso, relacionando a língua a seu uso. Pode-se entender como discurso o texto ou a fala, que são divididas em unidades menores denominadas sentenças. Cada uma dessas sentenças, podem ter significados que poderão gerar sentidos diferentes que possam comprometer totalmente o significado do texto analisado.

A etapa final do processamento pragmático é fazer a tradução, sempre que ocorrer a necessidade de uma representação, com base no conhecimento de um comando especifico a ser executado pelo sistema. Todos os tipos de processos são importantíssimos para sistemas que compreendem a LN. Porém, nem sempre estes programas são escritos com estes componentes, ocorrendo à omissão de dois ou mais processos, facilitando a criação de um subconjunto restrito de um língua, porém dificultando a ampliação de uma abrangência externa (OLIVEIRA, 2004).

2.3.2 Processamento Simbólico da Língua Natural

O processamento simbólico é o resultado do processamento, quando são aplicados conceitos da semiótica (uso de signos) no desenvolvimento de *softwares*. Podemos dizer que, qualquer *software* desenvolvido com este conceito utiliza processamento simbólico ao invés de processamento algorítmico (SEMIOTIC, 2005).

O processamento simbólico é voltado à manipulação de regras rígidas, tal como a gramática de qualquer LN. Como conseqüência deste fato, muitas das principais aplicações de programação simbólica possuem tais características como: raciocínio baseado em objetivos, análise de causas e análise gramatical. Por apresentar tais características, o processamento simbólico é o processo mais apropriado para este tipo de trabalho, visto que também existem abordagens voltadas a redes neurais (OLIVEIRA, 2004).

Apesar da programação simbólica possuírem estas características voltadas ao PLN, as redes neurais não podem ser consideradas ferramentas apropriadas a tal tarefa, devido as limitações existentes em processamento dependente de contexto. Sabe-se que as redes neurais são extremamente eficientes no reconhecimento de objetos isolados, cujo, reconhecimento já tenha sito treinado anteriormente. Mas se tratando de LN, é importante que sejam aplicadas regras dependentes de contexto, o que pode ser encontrado em processamento simbólico (OLIVEIRA, 2004).

2.4 Problemas de Ambigüidade

A ambigüidade entre as línguas é um problema encontrado tanto na tradução humana (TH) quanto na TA. Porém, na TA este problema se mostra gravíssimo, pois não existe um interpretador capaz de analisar os diferentes sentidos expressados por uma palavra ou sentença ambígua, o que não acontece na LN, onde o ser humano tem a capacidade interpretativa para definir o sentido muito próximo do contexto original da sentença ou palavra expressa na LF entre os interlocutores.

Há vários níveis de ambigüidade encontrada na tradução, seja na TA ou TH, por exemplo: ambigüidade léxica, ambigüidade sintática, ambigüidade semântica, ambigüidade contextual e pragmática (SPECIA e NUNES, 2004b).

A ambigüidade léxica (AL) ocorre quando em uma LA, uma palavra possui mais de um significado que pode ser usada na tradução de uma palavra na LF. Quando o significado desta palavra pertencer à mesma categoria gramatical, passa a ser chamada de ambigüidade

léxica de sentido (ALS), pois o significado muda apenas nas variações de sentido da LA. A opção de escolha por uma das possíveis palavras, pode não condizer com a acepção esperada, dando origem a proposições semanticamente incorretas e distintas. Com isso o resultado esperado pela tradução pode constituir uma sentença com um novo significado.

Conforme Specia e Nunes (2004b), o problema se mostra mais grave quando o significado encontrado possui existência de relações interlexicais, principalmente a homonímia, polissemia e a sinonímia.

A polissemia não é um defeito da língua como poderia se pensar, insistindo em dar os mesmos nomes a coisas diferentes. A polissemia é um recurso fundamental da linguagem humana, que na sua ausência, não funcionaria eficientemente. Não seria possível nomear separadamente cada coisa. A funcionalidade da polissemia é dar fundamentos a objetos nunca vistos antes, relacionando-os com outros já existentes. Pois, é assim que a linguagem e a mente humana trabalham, tentando assimilar o desconhecido a uma categoria já conhecida. A polissemia flexibiliza a linguagem humana à necessidade de exprimirem todos os aspectos da realidade. Conseqüentemente, a maioria das palavras e sentenças são polissêmicas em algum grau, sendo assim uma inconveniência, mas necessária (PERINE, 2002).

A polissemia é responsável pelo sentido das palavras, relacionando o significado da palavra de acordo com o contexto da frase. Numa tradução, uma palavra polissêmica na LF pode corresponder a dois ou mais significados na LA relacionados entre si. Por exemplo, à palavra do inglês "*know"* que pode corresponder a duas palavras no português, "saber" e "conhecer". Esta diferença de significado pode mudar o sentido da tradução, pois os tradutores ainda não são capazes de identificar qual dos significados corresponde ao contexto da frase, fazendo com que seja utilizada somente a forma mais comum (PERINE, 2002).

A homonímia entre duas línguas, para uma mesma palavra na LF pode ter dois ou mais significados na LA, que não estão relacionados entre si, por exemplo, a palavra do inglês *honey* que pode significar "mel" ou "querido" na língua portuguesa (PERINE, 2002).

A sinonímia ocorre quando uma palavra tem sentido idêntico ou quase idêntico. Geralmente isso ocorre com palavras onde qualquer um dos sinônimos escolhidos corresponde à palavra dada. No entanto, os sinônimos perfeito ou idênticos não representa um grande problema para os tradutores, já que a ocorrência deste é rara, por exemplo, "mãe" e "mamãe" são consideradas sinônimos, mas não é comum dizer "Maria é mamãe de Aline". O problema ocorre quando uma palavra LF corresponde a vários sinônimos na LA, sendo necessário escolher um dos possíveis sinônimos, por exemplo, a palavra do inglês "*dog"*, que

na tradução para o português é preciso escolher entre os sinônimos “cachorro” e “cão” (PERINE, 2002).

O problema de AL na tradução pode ser classificado como ambigüidade categorial (AC) e ambigüidade lexical de sentido (ALS). Não falaremos sobre AC, já que não é um grande problema para TA devido aos avanços do PLN. Já a ALS é um problema que se mostra gravíssimo em TA, pois exige uma análise bem minuciosa da semântica das palavras (SPECIA e NUNES, 2004a).

Para solucionar o problema de ALS é necessário fazer a desambiguação automatizada, sendo necessário à incorporação de um módulo de desambiguação léxica de sentido (DLS), constituído basicamente nas seguintes etapas:

- Especificar todas as palavras ambíguas de um texto a ser traduzido ou grupos de palavras, com base na similaridade sintático-semântico em um *corpus*, etc;
- Identificar todos sentidos possíveis para cada palavra ambígua, utilizando recursos como *corpus* paralelo, dicionários bilíngües, tradutores humanos, etc;
- Definir meios para atribuir o sentido mais apropriado a cada ocorrência da palavra, criando um modelo de DLS;
- Incorporar este modelo de DLS a um STA.

2.5 Etiquetagem (*Part-of-speech taggers* - POS *tagging*)

O etiquetador gramatical *Part-of-speech taggers* (ou POS *tagging*), é o sistema responsável pela identificação de cada um dos itens lexicais encontrados em uma sentença. Ele é responsável por decidir em qual categoria gramatical a palavra deverá pertencer, de acordo com a posição ocupada pela palavra na frase. Nesta etapa o etiquetador está preparando o texto para entrada na aplicação, deixando a questão da ambigüidade para uma outra etapa (VIEIRA e STRUBE, [entre 2002 e2004]).

Podem ser incluídas às etiquetas, ou parte do discurso: substantivo, verbo, pronome, preposição, advérbio, conjunção, particípio ou artigo. Estas etiquetas deverão variar de acordo com a aplicação a qual o texto etiquetado será utilizado. Por exemplo, o *corpus Penn Treebank* utiliza 45 etiquetas, já o *Brown Corpus* utilizam 87 etiquetas, sendo estas, duas importantes coleções de textos da língua inglesa, pois tratam de dois importantes *corpus* de sentenças corretamente analisados e marcados (VIEIRA e STRUBE, [entre 2002 e 2004]).

Os etiquetadores são processos que não possuem a mesma eficiência do analisador léxico-morfológico convencional, mesmo assim a quantidade de informações disponibilizadas é muito grande, fazendo com que este processo de analise seja bastante difundido. O fato de se poder saber a classificação gramatical de uma determinada palavra, por exemplo, um pronome possessivo, facilita a prever sobre que palavras podem sucedê-las.

O processo de etiquetagem tem como finalidade ser um assinalador de marcador de classe gramatical de cada palavra, em um *corpus*. Este processo é semelhante a "*tokenização*" em linguagem de programação. A etiquetagem trabalha com um grande número de situações ambíguas, porque estamos tratando de LN. A etiquetagem tem como ponto de entrada os itens lexicais, em um conjunto específico de etiquetas. A saída terá a melhor etiqueta associada aos itens lexicais analisados. O processo de etiquetagem limita-se exatamente em solucionar o processo de ambigüidade (VIEIRA e STRUBE, [entre 2002 e 2004]).

Há dois modelos de algoritmos para etiquetagem mais utilizados: os baseados em regras e os estocásticos. Os algoritmos baseados em regras, utilizam regras para encontrar a categoria de um certo item lexical. Neste caso quando surgem novas regras, estas vão sendo anexadas à base de dados, a medida que novas situações de uso do item surgirem. Já os algoritmos estocásticos costumam solucionar seus problemas utilizando um *corpus* de treino, calculando a probabilidade de uma certa palavra ou item lexical precisar de uma determinada etiqueta que contenha um contexto condizente ao item ou palavra pesquisada (VIEIRA e STRUBE, [entre 2002 e 2004]).

O processo organizacional de etiquetagem prevê a existência de pelo menos dois corpora: um *corpus* de treino, onde o etiquetador aprenderá as regras e o *corpus* de texto, onde serão analisados. A eficiência do sistema dependerá da quantidade e qualidade dos dados treinados, a quantidade de etiquetas geradas (quanto mais etiquetas geradas, mais especifico é o resultado, porém terá maior possibilidade de ocorrer ambigüidade), similaridade e diferenças entre o *corpus* de treino e teste (se o *corpus* a ser etiquetado for muito diferente em estilo e gênero do *corpus* que irá treinar o etiquetador, haverá degradação na precisão da marcação) e a existência de palavras ou construções desconhecidas pelos corpora (VIEIRA e STRUBE, [entre 2002 e 2004]).

2.6 Processamento de Corpus

Atualmente os trabalhos realizados na área de lingüística computacional, tais como PLN e desenvolvimento de dicionários, são baseados em conteúdos de corpus, onde pode ser

encontrada uma gama de repositórios, da linguagem escrita ou falada, natural e espontânea para servirem de base para pesquisas da lingüística computacional e do aperfeiçoamento de tradutores. Esta forma de trabalho só foi possível após os anos 60 com auxilio do computador.

Atualmente surgiram novos tipos de repositórios capazes de inserirem informações lingüísticas ao corpus. A inserção de informações lingüística falada ou escrita em um corpus eletrônico e chamado de anotação de corpus. Um caso familiar de um tipo de anotação de corpus e o P*art-of-speech taggers* (POS *tagging*) ou etiquetagem, já comentado anteriormente no ítem (2.5). Neste caso cada palavra no corpus e representada por uma etiqueta que faz referência a classe gramatical da palavra dada. Assim como os estudos lingüísticos e os problemas encontrados em lingüística computacional são divididos em níveis, o corpus, também divide suas anotações em morfológica ou gramatical, sintática, semântica e de discurso (VIEIRA e STRUBE, [entre 2002 e 2004]).

2.6.1 O Corpus Compara

O Compara é um corpus desenvolvido no campo de atuação da Linguateca, sendo ele bi-direcional e extensível, tendo em sua base de dados coleções de textos originais e traduções de inglês e português. O Compara é extensível porque tem capacidade de incorporar cada vez mais o número de traduções à medida que vão sendo armazenados novas informações e bi-direcional porque pode reconhecer textos do português para o inglês e vice versa (COMPARA, 2002).

A principio o Compara foi desenvolvido para reunir uma coleção de fragmentos de ficção da língua portuguesa e inglesa e suas respectivas traduções. Atualmente, após a última atualização versão 6.0 do Compara, realizada em março de 2005, disponibiliza de autorização para incluir fragmentos de 62 pares de textos de ficção contemporâneos e não contemporâneos de autores e tradutores da África do Sul, Angola, Brasil, Moçambique, Estados Unidos, Reino Unido e Portugal, sendo que muitos destes textos já estão disponíveis no corpus (COMPARA, 2002).

O Compara tem a finalidade de auxiliar o estudo de traduções e comparações do português e inglês através de pesquisas automáticas, auxiliando como verificar diferentes palavras ou expressões as quais foram submetidas a traduções.

Atualmente o corpus Compara possui mais de 1 milhão de palavras originais sem contar com as traduções, cinqüenta e cinco textos originais, mais de 75.000 unidades de

alinhamento (sentido da tradução), que estão disponíveis na ultima versão (COMPARA, 2002).

2.7 Linguagem Prolog

A linguagem *Programming in Logic* (Prolog*)*, é uma linguagem de programação para computadores, que tem como base um subconjunto de lógica de predicado de primeira ordem, que trabalham sobre diferentes versões. O principal foco da programação lógica é identificar a noção de computação com a noção de dedução.

Seu funcionamento é interpretado como chamadas de procedimentos, fornecendo uma interpretação procedimental para lógica de primeira ordem. Prolog é uma maneira particular da linguagem em se obter provas automatizadas dos teoremas. É uma linguagem desenvolvida para solucionar problemas que envolvam objetos e relações entre objetos. Por exemplo, quando se diz que "*Maria possui um carro*", declara-se a situação de posse do objeto "*Maria"* com o objeto "*carro*". Em Prolog este exemplo fica da seguinte forma: *Maria possui carro* ou *possui (Maria, carro)*. Prolog não é uma linguagem imperativa como (C, C++, FORTRAN, etc.), que tem orientação para a máquina, isso significa que é especificado um fluxo de controle para a máquina executar uma determinada tarefa, ao contrario Prolog utiliza princípios declarativo, onde o fluxo de controle para a execução de uma tarefa é dirigido pelo interpretador, de modo que a descrição do domínio da aplicação está relacionada há seus objetos e suas relações.

Prolog é a linguagem mais utilizada em PLN, porque está fundamentada na idéia de que o computador deverá executar instruções que o programador lhe forneceu, ao invés do programador precisar pensar em termos operacionais da máquina. Para isso, a linguagem precisa de um conjunto de mecanismos automáticos e algumas funções predefinidas, para efetuar o raciocínio sobre o conhecimento que o programador possui sobre o problema (MARQUES, 2003).

Tratando de PLN, é fundamental a escolha de Prolog como linguagem de programação, pois é a única linguagem que possui maior número de características que facilitam a implementação de programas processadores de línguas naturais. Veja as principais:

- Facilidade em modificar estruturas de dados grandes e complexas: facilita o armazenamento de estruturas sintáticas, semânticas e entradas léxicas, que são elementos presentes em programas que manipulam LN;

- Capacidade de auto-análise e automodificação de programas: suporte a metaprogramação, facilitando a adoção de modelos abstratos de programação;
- Algoritmo de busca *depth-first* embutido: é um algoritmo de busca de informações em profundidade de fatos e regras em programa *Prolog*;
- *Definite Clause Grammar* (DCG) incorporado: é um tipo de formalismo estendido às gramáticas livres de contexto que identifica a relação entre os componentes de uma regra gramatical. O DCG não deixa sobrecarregar o processamento de sentenças de tamanho considerável (OLIVEIRA, 2004).

3 MATERIAIS E MÉTODOS

Devido ao grande crescimento dos meios de comunicação após a década de 80, e principalmente com a disseminação da *Word Wide Web* em meados dos anos 90, constatamos que as distancias entre os diferentes pontos do mundo ficaram no aspecto imaginário, mais curto. Pois a facilidade de qualquer pessoa poder comunicar com uma outra pessoa independentemente do local em que esta esteja geograficamente localizada, já não era mais considerado um grande problema, já que estes estão virtualmente próximos.

A partir destas mudanças, muitos projetos e STA que já existiam e que estavam parados ou restritos a alguns sistemas específicos, foram reativados com a finalidade de poder auxiliar os milhões de usuários que estavam acessando *sites*, manuais, etc, dentre outras finalidades, onde poderiam ser feitas as traduções, facilitando a comunicação entre diferentes idiomas.

Como a maioria dos STA ainda não oferece traduções de qualidade e o constante crescimento por procura de traduções *on-line* vem crescendo, muitos estudos vêm sendo elaborados nesta área para que possam ser construídos sistemas mais eficientes. A pesar da complexidade de desenvolvimento destes tipos de sistemas ainda ser grande, surgiu à idéia de fazer um estudo sobre alguns sistemas disponíveis no mercado, para poder verificar a sua eficiência.

Partindo destas pesquisas foi feito um estudo sobre alguns substantivos ambíguos existentes na língua inglesa com relação a sua tradução para a língua portuguesa.

Atualmente existem vários sistemas de traduções automáticas disponíveis na *Internet*, dentre eles muitos são livres e podem ser utilizados por qualquer usuário sem nenhum custo adicional pelo seu uso.

Os tradutores escolhidos para estudo, sejam *free ou* comerciais, não tem influência para este trabalho, o objetivo foi avaliar o desempenho das respostas obtidas pelas ferramentas escolhidas, para poder comparar a freqüência de erros de cada tradutor automático, levando em consideração apenas a ambigüidade dos substantivos analisados para o propósito de avaliação, sendo assim, qualquer outro tipo de erro encontrado na tradução não será levado em consideração.

Após as respostas obtidas, será feito um relatório distribuído em quadros, para poder estar analisando os erros de cada um dos tradutores escolhidos.

Serão selecionados dez substantivos que apresentam problemas de ambigüidade, para ser integrados na avaliação. Nenhum substantivo que não tenha ambigüidade na tradução da LF para a LA, fará parte do estudo, até porque estes não são problemáticos em TA. Os substantivos serão selecionados por tradutores profissionais, que escolherão os substantivos que apresentem tais problemas. Após a definição destes substantivos será utilizado o corpus Compara, para que possa ser feita uma mineração na sua base de dados, com o objetivo de retorno de sentenças que tenha o substantivo inserido na pesquisa, onde será selecionado dentre estas sentenças, algumas para poder fazer a tradução nos tradutores escolhidos, e depois ser feita a tradução por tradutores profissionais para comparar a melhor tradução.

Foram escolhidas sentenças não muito longas com no máximo trinta palavras para serem analisadas, porque se tratando de substantivos ambíguos quanto maior a sentença, maior e o grau de dificuldade de um tradutor traduzir uma sentença da LF para a LA corretamente, retornando uma tradução que seja coerente com a sentença original na LF. Esta delimitação na quantidade de palavras diminuirá a complexidade na tradução, diminuindo a quantidade de erros retornados na sentença, já que não está no escopo do trabalho detectar todos os erros de tradução que eventualmente surgirá depois de concluída.

Analisar-se á capacidade de interpretação do tradutor automático já que a busca pela melhor acepção ficará restrita a uma quantidade menor de palavras. Assim será verificado como estes tradutores escolheram a melhor acepção para o substantivo designado, se foi pelo freqüente uso dos substantivos ou se existiu critérios técnicos na implementação do *software* especifico (tradutor), para o significado retornado.

3.1 Seleção dos Tradutores Automáticos

Estão disponíveis atualmente diversos sistemas de TA gratuitos na *internet*, podendo ser utilizados por usuários *on-line*, para facilitar a tradução de textos e páginas da *Web*, que não esteja no idioma correspondente do usuário. Utilizou-se neste trabalho, tradutores que estão disponíveis gratuitamente na *Web* para se fazer as traduções necessárias para atingir os objetivos propostos. Os tradutores que serão utilizados são: Systran, FreeTranslation e E-Translation Server.

3.1.1 Sistema de Tradução Automática "Systran"

O sistema Systran utiliza arquitetura de tradução modular. Possui tecnologia revolucionária em tecnologias de tradução para Internet, PCs e infra-estrutura de rede. Facilitando a comunicação de 36 pares de língua, sendo que 20 são especialidades de domínio, permitindo que possam ser traduzidas até 150 palavras simultaneamente (SYSTRAN, 2004).

O Systran permite que usuários possam fazer traduções de arquivos Microsoft® Word, Power Point, Exel, Web pages, etc.

A tecnologia do Systran foi desenvolvida sobre Linux, mas rodam em todas as plataformas Unix e Microsoft Windows. Todo o sistema Systran utiliza PLN, no desenvolvimento do *software*. Este sistema tem capacidade de integralizar a diversas bases de conhecimento lingüísticos que estão disponíveis em aplicações *Web* (SYSTRAN, 2004).

A pesar deste sistema ser um dos mais utilizados atualmente, suas traduções ainda possuem inúmeros erros, principalmente de semântica, isto implica que as acepções esperadas nas traduções da LF para a LA, na maioria das vezes não condizem com as acepções esperadas.

3.2 Seleção dos Substantivos

Os substantivos foram selecionados por profissionais, que sugeriram que fosse trabalhado substantivos pouco ambíguos, devido a grande dificuldade encontrada para a obtenção de resultados satisfatórios.

Os substantivos escolhidos foram os seguinte: *board, bond, council, form, interest, issue, point, stand, stock e subject*. Conforme Fontes (2001), estes substantivos possuem os seguintes significados, veja exemplo Quadro 1.

SUBSTANTIVOS AMBIGUOS	
Substantivos em Inglês	Substantivos em Português
Board	tábua, quadro, comida e conselho.
Bond	laço e vinculo.
Council	conselho, câmara e assembléia.

Form	forma, tipo, formulário, série e banco.
Interest	interesse, juro e participação.
Issue	número e assunto.
Point	ponta, cabo, ponto, instante, apontar, característica, rumo e questão.
Stand	posição, suporte, estande, tribuna, arquibancada, barra, duração e reputação.
Stock	estoque, sortimento, gado, ações, caldo, cabo e coronha.
Subject	súdito, assunto, matéria, motivo, tema e sujeito.

QUADRO 1: Substantivos Ambíguos.

3.3 Utilizando o Corpus Compara para extração das sentenças

Foi escolhido o corpus Compara para procurar sentenças que tenham em seu contexto as palavras em Inglês, apresentadas no Quadro 1. Dentro do corpus Compara selecionamos o tipo de pesquisa simples que permite fazer uma busca em todos os textos disponíveis no corpus para procurar por sentenças que tenham estes substantivos. Dentre todas as sentenças retornadas, foi escolhida uma sentença para cada substantivo proposto, conforme Quadro 1.

- Sentença selecionada com o substantivo "*board*":

Sentença em Inglês.

*"Bob Busby is still busy at his bulletin **board**, rearranging old notices around the new one, like a fussy gardener tidying a flower bed."*

Tradução proposta pelo corpus para o Português.

"Bob Busby ainda está atarefado com o quadro, recolocando avisos velhos à volta dos novos, como um jardineiro esmerado a limpar um canteiro."

- Sentença selecionada com o substantivo "*bond*":

Sentença em Inglês.

*"He stood aghast at this snap of their great **bond**, at the renouncement that rang out in the word she so expressively sounded."*

Tradução proposta pelo corpus para o Português.
"Ele ficou consternado com a ruptura dos fortes **laços** que os uniam, com a renúncia patente nas palavras que ela acabava de pronunciar."

- Sentença selecionada com o substantivo "*council*":
Sentença em Inglês.
*"Then a Mayor of Rouen who was keen on statues rediscovered the original plaster cast -- made by a Russian called Leopold Bernstamm -- and the city **council** approved the making of a new image."*

Tradução proposta pelo corpus para o Português.
"Depois, um presidente da câmara de Rouen, que gostava muito de estátuas, descobriu o molde original de gesso -- feito por um russo chamado Leopold Bernstamm -- e a **assembléia** municipal aprovou o projeto de uma nova estátua."

- Sentença selecionada com o substantivo "*form*":
Sentença em Inglês.
*"Sport used to be my chief **form** of therapy, though I didn't call it that."*

Tradução proposta pelo corpus para o Português.
"O esporte era a minha principal **forma** de terapia, mesmo que não o visse dessa maneira."

- Sentença selecionada com o substantivo "*interest*":
Sentença em Inglês.
*"Gradually, over the years, Philip's own **interest** in the physical side of marriage declined, but he persuaded himself that this was only normal."*

Tradução proposta pelo corpus para o Português.
"Gradualmente, com os anos, o **interesse** de Philip pelo aspecto físico do casamento diminuiu, mas convenceu-se de que isso era normal."

- Sentença selecionada com o substantivo "*issue*":
Sentença em Inglês.

*"Wily produced the current **issue** from one of his capacious pockets."*

Tradução proposta pelo corpus para o Português.
"Wily tirou o último **número** de um dos seus bolsos imensos."

- Sentença selecionada com o substantivo "*point*":
*"The **point** at which you suspect too much is being read into a story is when you feel most vulnerable, isolated, and perhaps stupid."*

Tradução proposta pelo corpus para o Português.
O ponto em que suspeitamos que estamos a ler demais numa história é quando nos sentimos mais vulneráveis, isolados e talvez estúpidos.

- Sentença selecionada com o substantivo "*stand*":
Sentença em Inglês.
*"Which is surprising, in a way, when you consider that, as explained above, a principal reason why they are all gathered here is that they believe it will **stand** them in good stead in the next world."*

Tradução proposta pelo corpus para o Português.
"O que é surpreendente, de certa maneira, quando pensamos que, como acima ficou dito, a razão principal por que aqui estão todos reunidos é que acreditam que isso os coloca em boa **posição** no outro mundo."

- Sentença selecionada com o substantivo "*stock*":
Sentença em Inglês.
*"Morris Zapp was standing at the window of his office at Rummidge, smoking a cigar (one of the last of the **stock** he had brought with him into the country) and listening to the sound of footsteps hurrying past his door."*

Tradução proposta pelo corpus para o Português.
"Morris Zapp estava de pé em frente à janela de sua sala em Rummidge, fumando um charuto (um dos últimos do **estoque** que trouxera para a Inglaterra) e ouvindo o som de passos ligeiros pelo corredor."

- Sentença selecionada com o substantivo "*subject*":

Sentença em Inglês.

"I said it was unrealistic to suppose that an educated middle-class couple would discuss the possible pregnancy of their schoolgirl daughter without mentioning the ***subject****."*

Tradução proposta pelo corpus para o Português.

"Eu disse que era irrealístico, pensar que um casal culto de classe média, ao discutir a possível gravidez de sua filha adolescente, não mencionasse o **assunto**."

4 RESULTADOS E DISCUSSÕES

Após a definição de todos os substantivos e selecionado as sentenças no corpus Compara, e inseridas as sentenças nos tradutores e chegou-se aos seguintes resultados. As respostas obtidas pelos três tradutores não foram alteradas, nem corrigidos erros de sintaxe, permanecendo idênticas as retiradas do corpus.

Traduções com o substantivo "*board*", retirada no corpus Compara, texto EBDLZ (1004):

Sentença selecionada no Corpus Compara	*Bob Busby is still busy at his bulletin **board**, rearranging old notices around the new one, like a fussy gardener tidying a flower bed.*
Systran	O busby de Bob é ainda ocupado em sua **placa** de boletim, rearranjando observações velhas em torno do novo, como um gardener fussy que arruma uma cama da flor
FreeTranslation	Balouce-se Busby é imóvel ocupado nseu **quadro** de avisos, redistribuindo avisos velhos ao redor do novo, como um jardineiro inquieto arruma um canteiro.
E-Translation Server	Bob Busby está ainda ocupado ao seu **quadro** de avisos, reajustando notificações velhas em torno ao novo, como um jardineiro exagerado que põe em ordem uma cama de flor.
Tradução humana	Bob Bushy ainda está ocupado com seu **quadro** de avisos reorganizando velhos avisos entorno dos novos, como um caprichoso jardineiro fazendo uma cama de flores.

QUADRO 2: Substantivo "*board"*.

Traduções com o substantivo "*bond*", retirada no corpus Compara, texto EBJT3 (535):

Sentença selecionada no Corpus Compara	*He stood aghast at this snap of their great **bond**, at the renouncement that rang out in the word she so expressively sounded.*
Systran	Estêve aghast nesta pressão de sua **ligação** grande, no renouncement que soou para fora na palavra que soou assim expressively.
FreeTranslation	Ficou estupefato neste estalo de seu grande bônus, no renouncement que tocou fora na palavra ela tão expressivamente soou.
E-Translation Server	Ele ficou aterrorizado por este estalo do seu **vínculo** grande, ao renouncement que tocou na palavra que ela tocou tão expressively.
Tradução humana	Ele instantaneamente se pôs contra o repentino o grande **vínculo** (compromisso) deles na renúncia que saiu nas notícias e que ela tão expressivamente falou.

QUADRO 3: Substantivo "*bond*".

Traduções com o substantivo "*council*", retirada no corpus Compara, texto EBJB (17):

Sentença selecionada no Corpus Compara	*Then a Mayor of Rouen who was keen on statues rediscovered the original plaster cast -- made by a Russian called Leopold Bernstamm -- and the city **council** approved the making of a new image.*
Systran	Então um mayor de Rouen que era afiado em estátuas rediscovered o emplastro original moldado -- feito por um russian chamado Leopold Bernstamm -- e o **conselho** de cidade aprovou fazer de uma imagem nova
FreeTranslation	Então um Prefeito de Rouen que estava afeiçoado a estátuas redescobriu a moldagem original -- feito por russo chamou Bernstamm de Leopold -- e o **conselho** de cidade aprovou a fabricação de uma nova imagem.
E-Translation Server	Então um prefeito de Rouen que era apaixonado por estátuas redescobriu a engessadura -- original feita por um Leopold Bernstamm -- chamado russo e o **conselho** de cidade aprovou a execução de uma nova imagem.

Tradução humana	Então, o prefeito de Rouen, o qual era entusiasmado com estátuas, redescobriu o gesso original, feito por um russo chamado Leopold Bernstamm, e a **câmara** municipal, aprovou o feitio de uma nova imagem.

QUADRO 4: Substantivo "*council"*.

Traduções com o substantivo "*form*", retirada no corpus Compara, texto EBDL1T2 (501):

Sentença selecionada no Corpus Compara	*Sport used to be my chief **form** of therapy, though I didn't call it that.*
Systran	O esporte usou-se ser meu **formulário** principal da terapia, embora eu não o chamei isso.
FreeTranslation	O atletismo era minha **forma** principal de terapia, embora eu não chamei isso.
E-Translation Server	O esporte era a minha **forma** principal de terapia, embora eu não o chamasse aquilo.
Tradução humana	O esporte, era minha principal **forma** de terapia, embora eu não o chamasse assim.

QUADRO 5: Substantivo "*form*".

Traduções com o substantivo "*interest*", retirada no corpus Compara, texto EBDL3T1(240):

Sentença selecionada no Corpus Compara	*Gradually, over the years, Philip's own **interest** in the physical side of marriage declined, but he persuaded himself that this was only normal.*
Systran	Gradualmente, sobre os anos, próprio **interesse** de Philip no lado físico da união declinou, mas persuadiu-se que esta era somente normal.
FreeTranslation	Gradualmente, sobre os anos, próprio **interesse** do Philip no lado físico de casamento inclinado, mas convenceu se que isto era único normal.
E-Translation	Gradualmente, durante os anos, o **interesse** próprio de Philip pelo lado

Server	físico de casamento baixou, mas ele se persuadiu que este era somente normal.
Tradução humana	Gradualmente, com o passar dos anos, o próprio **interesse** de Philip no lado físico do casamento declinou, mas ele se convenceu que aquilo era normal.

QUADRO 6: Substantivo "*interest*".

Traduções com o substantivo "*issue*", retirada no corpus Compara, texto EBDL3T1 (1078):

Sentença selecionada no Corpus Compara	*Wily produced the current **issue** from one of his capacious pockets.*
Systran	Wily produziu a **edição** atual de um de seus bolsos capacious.
FreeTranslation	Ardiloso produziu a **edição** atual de um dseus bolsos vastos.
E-Translation Server	Astuto produzido o **argumento** corrente de um dos seus bolsos vastos.
Tradução humana	Wily produziu a atual **edição,** a partir de um de seus espaçosos bolsos.

QUADRO 7: Substantivo "*issue*".

Traduções com o substantivo "*point*", retirada no corpus Compara, texto EBJB1 (209):

Sentença selecionada no Corpus Compara	*The **point** at which you suspect too much is being read into a story is when you feel most vulnerable, isolated, and perhaps stupid.*
Systran	O **ponto** em que você suspeita que demasiado está sendo lido em uma história é quando você sente o mais vulnerável, isolado, e talvez stupid.
FreeTranslation	O **ponto** em que você suspeita demais está sendo lido numa história é quando sente-se bem vulnerável, isolado, e talvez estúpido.
E-Translation Server	O **ponto** a qual você suspeita que demasiado esteja sendo lido num conto é quando você se sente mais vulnerável, isolado, e talvez estúpido.
Tradução humana	O **fato** no qual você duvida muito, está sendo atribuído a uma notícia, e é quando você se sente mais vulnerável, isolado e talvez tolo.

QUADRO 8: Substantivo "*point*".

Traduções com o substantivo "*stand*", retirada no corpus Compara, texto EBDL4 (202):

Sentença selecionada no Corpus Compara	*Which is surprising, in a way, when you consider that, as explained above, a principal reason why they are all gathered here is that they believe it will* ***stand*** *them in good stead in the next world.*
Systran	Qual está surpreendendo, em uma maneira, quando você considerar que, Como explicado acima, uma razão principal porque é tudo recolhida aqui é que acreditam ele os **estará** no stead bom no mundo seguinte.
FreeTranslation	Surpreende, de certa maneira, quando considera isso, como explicado acima, uma razão principal por que eles estão todos reunidos aqui está que acreditam que ele os **ficará** úteis no próximo mundo.
E-Translation Server	Qual é surpreendente, numa maneira, quando você considera que, como explicado sobre, uma razão principal por que eles são todos deduziu aqui é que eles acreditam que tolera-los-á em stead bom no mundo seguinte.
Tradução humana	O que é surpreendente, de alguma forma quando você considera que com explicado acima, que a razão principal porque eles estão todos reunidos aqui, e que eles acreditam que isso ira suporta-los firmemente no próximo mundo.

QUADRO 9: Substantivo "*stand*".

Traduções com o substantivo "*stock*", retirada no corpus Compara, texto EBDL3T2 (1458):

Sentença selecionada no Corpus Compara	*Morris Zapp was standing at the window of his office at Rummidge, smoking a cigar (one of the last of the* ***stock*** *he had brought with him into the country) and listening to the sound of footsteps hurrying past his door.*
Systran	Morris Zapp estava estando na janela de seu escritório em Rummidge, fumando um charuto (um do último do **estoque** tinha Trazido com ele no país) e escutando o som dos passos que apressam-se após sua porta.

FreeTranslation	Zapp de Morris ficava na janela dseu escritório em Rummidge, fumando um charuto (um do último do **estoque** ele tinha trazido com ele no país) e escuta o som de pegadas apressando-se após seua porta.
E-Translation Server	Morris Zapp estava ficando à janela do seu escritório a Rummidge, curando um charuto (um do último da **reserva** que ele tinha portado com ele no país) e escutando o som de passos que se apressam depois da sua porta.
Tradução humana	Morris Zapp, estava de pé na janela de seu escritório no Rummidge, fumando um charuto (um dos últimos do **estoque** que ele tinha trazido com ele para o país) e ouvindo ao som de passos apressados, passando na sua porta.

QUADRO 10: Substantivo "*stock*".

Traduções com o substantivo "*subject*", retirada no corpus Compara, texto EBDLT1T2 (1246):

Sentença selecionada no Corpus Compara	*I said it was unrealistic to suppose that an educated middle-class couple would discuss the possible pregnancy of their schoolgirl daughter without mentioning the* ***subject****.*
Systran	Eu disse que era unrealistic supôr que um par educado do middle-class discutiria a gravidez possível de sua filha do schoolgirl sem mencionar o **assunto**.
FreeTranslation	Disse que era irreal supor que um par burguês educado discutiria a possível gravidez de seua filha de schoolgirl sem mencionar o **assunto**.
E-Translation Server	Disse que era fantasioso para supor que um casal de classe média educado discutiria sobre a possível gravidez da sua filha de aluna sem mencionar o **sujeito**.
Tradução humana	Eu disse, que foi irrealista imaginar, que um educado casal de classe media, discutiria a possível gravidez de sua filha em idade escolar, sem mencionar o **assunto**.

QUADRO 11: Substantivo "*subject*".

Conforme observado o STA E-Translation Server conseguiu a melhor acepção para quatro substantivos nas sentenças analisadas, assim como a melhor tradução para a maioria das sentenças apresentadas comparada a tradução humana.

O STA FreeTranslation obteve a melhor acepção para seis substantivos, porém as traduções para a sentença completa foram inferiores a apresentada pelo E-Translation Server, comparadas a tradução humana.

O Systran conseguiu traduzir corretamente quatro substantivos, porém teve o pior desempenho comparado à sentença completa.

Verificou-se também que o problema de ambigüidade afeta não somente a tradução automática, pois a falta de conhecimento de toda a sentença prejudica até mesmo a tradução feita por humanos, como pode ser visto na tradução do substantivo "*stand*", cuja tradução humana não ficou adequadamente correta devido à falda do conhecimento do texto antecessor e sucessor.

Os resultados obtidos mostram a existência de ambigüidade na tradução feita por tais tradutores. Tal problema é causado pela deficiência destas ferramentas em não propiciar a desambiguidade tanto dos substantivos, quanto da própria sentença. A falta de desambiguadores eficientes nestas ferramentas dificulta a busca no dicionário, da palavra que tenha o melhor sentido na tradução da sentença avaliada.

4.1 Utilizando o PLN para superar as dificuldades apresentadas pela ambigüidade

As ferramentas utilizadas apresentaram dificuldades relacionadas ao tratamento da ambigüidade. Uma possível otimização destas ferramentas pode ser feita utilizando a técnica de Processamento de Linguagem Natural. Entretanto, tal técnica, também apresenta limitações relacionadas ao tratamento da ambigüidade.

Como citado no item (2.3.1), a técnica de PLN é composta por algumas etapas, sendo as principais a análise morfológica, análise sintática, análise semântica, análise do discurso e processamento pragmático. A ambigüidade pode aparecer em cada um destas etapas.

Em domínios genéricos é difícil à implementação de tais etapas com certo grau de eficiência, devido ao pouco conhecimento técnico existente para solução de tais problemas. A eficiência da técnica de PLN cabe ao desenvolvimento de ferramentas que solucione a ambigüidade existente para cada uma das etapas, além de poder resolver não só a ambigüidade dos substantivos, mas também a da própria sentença.

O maior limite do PLN ainda é o desenvolvimento destas ferramentas, pois exige estudos profundos tanto da gramática quanto do conhecimento do comportamento humano relacionado à linguagem natural. A melhor utilização possível de PLN seria para domínios específicos de textos, por exemplo, computação, medicina, meteorologia, etc.

Desta forma a ambigüidade entre os substantivos poderia ser menor, devido a pouca utilização das variações de significado, o que melhoraria as traduções por estar tratando de textos mais técnicos e específicos, ao contrario se utilizados em textos literários não poderiam trazer resultados satisfatórios, devido os textos possuírem uma interpretação muito pessoal para cada leitor.

O PLN está em processo evolutivo, pois é uma técnica capaz de adquirir conhecimento a partir do momento que novas informações são inferidas, porém a complexidade também aumenta junto.

O tratamento da ambigüidade é um caso específico e de difícil resolução, mesmo em sistemas que utilizam PLN, este assunto é extremamente amplo e apresenta grande complexidade em qualquer gramática, dificultando mais ainda, se não considerar o conhecimento do contexto (conhecimento do mundo). Desta forma, sua utilização se mostra mais viável para domínios específicos onde há a possibilidade de limitar as palavras e suas variações, para serem utilizadas nos textos que constituem o assunto tratado.

Para a ambigüidade semântica poder tentar solucioná-la através de uma abordagem baseada em classificação *bayesiana*, que seria classificar o sentido dependendo do contexto. Através desta abordagem é possível representar conhecimentos incertos, neste caso pode-se a partir de uma palavra ambígua ***p*** encontrar o sentido ***s*** com a maior probabilidade de possuir o significado coerente. O problema é que teria que ser percorrido todo ou parte do texto para obter-se o conhecimento do significado, com isso aumentaria muito o processamento e a complexidade do sistema, sendo assim, seria melhor o desenvolvimento de *parsers.*

Uma outra possível solução para ambigüidade semântica seria a abordagem baseada em dicionário, porém este método não é o melhor para traduções de textos genéricos, porque eles apenas fazem à substituição das palavras traduzidas, mas poderiam ser utilizados em traduções de textos de conhecimentos restritos, como por exemplo o STA Taum Meteo, comentado no item (2.2.2) não tendo capacidade interpretativa da sentença, utilizando um dicionário bilíngüe para descobrir o significado mais comum para aquela palavra, sem preocupar com a ambigüidade.

Se forem considerados o domínio da tradução poderão ser desenvolvidos *parsers,* que é a transformação de uma frase potencialmente ambígua em uma não ambígua.

Considerando esta situação, a ambigüidade será tratada, durante a análise e projeto do sistema de tradução. Esta ambigüidade poderá ser limitada com a implementação de *parsers* que reduziria tais problemas facilitando o tratamento das sentenças, para aplicação da técnica de PLN. Tais propostas poderiam ser implementadas utilizando a linguagem Prolog, que apresenta algumas características que facilitaria a implementação de uma ferramenta para auxiliar na otimização dos STA. Para estes casos a escolha do Prolog seria incontestável, devido a estas características que atenderiam os requisitos mínimos que outras linguagens não o fariam, conforme especificado no item (2.7).

5 CONCLUSÃO

Devido a grande demanda pela utilização da *Word Wide Web* (Internet), fez se necessária a disponibilização de sistemas de tradução automática para auxiliar na tradução de textos e páginas nela encontrados.

Após a realização da traduções em três diferentes tipos de sistemas de tradução automática, foi possível fazer uma avaliação das respostas obtidas. Tratando da análise de substantivos, em muitas das sentenças propostas, os tradutores fizeram algumas traduções com a resposta esperada, entretanto isso não significou que a tradução da sentença ficasse com o sentido correto. Verificou-se que tais tradutores fazem as traduções buscando os significados das palavras com maior freqüência de utilização na tradução, assim, foi possível concluir que cada uma das palavras utilizadas na tradução são aquelas com maior ocorrência na escrita de texto na língua alvo (LA).

Pressupôs-se que, estes tradutores não possuem desambiguadores eficientes embutidos no sistema, até porque, aumentaria muito sua complexidade. Com isso, o processamento das informações cresceria exponencialmente para obter a melhor acepção para a sentença, aumentando o tempo da tradução e o custo de desenvolvimento destes *softwares*.

Recomenda-se que em uma sentença avaliada, para obtenção de uma resposta mais coerente, seja necessário, que toda a sentença seja considerada, ou seja, avaliar o contexto geral, para que as palavras ambíguas, após submetidas ao desambiguador, propiciem ao sistema a capacidade de encontrar o melhor sentido da palavra em seu dicionário.

Concluiu-se, que muito terá que ser feito, para se obter um tradutor, que consiga traduzir sentenças de uma forma próxima da apresentada na tradução humana. É recomendável, que o desenvolvimento destes tradutores seja voltado para áreas específicas, dividindo o campo de atuação. Isso diminuiria a complexidade na busca da melhor acepção, porque a procura de um significado em uma base de dados mais enxuta seria menos complexa, obtendo respostas mais coerentes.

Como trabalho futuro, a partir do estudo realizado, existe a possibilidade de modelar e implementar um desambiguador para substantivos, utilizando a linguagem de programação Prolog, juntamente com a técnica de PLN. Portanto o desenvolvimento desta ferramenta

diminuirá a dificuldade que o STA, terá em definir qual a melhor acepção para a palavra traduzida.

REFERÊNCIAS

ALFARO, C. **Descobrindo, Compreendendo e Analisando a Tradução Automática.** Rio de Janeiro, 1998. Disponível em: <http://www.tecgraf.puc-rio.br/~carolina/monografia/apresentacao.html> Acesso em: 18/03/2005.

APPI... **Interlíngua**. São Paulo, 2005. Disponível em: <http://wco.sites.uol.com.br/ > Acesso em: 30/04/2005.

ARARIBÓIA, G. **Inteligência Artificial - um curso prático.** Rio de Janeiro: LTC, 1988. p. 149-237.

BITTENCOURT, G. **Breve história da Inteligência Artificial.** Santa Catarina, 2004. Disponível em: <http://www.das.ufsc.br/gia/history/ > Acesso em: 27/03/2005.

BRAGA, C. **Modularidade na tradução.** Rio de Janeiro, 2005. Disponível em: <http://www.ic.uff.br/~cbraga/comp/2005.1/material-didatico.html > Acesso em: 08/05/05.

COMPARA. Lisboa, 2002. Disponível em: <http://www.linguateca.pt/COMPARA/Bem-vindos.html > Acesso em: 22/04/2005.

FERNANDES, A. M. R. **Inteligência Artificial: noções gerais.** Florianópolis: Visual Books Ltda, 2003.

FILHO, I. W. R.. **Processamento de Linguagem Natural.** Santa Catarina, 2004. Disponível em: <http://www.inf.ufsc.br/~ilson/> Acesso em 20/08/2004.

GARRÃO, M. U. **Tradução Automática: ainda um enigma multidisciplinar.** Rio de Janeiro, [entre 2000 e2004]. Disponível em: <http://www.filologia.org.br/vcnlf/anais%20v/civ11_05.htm > Acesso em: 19/04/2005.

GIGAFLOPS... **Linguagem Pascal.** 2005. Disponível em: <http://gigaflops.tripod.com/page/lp/pascal/pascal.html > Acesso em: 08/05/05.

MARQUES, M. L. **Introdução a Linguagem *Prolog.*** Brasília, 2003. Disponível em: <http://portal.cid.unb.br/pipermail/reprecon/attachments/20031030/e39dee12/ApostilaPROLOGProfMamede-0001.bin > Acesso em: 16/12/2004.

MARTINS, R. T. **Tradução Automática e Estudos da Tradução: um conflito paradigmático.** São Paulo, 2003. Disponível em: http://www.nilc.icmc.usp.br/til2003/oral/RonaldoMartins_31.pdf > Acesso em : 18/04/2005.

OLIVEIRA, F. A. D. **Processamento de Linguagem Natural: princípios básicos e a implementação de um analisador sintático de sentenças da língua portuguesa.** Rio Grande do Sul, 2004. Disponível em: <http://www.inf.ufrgs.br/procpar/disc/cmp135/trabs/992/Parser/parser.html#_Toc470452821> . Acesso em: 17/08/2004.

PERINE, M. A. **Gramática Descritiva do Português.** 4.ed. São Paulo: Ática, 2002. p. 250-252.

RUSSEL, S; NORVIG, P. **Inteligência Artificial.** 2.ed. Rio de Janeiro: Campus, 2004. p. 3-31, 823-825.

SANTOS, D. **Introdução ao Processamento de Linguagem Natural através das aplicações.** Lisboa, 2005. Disponível em: <http://www.linguateca.pt/Diana/download/aplicacoes.rtf > Acesso em: 30/04/2005.

SEMIOTIC..., **Processamento Simbólico.** São Paulo, 2005. Disponível em: <http://www.semiotic.com.br/conceito/proc_simb.htm > Acesso em: 01/05/05.

SIQUEIRA, R. ***InTranslator*: Sistema de Tradução de Idiomas *Insite*.** São Paulo, 2001. Disponível em: <http://linguistica.insite.com.br/intranslator/paper1.phtml> Acesso em: 17/08/2004.

SIQUEIRA, R. **Processamento de Linguagem Natural.** São Paulo, 2001. Disponível em: <http://linguistica.insite.com.br/nlp.phtml> Acesso em: 17/08//2004

SPECIA, L.; NUNES, M. das G. V. **A Ambigüidade Lexical de Sentido na Tradução do Inglês para o Português – um recorte de verbos problemáticos.** São Carlos, 2004a. Disponível em: <http://www.nilc.icmc.usp.br/~specia/publications/TR0401-SpeciaNunes.pdf> Acesso em 15/03/2005.

SPECIA, L; NUNES, M. das G. V. **O Problema da Ambigüidade Lexical de Sentido na Comunicação Multilingüe**. São Carlos, 2004b. Disponível em: < http://www.nilc.icmc.usp.br/nilc/download/TIL2004-SpeciaNunes.pdf> Acesso em 14/11/2004.

SPECIA, L. **Modelagem de um Interpretador Lexical para a Linguagem DART.** Cascavel, 2000c. Disponível em: <http://www.nilc.icmc.usp.br/~specia/publications/MonoGrad2000-Specia.pdf > Acesso em: 30/04/2005.

SYSTRAN. 2004. Disponível em: <http://www.systransoft.com/index.html > Acesso em: 22/04/2005.

VIEIRA, R; STRUBE, V. L.. **Lingüística Computacional: princípios e aplicações.** Rio Grande do Sul, [entre 2002 e 2004]. Disponível em: <http://www.inf.unisinos.br/~renata/laboratorio/publicacoes/jaia12-vf.pdf> Acesso em 05/04/2005.