CENTRO UNIVERSITÁRIO FEEVALE

LUCAS BRANDT

EXTRAÇÃO DE INFORMAÇÕES EM SALAS EM BATE-PAPO

NOVO HAMBURGO
2009

LUCAS BRANDT

EXTRAÇÃO DE INFORMAÇÕES EM SALAS DE BATE-PAPO

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de informação pelo centro Universitário Feevale

ORIENTADOR: RODRIGO RAFAEL VILLARREAL GOULART

NOVO HAMBURGO
2009

AGRADECIMENTOS

Agradeço a todas as pessoas que tem ajudado a idealizar este trabalho.

RESUMO

Esse trabalho apresenta a necessidade atual do ser humano em comunicar-se e estar informado. Atualmente a comunicação é realizada de diversas maneiras, entretanto a que mais se destaca é a comunicação eletrônica, emails, messengers, chats entre outros. Este meio de comunicação utilizado pelo um numero cada vez maior de pessoas, faz com que sejam gerados inúmeras quantidades de informações não estruturadas. Tal cenário possibilita o desenvolvimento de diversas ferramentas para estruturar e organizar estes dados. Sendo que este trabalho propõem uma ferramenta capaz de extrair informações em salas de bate-papo, devido que a mesma atualmente é utilizado não somente como entretenimento, mas como ferramenta de apoio ao ensino. Para idealizar o desenvolvimento desta ferramenta, será necessário a utilização de técnicas de PLN e extração de informações. Sendo assim, o presente trabalho descreve a metodologia que será empregada de ambas as áreas, bem como propõem o desenvolvimento de um corpus, direcionado a analise das sessões das salas de bate papo.

Palavras-Chave: PLN, Extração, Informação, Comunicação, Bate-Papo.

ABSTRACT

This paper presents the current need of mankind to communicate to each other and be informed about everything around. Nowadays communication is done in many ways, however the one that most stands out is the electronic communication, emails, messengers, chats and many others. This type of communication is being used by an increasing number of people every day, makes them generated numerous amounts of unstructured information. This scenario allows the development of various tools to build up and organize this information. Since this paper proposes a tool capable of extracting information in chat rooms, because that is currently used not only as entertainment but also as a tool to support education and others. So, to build up the development of this tool will be necessary to use the techniques of PLN and extraction of information. Therefore, this paper describes the methodology to be employed in both areas, as well as proposes the development of a corpus, directed to the analysis of the chat room sessions.

key words: NLP, Extraction, Information, Comunication, Chats

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PLN	Processamento de Linguagem Natural
NLP	Natural Language Processor
MIT	Massachusetts Institute of Technology
HTML	Hyper Text Markup Language
PDF	Portable Document Format
XML	Extensible Markup Language
EPC-P	Extrator de Palavras-Chave por freqüência de Padrões
EPC-R	Extrator de Palavras-Chave por freqüência de Radicais
POS	Part-Of-Speech
ID	Identificador Único
NEAD	Núcleo de Ensino e Educação à Distância

SUMÁRIO

INTRODUÇÃO

Atualmente percebe-se que estamos inseridos em um mundo altamente globalizado, cuja principal arma para a sobrevivência é a informação. A mesma é de suma importância para sobrevivermos contra a concorrência, sabermos aonde ir, estarmos informados sobre o mercado e demais área de interesse pessoal.

Entretanto não temos a informação pronta de maneira fácil em nossas mãos. O ambiente o qual vivemos está repleto de informações, estamos imersos em um mar de dados não estruturados, vivenciamos uma era onde temos uma quantidade enorme informações em nossa volta.

Com o passar dos anos e subseqüente o desenvolvimento da sociedade, os dados e informações tem seu crescimento multiplicado em escalas extraordinárias. Então como podemos sobreviver a essa imensidão? Como filtrar esta enorme diversidade e garantir para nós somente o que é importante? Eis que surge a informática para sobrevivermos a tal desafio imposto. Apesar do surgimento da informática, para nos auxiliar a lidar com o problema das grandes massas de dados existentes, a mesma acabou facilitando e geração de uma quantidade maior de dados. Graças o surgimento da internet, e de sua popularização, as massas de dados foram aumentando exponencialmente, não apenas em tamanho, mas se disseminando-se em diversos lugares, então a mesma nos impõem mais um obstáculo. Este gigantesco mar que é a internet hoje notamos além da diversidade, percebemos também um redundância de informações, muitas vezes algumas delas estão incorretas, entretanto cabe a nós sabermos escolher e identificar o que é valido.

Outro ponto importante a se destacar em nossas vidas atualmente, além da necessidade de estarmos constantemente informados, é comunicação, assim como os dados, os meios de comunicação também vêm sofrendo grandes modificações, tanto tecnológicas ou não. Por exemplo, há um tempo atrás, podíamos nos comunicar apenas por carta, após um tempo passamos a nos comunicar por telefone, e hoje temos varias opções, como ligações telefônicas via internet, sites de relacionamento, mensageiros instantâneos, chats e entre outros disponíveis gratuitos ou não. Focalizando na comunicação via escrita como alguns exemplos citados anteriormente, podemos ter uma grande fonte de dados. Este meio de comunicação tem se consagrado muito entre as pessoas, que os utilizam para finalidades

múltiplas, como reuniões de negócio, disseminação do conhecimento entre grupos de pessoas, entretenimento e até mesmo estudos.

Focalizando as salas de bate-papo disponíveis na internet, é de grande validade para as pessoas, como por exemplo, o chat disponível no site NEAD(Núcleo de Educação e Ensino à Distância) da Feevale, sendo que sua utilização é possível somente para disciplinas cursadas a distância, ou também a disciplinas presenciais, mas que necessitem de encontros fora os presenciais. Notamos que a mesma é de suma importância para a troca de informações e conhecimento entre os alunos. As sessões de bate-papo que ocorrem disponibilizam uma grande fonte de dados semi-estruturados.

Esta modalidade de cursos e cadeiras tem se tornado muito popular entre as universidades. A possibilidade de ter aulas à distância tem agradado em muito as pessoas. Perante tal cenário de interesse, notamos que têm surgido muitas universidades que disponibilizam o curso integralmente à distância. Favorecendo as entidades provedoras de ensino, no sentido de não necessitar manter um patrimônio físico grande para prover o ensino, as mesmas necessitam de ferramentas tecnológicas para poderem oferecer seus cursos. Entre várias a ferramentas podemos citar o ambiente de ensino a distancia, que esteja disponível na internet, e neste ambiente é necessário a presença de uma sala de bate-papo. Entretanto o uso do chat em cursos a distancia atualmente tem se limitado a comunicação momentânea, porém atualmente não se tem notado a sua riqueza em informações. Podemos definir estas informações como semi-estruturadas, ao caso que cada mensagem enviada pelos participantes de uma sessão conterem idéias formuladas pelos mesmos, que se mistura a tantas outras mensagem enviadas por outro participantes da sala. A possibilidade de se estruturar de uma forma mais organizada, é possível somente com o desenvolvimento de uma ferramenta capaz de trabalhar com as mensagens registradas, sendo este o principal foco deste trabalho, que será possível somente com a utilização da área de PLN (Processamento de Linguagem Natural). O Capitulo I é dado uma introdução referente a área de PLN, bem como a historia de seu surgimento e evolução. No Capitulo II, é descrito o funcionamento de algumas ferramentas já existente, conseqüentemente o capitulo III aborda a descrição da estrutura das ferramentas. O conceito, algumas regras de geração de um corpos são descritos no Capitulo IV, após a conceitualização o Capitulo V exemplifica de descreve algumas ferramentas existentes para o desenvolvimento do mesmo. O Capitulo VI, aborda algumas metodologias para extração de informação.

1 PLN

Desde o seu surgimento até os dias atuais o PLN vem abrangendo cada vez mais, diversas e complexas áreas do conhecimento e da informação. Com a intenção de remover algumas barreiras impostas entre a relação ser humano e máquina, como por exemplo, a incapacidade do computador compreender a linguagem natural, isto é, a língua que falamos, a mesma vem abstraindo a necessidade de digitação de códigos complexos e muitas vezes incompreensíveis para operarmos um computador. Atualmente existem algumas ferramentas que diminuem essa complexidade; tomamos como exemplo a interface gráfica de um sistema operacional, que torna a usabilidade mais simples ao usuário. Entretanto a apesar dessa facilidade, o computador ainda é incapaz de compreender nossa língua, a Linguagem Natural. Existem diversos estudos na própria área do PLN, que tentam acabar com esse problema, entretanto esses estudos vêm caminhando a passos curtos, devido ao desafio imposto, que é fazer o computador compreender nossa linguagem.

Neste capítulo, será exposto um breve histórico de como surgiu o PLN, a metodologia envolvida, uma breve introdução aos diferentes tipos de conhecimento lingüístico para o tratamento da linguagem natural, arquiteturas que possibilitam a interpretação e geração de línguas naturais, e ainda um introdução de processamento de análise sintática automática.

1.1 PLN: HISTORIA E METODOLOGIA.

Como descrito no capítulo inicial, a comunicação tem evoluído constantemente nas últimas décadas, isso graças ao desenvolvimento das tecnologias envolvidas, e principalmente desde a introdução do computador em nossa sociedade, no inicio dos anos 40, além de trazerem avanços substâncias na área do conhecimento cientifico, os mesmos possibilitaram o início e desenvolvimento de pesquisas em diversas outras áreas. Devido a sua praticidade em auxiliar

os pesquisadores em seus estudos e descobertas, o computador, desde que inserido em nosso cotidiano tem modificado em muito nossas vidas.

Apesar de inúmeros benefícios, no âmbito de auxilio a desenvolvimento de tarefas complexas ao ser humano, os computadores impuseram um grande paradigma, perante seu surgimento: Como fazer com que eles possam entender as instruções por nos impostas, para serem executadas? A partir deste grande enigma que surgiram as linguagens de programação, como uma possibilidade imediata para a solução "parcial" deste problema.

Entretanto esta solução tem estado presente nos dias de hoje, e embora tenha sido uma solução inicial para o problema, as mesmas também vêm evoluindo constantemente. Para os envolvidos no campo do desenvolvimento de programas para computador, sua vida atualmente tem sido muito facilitada, com as linguagens de alto nível. Para solução inicial do enigma, surgiram as linguagens de programação de baixo nível, neste cenário podemos citar a assember, sendo que a mesma exigia do programador inúmeras e complexas linhas de instruções, no padrão da linguagem, para se executar uma pequena tarefa. Essa complexidade acabou virando uma necessidade para os programadores, devido que a linguagem de baixo nível, era muito complexa, exigia muito tempo de desenvolvimento dos algoritmos e muitas vezes o código era entendível somente ao autor do mesmo. Atualmente vemos inúmeras linguagens de programação que facilitam muito a vida dos desenvolvedores, como por exemplo, a linguagem Java.

Apesar desta grande evolução das "linguagens de máquina", presenciado por nós através de sua facilidade de aprendizagem e inteligibilidade, as mesmas ainda não solucionaram completamente a problemática da relação entre a máquina e o homem. Justificando-se ao fato de que para um algoritmo funcionar, independente da linguagem de programação, é necessário seguir um padrão de desenvolvimento imposto pelos fabricantes, por exemplo: Caso desenvolvermos um algoritmo e não escrevemos corretamente os comandos, o compilador simplesmente não executará o script.

Atualmente muito tem se ousado para resolver este desafio; eis que surge o PLN. Devido à preocupação em facilitar a comunicação o computador e o ser humano, existem muitos estudos nesta área. Muitos pesquisadores têm ousado, em fazer com que os computadores sejam capazes de aprender a linguagem natural. Sendo que uns dos primeiros experimentos foram na tradução de textos, e os mais ousados ainda criaram programas capazes de interagir verbalmente com o usuário, como por exemplo, os ChatterBots.

Hoje perante a ousadia, vastos estudos e experimentos bem ou mal sucedidos nesta área, faz com que o PLN apresente-se como um campo de estudo bem desenvolvido, porém muito fragmentado, com muito material teórico e diversas idéias de aplicação para a mesma.

Dentre as experiências realizadas na área e muito presente e utilizado nos dias de hoje, podemos destacar a tradução automática de texto. Conforme dito em assuntos anteriores, a mesma pode ser considerada como um dos marcos iniciais para a utilização do computador para a investigação de linguagens naturais.

Os primeiros estudos em PLN começaram pela década de 50, através da distribuição de uma carta escrita por Warren Weaver, que era um grande estudioso na área de criptografia de dados, sendo que esta carta era um convite as universidades e empresas em desenvolver algo que ele chamava de "tradução mecanizada". Apesar da preocupação que a carta passava sobre o estudo nesta área, apenas nos dois anos inicias à sua distribuição é que as universidades começaram a investir nos estudos nesta área (Silva et al. 2007, p8).

A primeira reunião sobre o assunto ocorreu na MIT (Massachusetts Institute of Technology), em 1953, e a primeira demonstração ao publico ocorreu dois anos depois. Esta demonstração consistia em um programa que era capaz de traduzir do Russo para o inglês, apenas 50 frases retiradas de um texto sobre química. O programa possui um dicionário de consulta de apenas 250 palavras e seis regras para a gramática. Após a demonstração a ferramenta foi considerada um sucesso, e muito foi investido pelos financiadores do projeto em aperfeiçoar e aumentar a capacidade da ferramenta (Silva et al. 2007, p8)..

Os sistemas desenvolvidos e existentes nesta época, apesar de serem um sucesso, apresentavam uma tradução de péssima qualidade, devido que o programa apenas listava uma serie de possibilidades de tradução para cada palavra do texto, sendo que não havia nenhuma analise gramatical; assim apesar de facilitar, o texto traduzido exigia muito das pessoas que o utilizavam, devido a ausência de concordância entre as palavras.

Esse falso sucesso do sistema, teve seu primeiro critico em 1964, um grande pesquisador sobre o tema, Bar-Hillel, incentivou constantemente a divulgação de um relatório produzido pelo Comitê Assessor de Processamento Automático das Línguas Naturais. O relatório continha uma avaliação completamente negativa ao sistema de tradução atual, relatando que ate agora não tinha se conseguido tradução automática de textos científicos de forma alguma, pelo fato de que o empreendimento necessitava de pessoas para a pré e pós tradução dos textos.

Este relatório foi um acontecimento catastrófico não somente ao sistema de tradução, mas também a todos os estudo que se iniciavam na área de PLN, devido que após sua divulgação vários dos investidores diminuíram, ou até mesmo cortaram os investimentos.

Após este desastre, os estudos na área se reiniciaram apenas na década de 70, através de Winograd, que desenvolveu um sistema em PLN, que consistia que o usuário digitasse comandos em língua natural em um programa, e os mesmos erram executados por um braço mecânico.

Para promover uma demonstração resumida da evolução do PLN, a tabela 1 lista os principais acontecimentos, descobertas e desenvolvimento na área em cada década.

Década	Evento
Década de 50	Tradução Automática
Década de 60	Novas Aplicações e criação de Formalidades
Década de 70	Retomada e consolidação dos estudos em PLN
Década de 80	Sofisticação dos sistemas existentes
Década de 90	Sistemas baseados em “representações do conhecimento”

Tabela 1 - Evolução do PLN (Silva et al. 2007, p25).

2 FUNCIONAMENTO DOS SISTEMAS DE PLN

Antes de descrever a estrutura dos sistemas de PLN, este capitulo apresenta, como estes sistemas obtêm informações e conhecimentos para tratarem as línguas naturais.

Em PLN o material de entrada para o processamento, são os textos; sendo que para a idealização da analise, o mesmo é divido em partes menores, para facilitar assim na detecção dos fenômenos da língua natural.

2.1 ESTRUTURA LINGÜÍSTICA

Conforme descrito anteriormente, o processador lingüístico de um sistema de PLN recorta o texto em diversas partes, sendo que estes recortes recebem o nome de sentenças(S), ou conhecidos por nós como frases ou orações. Como por exemplo: “O carro segue pela estrada”.

Tendo o texto recortado em sentenças, é possível descrever a sua constituição na língua natural, através da análise estrutural da frase é possível dívida em entidades menores, denominado constituintes, sendo possível organizá-las hierarquicamente. Esta organização recebe o nome de estrutura interna, cuja pode ser representada em forma de estrutura arbórea, a Figura 1 mostra um exemplo estrutura arbórea.

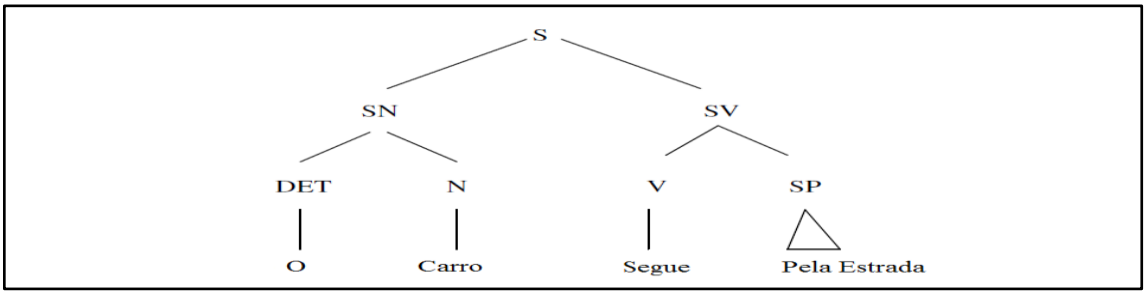

Figura 1 – Estrutura Arbórea (Silva et al. 2007, p16).

Na imagem anterior podemos notar que cada item da hierarquia da sentença foi indicado por SN, SV, SP, DET, N e V. Os elementos intermediários desta estrutura são a expressões que fornecem informações necessárias para composição da sentença, denominados Sintagmas e as de nível baixo definem qual a classe que a palavra esta.

Este esquema estrutural aqui exemplificado é valido para qualquer língua natural, mediante itens.

Segundo Silva et al. (2007):

> Os sintagmas são grupos de palavras organizados em torno de um núcleo sintático que o denomina. Assim, quando o núcleo de um sintagma é um nome (substantivo, adjetivo ou pronome substantivo) falamos de sintagma nominal (SN); quando é um verbo (ou locução verbal), há sintagma verbal (SV); preposição, sintagma preposicional (SP) e advérbio, sintagma adverbial (SAdv). As categorias gramaticais ou sintáticas, por sua vez, refletem as classes nas quais as palavras da língua são organizadas: determinante (DET), nome (N), verbo (V), advérbio (Adv), preposição (P) e assim por diante.

Ainda, Segundo Silva et al. (2007):

> Para a caracterização de cada um dos sintagmas e das categorias gramaticais da estrutura é necessária a compreensão da função de cada elemento na sentença. Para isso o sistema de processamento é alimentado pelos itens lexicais que carregam toda sorte de informações pertinentes para a sua operacionalização, isto é, aos itens lexicais (palavras da língua) são associadas informações de natureza fonético-fonológica, morfológica, sintática, semântica e pragmático-discursiva. A maneira como essas informações são combinadas para a disposição dos itens lexicais na sentença é dada através das diversas regras de tratamento lingüístico das quais falaremos em outras seções.

2.2 NÍVEIS DE PROCESSAMENTO DAS PALAVRAS.

Conforme demonstrado na seção 2.1, as palavras podem ser caracterizadas de varias maneiras, porém de acordo com o estatuto da descrição lingüística.

De acordo com Silva et al. (2007), podemos definir uma palavra pelo seu estatuto, devido ao fato de que as palavras possuem propriedades de natureza distinta, sendo que seu comportamento é refletido quando combinadas com outras palavras, na atividade comunicativa, isto é o sentido. Abaixo segue a forma pela qual podemos definir o estatuto das palavras:

- Fonético-fonológico: O sentido sonoro que a palavra implica na sentença.

- Morfológico: Isolamento das entidades mínimas, as palavras, para compreender a formação e flexão das mesmas.
- Sintático: Compreender a função desempenhada das palavras em uma sentença.
- Semântico: Possibilita a capacidade de se identificar objetos do mundo.
- Pragmático-discursivo: quando a força expressiva das palavras possibilita além da identificação de objetos, podendo-se identificar a quem se está referenciando.

Para maioria dos sistemas de PLN cada um dos níveis acima apresentados, cuja função, possibilitam descrever as palavras, constitui um modulo lingüístico, ou seja a etapa para o processamento da língua natural. Sendo que em cada um desses módulos as informações contidas são processadas de forma a se obter o melhor tratamento lingüístico, não apenas para reconhecimento, mas também para produção de sentenças.

Entretanto da Silva et al. (2007), enfatiza a necessidade das informações de que tratam cada modulo, sejam armazenadas juntamente às formas lingüísticas correspondentes.

3 ARQUITETURA DE UM SISTEMA DE PLN

Para o desenvolvimento de um sistema de PLN, não existe modelo definido a se seguir, o modelo pode variar de acordo com as necessidades da aplicação. Para exemplificar, será demonstrada a estrutura de um sistema de tradução automática.

3.1 ARQUITETURA PROPOSTA A UM SISTEMA DE TRADUÇÃO AUTOMÁTICA

Abaixo segue o que um sistema de tradução automática devera ser capaz de fazer:

- Reconhecer as palavras das sentenças da língua existente no texto de origem:
- Ser capaz de associar os atributos e funções sintáticas correspondentes a cada palavra da sentença, isto é, capacidade de analise sintática.
- Representar a sentença em uma forma intermediaria entre a língua origem e destino, ou seja, uma língua que o sistema agregue além das palavras as suas devidas características.
- Identificar o significado geral da sentença, ou seja, através das varias palavras existentes na mesma, identificar qual o significado; Analise semântica.
- Mapear e associar o significado extraído da linguagem intermediaria.
- Transformação da língua intermediaria na língua de destino.

Podemos perceber que um sistema de tradução automática é bastante complexo, simplesmente pelo fato da fase de interpretação, isto é, a passagem da língua original em uma intermediaria, e a fase de geração, que é a passagem da língua intermediaria na língua destino. A maioria dos sistemas de PLN apresenta somente uma delas, um exemplo a ser dado, é o sistema de consulta a base de dados, ao fato de que na interface, o sistema apenas retorna as perguntas realizadas pelo usuário.

Enfatizando um sistema de tradução automática, Silva et al. (2007) demonstra sua arquitetura na figura 2, em seguida é fornecida uma breve explicação de cada um dos modulo que a constituem.

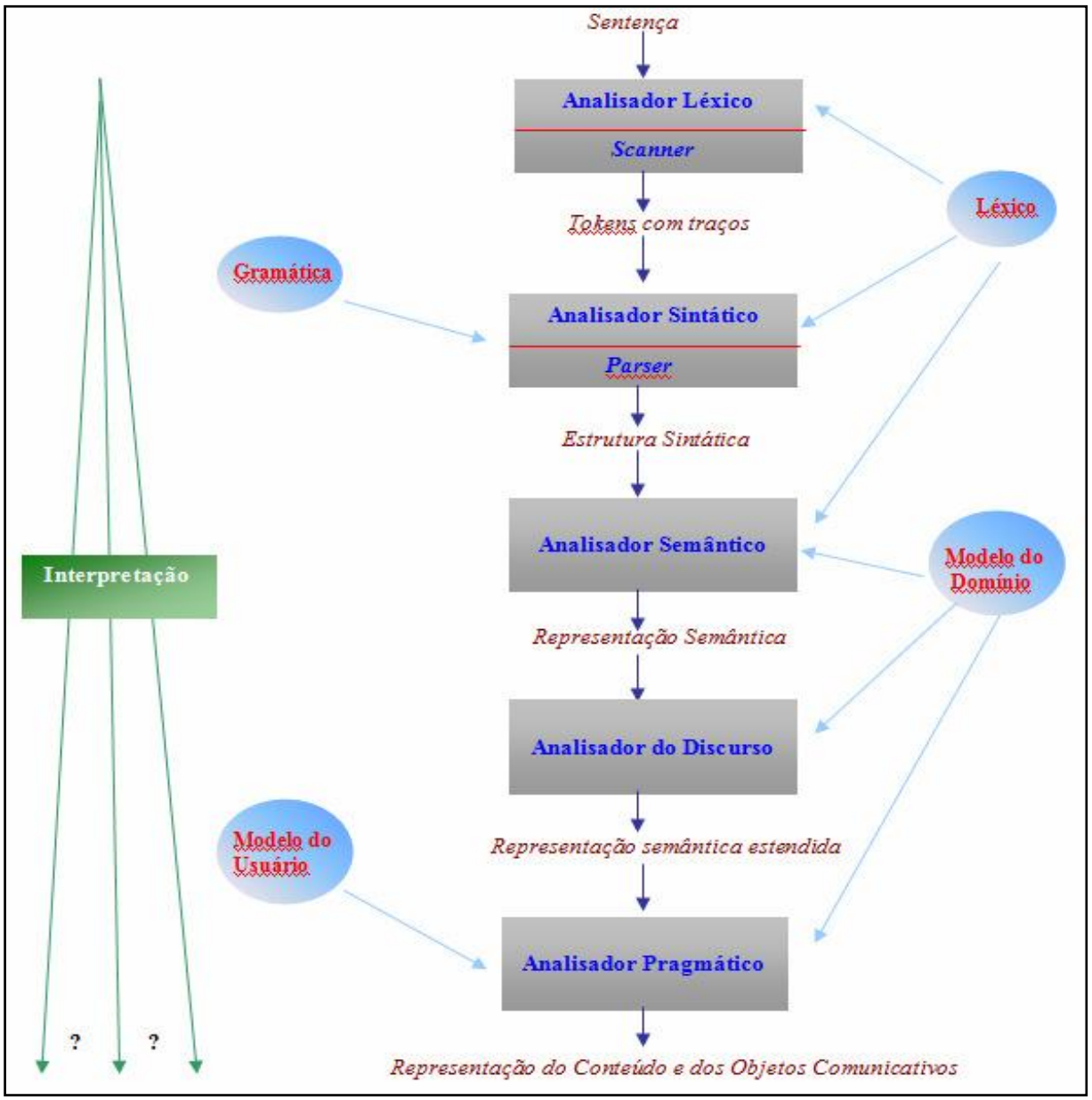

Figura 2 – Estrutura Sistema de Tradução Automática(Silva et al. 2007, p25).

- Analisador léxico (scanner): Processo comumente chamado de tokenização, este processo consiste em identificar e separar os componentes significativos da sentença, envolvendo desde as palavras, bem como a associação de atributos as mesmas. Este processo pode alcançar níveis de complexidade avançados, dependendo somente do resultado que se deseja do mesmo.

- Analisador Sintático (parser): Processo responsável por gerar ou recuperar a estrutura sintática de uma sentença, construído a estrutura baseando-se na gramática da língua de origem; geração da arvore sintática.
- Analisador semântico: Processo responsável pela interpretação da sentença em si ou de sentido global.
- Analisador de discurso: Processo responsável de dar sentido geral ao texto, isto é, faz com que o sentido do texto traduzido apresente uma maior fluidez, ao fato de que ele forma uma ligação de sentidos entre as sentenças antecedentes e procedentes.
- Analisador pragmático: Processo responsável pela identificação do contexto de uma sentença, por exemplo, o mesmo pode identificar se uma sentença interrogativa exige ou não uma resposta.

Apesar de a figura definir certa seqüência de execução dos processos, Silva et al. (2007), defende a idéia de que nem sempre os processos são executados um após o outro.

Silva et al. (2007) exemplifica:

> Considere, p.ex., a sentença "É o pote creme de molho inglês?" (exemplo extraído de Rich and Knight, 1993, p.437). durante sua análise sintática, é preciso decidir qual é o sujeito e qual é o predicado, dentre os três substantivos da sentença *(pote, creme e molho*) e dar a ela o formato "É x y?". Lexicamente, todas as seguintes delimitações da frase *pote creme de molho inglês* são possíveis: o pote, o pote creme, o pote creme de molho, o pote creme de molho inglês, creme de molho inglês, molho inglês, inglês. Entretanto, o processador sintático será incapaz de decidir quais, dentre essas formas, correspondem a estruturas sintáticas válidas, se não contar com algum modelo de mundo em que certas estruturas fazem sentido e outras não. Caso esse modelo exista no sistema automático, é possível obter-se uma estrutura que permita p.ex., a interpretação o pote de cor creme contém molho inglês, e não o pote é creme de molho inglês. Desse modo, as decisões sintáticas dependem da análise do discurso ou do contexto de uso e, portanto, os processos representados na Figura 4.1 interagem entre si. Não é difícil notar que a execução seqüencial dos processos de interpretação simplifica sobremaneira o projeto do sistema, se considerarmos que o resultado de uma fase constitui a entrada para a fase subseqüente. Neste caso, os processos se tornam modulares e, portanto, o controle é menos complexo. As decisões sobre a seqüencialização ou combinação dos processos dependem das características do projeto particular que se tem em mente.

3.2 Considerações Finais do Capitulo

Neste capitulo foi relatado de forma simplificada e resumida a historia do PLN, bem como a uma demonstração de uma estrutura básica de um sistema de tradução automática.

Os próximos capítulos irão demonstrar um dos desafios impostos no desenvolvimento deste trabalho, a construção do corpus, abordando as formas anotação e compilação do mesmo.

4 CORPUS

Para a realização deste trabalho, é necessário criar um corpus, para o sistema a ser desenvolvido possa funcionar. Entretanto antes de descrever uma metodologia de criação de corpus, primeiramente devemos conceitualizar o mesmo.

4.1 CONCEITO DE CORPUS

Segundo Almeida e Aluísio(2006, p.157) podemos definir corpus como: um conjunto finito de enunciados tomados como objeto de análise. Mais precisamente, conjunto finito de enunciados considerados característicos do tipo de língua a estudar, reunidos para servirem de base à descrição e, eventualmente, à elaboração de um modelo explicativo dessa língua. Trata-se, pois, de uma coleção de documentos quer orais (gravados ou transcritos) quer escritos, quer orais e escritos, de acordo com o tipo de investigação pretendido. As dimensões do corpus variam segundo os objetivos do investigador e o volume dos enunciados considerados como característicos do fenômeno a estudar. Um corpus é chamado exaustivo quando compreende todos os enunciados característicos. E é chamado seletivo quando compreende apenas uma parte desses enunciados. (Galisson e Coste, 1983).

Atualmente existem diversos corpus para uso, entretanto nem sempre os corpus disponíveis poderão abranger a necessidade do sistema. Existem inúmeras questões a serem avaliadas ao se utilizar um corpus já existente, como por exemplo, a língua e o assunto que o mesmo foi compilado. Um caso mais especifico se dá ao próprio desenvolvimento deste trabalho, que necessita de um corpus direcionado a língua portuguesa, e que possa tratar os inúmeros desafios impostos pela linguagem resumida e rica em gírias, conforme observado em várias sessões de bate-papo já observadas.

Essa necessidade de mudanças para a concepção de corpus, segundo Almeida e Aluisio(2006, p.157) deve-se a lingüística de corpus, tida como uma abordagem que se ocupa

da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador (Berber Sardinha, 2004).

Podemos dizer que um conjunto de livros, revistas, artigos e noticias de jornais poderiam ser considerados um corpus, entretanto esta idéia não é correta, devido ao fato de que o os dados lingüísticos contido, não estão em formato eletrônico, ou seja, não podem ser processado por um computador.

Tendo essa noção de o que é um corpus e como ele deve estar apresentado, Almeida e Aluísio(2006, p.157) definem 4 características fundamentais:

- Amostragem e representatividade: o corpus deve ter uma amostragem ou variedade da língua que quer ser analisada, suficiente para se obter um grau satisfatório de representatividade;
- Tamanho finito: o corpus deve ser uma quantidade finita, por exemplo, 500 mil palavras, entretanto o Corpus-Monitor é uma exceção, ao fato de que ao receber mais textos, ele possa ser maior, tendo uma amostragem maior.
- Formato eletrônico: O mesmo deve estar em forma eletrônico, ou seja, em arquivo eletrônico, para ser processado em maior velocidade pelo computador;
- Referencia padrão: O corpus deve conter um referencia, para que o mesmo alcance um grau de representatividade maior no assunto em que o mesmo foi compilado. Por exemplo: um corpus compilado com textos eruditos e poéticos, não terá grau de representatividade satisfatório ao se analisar textos técnicos na área da medicina;

Em relação às quatro características descritas anteriormente podemos destacar a questão do formato eletrônico, que não basta ele estar apenas gravado em arquivo, o mesmo deve estar compilado e arquivado, de forma que os sistemas possam ser utilizados de forma rápida, tendo fácil acesso as dados nele contidos. Tal cenário podemos citar a WEB, que apesar de estar em formato eletrônico, a mesma ainda se apresenta em formato de artigos ou meramente textos, sendo assim precisariam de tratamento para tornar-se um corpus.

O formato eletrônico do corpus é um caso a parte, na questão de que a partir dos anos 90, terem sido um ano de grande evolução na questão de estudos sobre PLN. Tendo uma

grande evolução, as ferramentas, podemos destacar, também as mesmas voltadas a língua portuguesa.

Tal evolução no desenvolvimento das ferramentas, e principalmente no surgimento do corpus podemos ter uma noção quase que precisa do comportamento lingüístico existentes em texto, não sendo necessário uma pré e pós avaliação da análise por pessoas. Sobretudo por meio de um corpus podemos observar informações morfológicas, semânticas, sintáticas e discursivas; que são peças fundamentais em uma pesquisa. Além disso, podemos, podemos explicar o emprego das palavras nos textos, os sentidos que a mesma exerce, além de descobrir fatos novos da língua em questão, cujas quais são de difícil intuição a um ser humano.

4.2 Questões Relevantes para o Desenvolvimento de um Corpus.

Para o desenvolvimento do corpus para o projeto, faz-se necessário a observação de alguns pontos para sua compilação, para que o mesmo atinja um nível de confiabilidade, necessário para sua utilização na pesquisa. Dentre os pontos a se observar podemos citar os seguintes: autenticidade, representatividade, balanceamento, amostragem, diversidade e tamanho, cujos os quais terão seus detalhes descritos abaixo:

- Autenticidade: os textos que formarão o corpus devem ser textos em linguagem natural, alem de que devem ser escritos por pessoas de linguagem nativa com a que se esta desenvolvendo o corpus;
- Representatividade: Os textos utilizados deverão ter uma representatividade com o assunto que se destina a pesquisa, no caso o desenvolvimento deste trabalho, será necessário utilizar textos “as conversas” das sessões de bate-papo de um determinado assunto. E como já comentado anteriormente, não se tem um nível satisfatório no sistema, ao se utilizar um corpus formado por poesias e textos épicos, para pesquisa nas áreas técnicas como medicina por exemplo.
- Balanceamento: Muito parecido com a representatividade, o balanceamento reforça a questão de se escolher o assunto que o corpus abrange, por exemplo,

seria confuso misturar diversas áreas em um corpus, sendo assim poderia se confundir o sistema de PLN, no caso de avaliação de uma determinada palavra ser utilizada de uma certa maneira em um tema, e de um jeito completamente deferente em outro.

- Amostragem: Apesar de a construção do corpus necessitar focar um tema, o mesmo não impede que os textos sejam retirados apenas de uma fonte, as mesmas podem variar desde livros, notícias, revistas e artigos acadêmicos.
- Diversidade: Apresenta muita semelhança com a amostragem; a diversidade reforça a idéia de se retirar os textos de meios de propagação da informação e conhecimentos, devido ao fato de enriquecer o corpus pelo fato do uso diferenciados na colocação das palavras nos textos.
- Tamanho: O corpus que se deseja desenvolver deve ter uma quantidade adequada para a realização de pesquisa na área em que se focar. Esta quantidade não se resume apenas a quantidade de palavras ou tipos da mesma. O corpus deve contemplar uma quantidade diversificada de textos, de diversas fontes, bem como autores, tipos discursivos e datas. Dependendo do tipo de pesquisa que se deseja realizar, existem corpus de 100.000 palavras que dão conta do recado, entretanto existem outras pesquisas, que necessitam de um numero maior. Atualmente o corpus de maior diversificação é o da Bank of English, com mais de 530 milhões de palavras.

Para Almeida e Aluisio(2006, p.157) a elaboração de um corpus é um processo que avança em ciclos: inicia-se a escolha de textos baseada em critérios externos culturalmente aceitos (tipologia de gêneros e tipos de textos, por exemplo), depois se prossegue com investigações empíricas da língua ou variedade lingüística sob análise (também denominados critérios internos) e, finalmente, procede-se com a revisão de todo o projeto(Biber, 1993) .

4.3 Etapas para a Compilação de um Corpus

Para o desenvolvimento de um corpus é necessário que seja seguido algumas etapas, entretanto lembrando que as etapas abaixo descritas não são necessariamente obrigatórias, ou

que estão formalizadas em algum site de PLN, instruindo que a compilação do corpus deve seguir a metodologia a ser apresentada abaixo.

4.3.1 Seleção do texto

Para a compilação de um corpus, a primeira etapa a ser executada é a seleção de textos, sendo que esta etapa deve se ter conhecimentos em mente as questões relevantes para o desenvolvimento de corpus, conforme descrito na seção 4.2. Tendo conhecimentos das questões anteriores, podemos selecionar os textos que farão parte do corpus e passar ao próximo passo descrito na seção 4.3.2.

4.3.2 Compilação e manipulação

Segundo Almeida e Aluisio(2006, p.160), a compilação consiste no armazenamento em arquivos predeterminados de todos os textos selecionados.

Os textos que irão formar o corpus poderão ser provenientes de diversas partes, como por exemplo, revistas, jornais, artigos e livros; sendo que estes deverão ser digitalizados por meio de um scanner com OCR. Existe também a possibilidade de se obter textos da WEB, sendo que os mesmo devem passar por programas que eliminem as TAGS em HTML (Hyper Text Markup Language). O corpus a ser desenvolvido terá seus textos extraídos em sessões de bate-papo cujos os quais contem tais TAGS em HTML.

A manipulação do corpus consiste em tratar os textos selecionados, por exemplo, textos contidos em arquivos PDF (Portable Document Format), DOC e dentre outros devem passar por uma transformação de sua extensão de arquivos atual para arquivos de texto puro. Tendo o os textos, devemos organizá-los em arquivos, seguindo um padrão de nomeação dos mesmos, para facilitar a sua manipulação.

4.3.3 Direitos autorais

No caso da compilação de um corpus pessoal, com algum já existem, ou até mesmo se utilizar de um corpus já existente, devemos esta ter a autorização para a utilização dos mesmos, junto aos criadores do mesmo.

4.3.4 Anotação

A anotação consiste basicamente em duas representações de informação: a anotação estrutural e a anotação lingüística.

Almeida e Aluisio (2006, p.160) definem a *anotação estrutural*:

> [...] compreende a marcação de dados externos e internos dos textos. Como dados externos entendemos a documentação do corpus na forma de um cabeçalho que inclui os metadados textuais (ou dados estruturados sobre dados), isto é, dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, a tipologia textual e informação sobre a distribuição do corpus. Como dados internos temos a anotação de segmentação do texto cru, que envolve: a) marcação da estrutura geral – capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras, e b) marcação da estrutura de subparágrafos – elementos que são de interesse lingüístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e ênfases tipográficas do tipo negrito, itálico, sublinhado, etc.

Sendo que tendo estas informações, fica fácil no momento do processamento, selecionar um trecho de determinado autor ou data de publicação entre outras possibilidades.

Ainda, Almeida e Aluisio(2006, p.160) definem também a *anotação lingüística*:

> [..] pode ser em qualquer nível que se queira, isto é, nos níveis morfossintático, sintático, semântico, discursivo, etc., sendo inserida de três formas: manualmente (por lingüistas), automaticamente (por ferramentas de Processamento de Língua Natural – PLN) ou semi-automaticamente (correção manual da saída de outras ferramentas). Essa última é comprovadamente mais eficiente, pois revisar é mais rápido e gera dados mais corretos do que anotar pela primeira vez.

Tendo em mente o conceito, e as formas de estrutura de anotação de um corpus, o capitulo seguinte demonstrará algumas das ferramentas existentes e ainda demonstrara um método que vem se consagrando em questões de anotação, o uso de XML (Extensible Markup Language).

4.4 TAGSETS

TagSets, são um conjunto de etiquetas utilizadas no processo de anotação de um corpus, nesta seção será demonstrado dois exemplo de TagSets, disponíveis para a língua portuguesa.

4.4.1 NILC

Tagset proposto por AIRES, 1998 para o português que possuem várias formas de classificação para as funções sintáticas, facilitando o processo de parsing, devido ao fato de fornecer *tags.*

Adjetivo	ADJ
Advérbio	ADV
Artigo	ART
Número Cardinal	NC
Número Ordinal	ORD
Outros Números	NO
Substantivo Comum	N
Nome Próprio	NP
Conj. Coordenativa	CONJCOORD
Conj. Subordinativa	CONJSUB
Pronome Demonstrativo	PD
Pronome Indefinido	PIND
Pronome Oblíquo Átono	PPOA
Pronome Pessoal Caso Reto	PPR
Pronome Possessivo	PPS
Pronome Relativo	PR
Pronome Oblíquo Tônico	PPOT
Pronome Interrogativo	PINT
Pronome Apassivador	PAPASS
Pronome de Realce	PREAL
Pronome Tratamento	PTRA
Preposição	PREP
Verbo Auxiliar	VAUX
Verbo de Ligação	VLIG

Verbo Intransitivo	VINT
Verbo Transitivo Direto	VTD
Verbo Transitivo Indireto	VTI
Verbo Bitransitivo	VBI
Interjeição	I
Locução Adverbial	LADV
Locução Conjuncional	LCONJ
Locução Prepositiva	LPREP
Locução Pronominal	LP
Palavra Denotativa	PDEN
Locução Denotativa	LDEN
Palavras ou Símbolos Residuais	RES
.	.
:	:
-	-
(	(
!	!
?	?
...	...
)	)
"	"
[	[
]	]
{	{
}	}
'	'
,	,

Tabela 2 – Lista de TagSet da NILC (AIRES, 1998)

4.4.2 VISL

O conjunto de *tags* do VISL, proposto COLOCARAKI por é bem completo e proporciona boa classificação individual das palavras. Além das *tags* principais há outras que possibilitam fornecer informação morfológica sobre a palavra, tal como gênero, número, pessoa, etc.

Classificação	TAG
$	Pontuação
Adj	Adjetivo
Adv	Advérbio
Det	determinador (artigo+pronome)
In	Interjeição
Kc	conjunção coordenativa
Ks	conjunção subordinativa
N	Substantivo
Num	Numeral
Pers	pronome pessoal
Prop	substantivo próprio
Prp	Preposição
Spec	Especificador
V	Verbo
v-cond	não informado
v-fut	verbo no futuro
v-ger	verbo no gerúndio
v-imp	não informado
v-impf	não informado
v-inf	verbo no infinitivo
v-mqp	verbo no pretérito mais-que-perfeito
v-pcp	verbo no particípio
v-pr	não informado
v-os	não informado
v-ps/mqp	não informado

Tabela 3 – Lista de TagSet - VISL

5 FERRAMENTAS PARA ANOTAÇÃO

Sabemos muito bem que a língua portuguesa é uma das línguas mais ricas e complexas, entretanto a quantidade de ferramentas para o PLN na língua, é muito pequena, algumas ferramentas podem ser conferidas no site da Linguateca (http://www.linguateca.pt/ferramentas.html).

Conforme descritos os conceitos e etapas para a compilação de um corpus anteriormente, este capitulo descreve algumas das ferramentas para a anotação do corpus, que se dá de forma automática ou manual. A seguir serão descritas ferramentas de cada uma das maneiras:

5.1 ANOTAÇÃO MANUAL - MMAX

A ferramenta MMAX, é uma ferramenta de anotação manual de informações em textos e diálogos. A ferramenta possui uma interface gráfica, pela qual o usuário pode partes de texto que o usuário achar ter uma relação anafórica, para então serem anotadas, gerando como saída um arquivo XML chamado Markables, com os textos anotados. Sendo que na estrutura deste arquivo, existem os seguintes atributos:

- id: identificador da marcação
- span: início e fim do sintagma nominal
- pointer & meber: codificam informações a respeito dos markables.

5.2 ANOTAÇÃO AUTOMÁTICA – PALAVRAS

O analisador sintático PALAVRAS permite, que a anotação do corpus de língua portuguesa seja feita de modo automática. O texto de saída após a analise, possuem associações de informações tais como substantivo, verbo, adjetivo, preposição - suas flexões de gênero e número, e, ainda, em alguns casos, seu tipo semântico). Ainda o Analisador PALAVRAS pois três formatos saídas diferentes. A primeira se apresenta-se graficamente, em forma arbórea, representando o texto, a segunda assemelhando-se muito com a primeira, apresenta em formato texto, e a terceira igual a forma da segunda, entretanto estruturado em tags XML (formato Tiger-XML)

6 EXTRAÇÃO DA INFORMAÇÃO

O presente capitulo tem como o objetivo, demonstrar algumas das heurísticas, para a extração de informações, as mesmas terão seus conceitos descritos nos sub-capitulos subseqüentes.

6.1 PLN E XML

Atualmente temos percebido que o XML, tem oferecido muitas vantagens quanto a sua utilização, na vida dos desenvolvedores de software. Na area do PLN, o estrutura oferecida pelo XML, pode apresentar grandes funcionalidades no desenvolvimento de ferramentas. Ao fato que alem de poder estruturar as informações, pode-se anexar outras informaçoes as mesmas, e ainda, de forma organizada e padronizada.

A partir desses inúmeros benefícios oferecidos pelo XML, Gasperin et al (2003), desenvolveram uma ferramenta capaz de simplificar a extração e o uso das informações, oferecidas pelo parser PALAVRAS, o PALAVRAS Xtrator. Através da saída gerada pelo parsser PALAVRAS, o Xtrator cria três arquivos XML. As funcionalidades de todos os arquivos estão descritos na tabela 4.

Arquivo	Funcionalidade
WORDS	Neste arquivo, cada palavra, possui uma representação, por um elemento Word, e cada elemento possui um ID
POS (Part-Of-Speech)	Este arquivo contêm as informações morfossintáticas de cada palavra.
CHUNKS	Este arquivo apresenta as informações sintáticas sobre a estrutura do texto.

Tabela 4 – Funcionalidade Arquivos de Saída Palavras Xtrator

Gasperin et al (2003, p5) exemplificam a estrutura de cada arquivo, através da frase: “Três acidentes graves marcaram o fim de semana.”, abaixo são exibidas figuras com a estrutura de cada arquivo, com a frase mencionada.

```
<words>
  <word id="word_1">Três</word>
  <word id="word_2">acidentes</word>
  <word id="word_3">graves</word>
  <word id="word_4">marcaram</word>
  <word id="word_5">o</word>
  <word id="word_6">fim_de_semana</word>
  <word id="word_7">.</word>
</words>
```

Figura 3 - Arquivo Words (GASPERIN et al. 2003)

```
<words>
  <word id="word_1">
    <num canon="três" gen-
der="M" number="P">
      <secondary_num tag="card"/>
    </num>
  </word>
  <word id="word_2">
    <n canon="acidente" gen-
der="M" number="P"/>
  </word>
  <word id="word_3">
    <adj canon="grave" gen-
der="M" number="P"/>
  </word>
  <word id="word_4">
    <v canon="marcar">
      <fin tense="PS/MQP" per-
son="3P" mode="IND"/>
    </v>
  </word>
  <word id="word_5">
    <art canon="o" gender="M" num-
ber="S">
      <secondary_art tag="artd"/>
    </art>
  </word>
  <word id="word_6">
    <n canon="fim_de_semana" gen-
der="M" number="S"/>
  </word>
</words>
```

Figura 4 - Arquivo POS (GASPERIN et al. 2003)

```
<paragraph "paragraph_1">
  <sentence id="sentence_1" span="word_1..word_14">
    <chunk id="chunk_1" function="subj" form="np" span="word_1..word_3">
      <chunk id="chunk_2" function="h" form="n" span="word_2"/>
    </chunk>
...
  </sentence>
</paragraph>
```

Figura 5 - Arquivo Chunks (GASPERIN et al. 2003)

6.2 EXTRAÇÃO DE INFORMAÇÃO A PARTIR DE MÉTODOS ESTATÍSTICOS

Os seguintes itens descreverão dois algoritmos (EPC-P & EPC-R) propostos por Nunes et al. (2000), para a extração de palavras chave.

Nunes et al. (2006, p.160) defendem de que a extração de palavras chaves:

> [...]podem ser úteis em diversas aplicações computacionais, em especial aquelas que necessitam indexar documentos para buscas posteriores. [...] Técnicas extrativas baseadas na seleção de frases do texto são consideradas simples se comparadas a técnicas que incluem compreensão de texto (e, portanto, muito complexas), mas podem ser sofisticadas com algum conhecimento lingüístico e assim alcançarem índices razoáveis de eleição de palavras-chave.

6.2.1 Extração Baseada em freqüência de padrões

Tendo delimitado a língua e o gênero dos textos que serão utilizados na analise, a eficácia deste método se torna mais eficaz, ao fato que temos alguns padrões pré-estabelecidos. O método pode ser chamado de EPC-P, sendo que o mesmo efetua uma busca de palavras que casam com padrões morfossintáticos. Resumindo sua funcionalidade, o método encontra todas as frases que apresente uma ligação, com os padrões pré-definidos, tendo uma listagem das frases, são utilizados métodos estatísticos, filtrando somente a que possuem uma maior relevância.

6.2.2 Extração baseada em freqüência de Radicais

Este método também conhecido com EPC-R, apresenta uma grande semelhança ao descrito anteriormente.

Este método é baseado no algoritmo de Extractor [Turney, 1999], tendo seu funcionamento de maneira bem simples, o mesmo utiliza somente a freqüência de radicais no texto, não importando se os mesmos se encaixam nos padrões predeterminados. Entretanto este método se diferencia, pelo uso constante de listas de stop-words, para remoção das palavras irrelevantes, e também pela formação de três listas de palavras; a primeira para palavras simples, a segunda para as duplas e a terceira as triplas e juntamente a estas listas são armazenadas as freqüência que cada uma apareceu no texto.

6.2.3 Algoritmo Extrator

Conforme descrito anteriormente o algoritmo Extractor, é um algoritmo de extração de palavras-chave, levando em conta somente a freqüência dos radicais em um texto. Podemos exemplificar da seguinte maneira, o algoritmo conta quantas vezes as seqüências de palavras sejam elas simples, duplas ou triplas, apareceram no texto. O algoritmo pode fazer uso de uma lista de stop-words, ou seja, uma lista contendo as palavras irrelevantes ao estudo, como por exemplo, pontuação, preposições entre outras palavras de classe fechada.

CONCLUSÃO

Neste presente trabalho é realizado um estudo e análise do material bibliográfico adquirido durante as pesquisas. Este estudo proporcionou um maior compreendimento sobre PLN, bem como a área de extração de informações.

Sem dúvida o PLN é uma área admirável e impressionante, pelo fato como a mesma pode tratar a linguagem natural, proporcionando uma serie de soluções para o que se deseja desenvolver, no que se diz respeito ao Processamento da Linguagem Natural. Neste trabalho pode-se destacar que um dos maiores desafios a serem superados é o desenvolvimento de um corpus adaptado para salas de bate-papo, sendo que o mesmo é inédito, ao caso que em nao foi localizado nada assemelhado em núcleos de pesquisa sobre PLN. Apesar de se ter uma quantidade razoável de material para o desenvolvimento deste trabalho, percebesse que é necessário ampliar as pesquisas tanto na área de PLN, quanto de Extração de Informações, para idealiza completa do trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

AIRES, Rachel. Nilc TagSets. 1998. Disponível em: http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html. Acesso em 18 de junho de 2009.

ALUÍSIO, Sandra Maria; ALMEIDA, Gladis Maria de Barcellos. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística.** Dezembro de 2006. Disponível em:< http://www.icmc.usp.br/~taspardo/NILC-TR-09-08.pdf >. Acesso em 30 de maio de 2009.

BRUCKSCHEN, Miriam et al. **Anotação Lingüística em XML do Corpus PLN-BR.** Núcleo Interinstitucional da Lingüística Computacional: Agosto de 2007. Disponível em: < http://www.icmc.usp.br/~taspardo/NILC-TR-09-08.pdf >. Acesso em 15 de maio de 2009.

HERNANDES, Carlos A. M.; SANTANA Roberto A. S.; FALCÃO Sérgio D. Sobre o uso do chat como ferramenta auxiliar de ensino e aprendizagem no curso de Mestrado em Informática da Universidade Católica de Brasília. Universidade Católica de Brasília. Disponível em: < http://carlosmamede.org/Artigo%20sobre%20chat%20na%20UCB%20-%20publicado.pdf >. Acesso em 15 de Março de 2009.

GASPERIN, C.; Vieira, R.; GOULART, R.; P.QUARESMA**. Extracting XML Syntactic Chunks from Portuguese Corpora**. TALN 2003,. Disponível em: <http://www.rodrigo.goulart.nom.br/publicacoes/gasperin2003a.pdf>. Acesso em 17 de Maio de 2009.

SILVA, Bento C. D.; MONTILA, Gisele; RINO, Lucia H. M.; SPECIA, Lucia; NUNES, Maria das G. V.; OLIVEIRA JR, Osvaldo N.; MARTINS, Ronaldo T.; PARDO, Thiago A. S., **Introdução ao Processamento das Línguas Naturais e Algumas Aplicações**. Núcleo Interinstitucional da Lingüística Computacional: Agosto de 2007. Disponível em: < http://www.icmc.usp.br/~taspardo/NILCTR0710-DiasDaSilvaEtAl.pdf >. Acesso em 12 de Fevereiro de 2009.

SYMBOLSET MANUAL. 2009. Disponível em: http://beta.visl.sdu.dk/visl/pt/info/symbolset-manual.html. Acesso em 18 de junho de 2009.