UNIVERSIDADE FEDERAL DE RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
CURSO DE ENGENHARIA DE COMPUTAÇÃO

Giovana Jaskulski Gelatti

Text-mining Aplicado a Geração de uma Interlíngua Português-LIBRAS

Rio Grande, 2013

UNIVERSIDADE FEDERAL DE RIO GRANDE
CENTRO DE CIÊNCIAS COMPUTACIONAIS
CURSO DE ENGENHARIA DE COMPUTAÇÃO

Giovana Jaskulski Gelatti

Text-mining Aplicado a Geração de uma Interlíngua Português-LIBRAS

Monografia apresentada como requisito parcial para obtenção de título Bacharel em Engenharia de Computação ao Centro de Ciências Computacionais da Universidade Federal de Rio Grande.

Orientadora : Dra. Karina Machado
Co-orientadora: Me. Carla Imaraya Meyer de Felippe

Rio Grande, 2013

LISTA DE SIGLAS E ABREVIATURAS

ADA	*American with Disabilities Act*
BIAP	*Bureau International d'Audiophonologic*
CAA	Comunicação aumentativa e alternativa
CSA	Comunicação aumentativa suplementar e alternativa
KDD	*Knowledge Discovery in Databases*
LIBRAS	Língua Brasileira de Sinais
PLN	Processamento da Língua Natural
RI	Recuperação de Informação
SIGNUM	Palavra do latim, significa sinal, é o nome dado ao Projeto que este trabalho faz parte
TA	Tecnologia(s) Assistiva(s)

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

Agradecimentos

Este trabalho foi apresentado como trabalho de conclusão do curso de Engenharia de Computação da Universidade Federal do Rio Grande. Agradeço primeiramente a orientadora deste trabalho, Karina Machado, pelo apoio, dedicação e incentivo a realização do mesmo. A tua orientação foi primordial para a realização do trabalho. Obrigada por me apresentar a área de mineração de dados, a qual me identifiquei e pretendo seguir contribuindo. A co-orientadora Carla de Felippe e, em seu nome, ao Núcleo de Estudos e Ações Inclusivas da FURG por expor o problema e incentivar a solução guiando-me e inserindo-me no meio da acessibilidade.

Muitas pessoas estimularam e incentivaram a entrada e permanência no curso de Engenharia de Computação. Ao ingressar, as dificuldades de começar em uma nova cidade, cursar engenharia apareceram logo após a mudança. Permanecer no desenvolvimento de um sonho deve-se ao apoio da minha família. A eles, o meu agradecimento mais profundo por me mostrar os caminhos e então confiar para que eu mesma construísse e descobrisse os meus, pelo apoio e compreensão nos momentos de dedicação ao curso e a este trabalho e por sempre estarem presentes, apesar da distância.

Aos colegas de turma, pelos cinco anos que cursamos e pela amizade durante estes. Aos professores, grandes mestres, que admiro e respeito pelo ensinamento passado que me proporcionou chegar ao final do curso.

Aos membros do projeto SIGNUM, por acreditarem na execução dele, pela dedicação no desenvolvimento e discussões técnicas para encontrar a melhor solução exequível para o projeto. A tradutora e intérprete Cristiane Fernandes, pela revisão das regras LIBRAS e por aceitar participar do projeto.

Ao João Silva e Sara Silveira do grupo NLX da Universidade de Lisboa, pela disponibilização de sistemas utilizados no trabalho e consideração ao se prontificar na solução de um problema no sistema quando este se encontrou indisponível.

As empresas TokenLab e ETEG Tecnologia da Informação pela compreensão e apoio, viabilizando a disponibilização de tempo na empresa, para a conclusão deste trabalho.

A todos que acreditaram e contribuíram na realização deste trabalho.

Resumo

O trabalho tem finalidade de ser um recurso de acessibilidade na leitura de páginas web por pessoas com as necessidades específicas de surdez e/ou mudez. Para usuários surdos e mudos brasileiros, a utilização e leitura de páginas web são dificultadas, pois a sua língua é LIBRAS (Língua Brasileira de Sinais). Com a finalidade de facilitar a leitura de páginas web para falantes LIBRAS, foi criado o Projeto SIGNUM. Este projeto tem objetivo de realizar a tradução de páginas web em português para LIBRAS através de uma extensão para o navegador Mozilla Firefox. São apresentados trabalhos relacionados como contribuição e inspiração do trabalho. Para traduzir o texto em português para LIBRAS, que são línguas muito diferentes, foi feito um processamento do texto utilizando técnicas de *text-mining*, ferramentas de Tagger e Parser e a implementação de regras que baseiam-se nestas técnicas, assim como a pesquisa, definição e análise destas. Elas também podem ser utilizadas para encontrar os principais termos, simplificar as frases e reorganizar a sintaxe. O trabalho desenvolvido faz o processamento do texto e tem como saída uma interlíngua Português-LIBRAS. Esta interlíngua é um texto onde cada palavra ou termo significa um gesto em LIBRAS. Por fim, foram feitos estudos de caso para avaliar a ferramenta desenvolvida. A avaliação foi feita com linguistas falantes LIBRAS a partir do uso da ferramenta em textos simples.

1. Introdução

A tecnologia tem como objetivo facilitar a vida dos usuários. Assim sendo, as tecnologias assistivas auxiliam pessoas com necessidades específicas, tornando possível e/ou facilitando tarefas diárias e aumentando a independência. De acordo com Carmo (2005), são detalhes que auxiliam a vida de muitas pessoas, principalmente a das pessoas com necessidades específicas, quando precisam lidar com computadores.

Segundo Vygotsky (1989), a linguagem é uma ferramenta psicológica importante para organização do pensamento: "*Uma vez que as crianças aprendem a usar, efetivamente, a função planejadora de sua linguagem, o seu campo psicológico muda radicalmente*", . Por este conceito vê-se a necessidade de possuir uma linguagem eficaz, que possa ser compreendida pelas pessoas falantes dela.

O usuário que apenas se comunica por meio da linguagem LIBRAS possui dificuldade em ler textos em português pois são línguas diferentes. Como explica Moura (2008), LIBRAS é de modalidade visuo-espacial, marcada por uma estrutura simultânea, e Português é de modalidade oral-auditiva (devido à ligação com a oralidade das palavras). Sendo assim, o texto português é compreendido através de fonemas que as palavras produzem. Com este recurso dificultado, as pessoas com necessidade específica auditiva (surdez) e/ou oral (mudez) acabam não lendo o texto ou aqueles que foram coagidos à alfabetização em português, possuem dificuldades em entender o texto não traduzido para LIBRAS.

Dessa forma, a leitura de textos em português tanto em livros, apostilas, revistas quanto em páginas Web é difícil para usuários com necessidade específica auditiva e/ou oral que foram alfabetizados e utilizam LIBRAS ou para os que a alfabetização na língua portuguesa foi forçada. Por isso, atualmente existem tradutores e intérpretes de LIBRAS e pesquisas envolvendo o estudo que auxilie este público em especial.

Entre as ferramentas estudadas, não foi encontrada uma que permitisse seu uso em conjunto com os navegadores Web atuais e que facilmente os usuários pudessem traduzir textos da internet para a linguagem LIBRAS gratuitamente. A partir deste conceito surgiu o projeto SIGNUM (palavra do latim, significa sinal). O objetivo do projeto é desenvolver um software que possa ser utilizado em conjunto com navegadores Web onde o usuário falante LIBRAS possa ler textos da Web em português por meio de um vídeo que exibe o conteúdo do texto na linguagem LIBRAS.

A ideia do projeto surgiu primeiramente para auxiliar os falantes de LIBRAS para a leitura de páginas web. Ao decorrer do desenvolvimento foi visto que o escopo pode ser maior que o objetivo inicial. Com a extensão falantes da língua portuguesa também podem aprender uma nova língua, LIBRAS. Abrangendo não apenas a área de computação, o SIGNUM é um projeto que foi construído interdisciplinarmente com integrantes das áreas de sistemas de informação, letras e psicologia.

Com a integração destas áreas, foi possível identificar o problema e oferecer um tradutor

de textos em LIBRAS para Web , propor soluções e verificar qual seria a mais apropriada para falantes LIBRAS. Para alcançar o objetivo de exibir no navegador Web o conteúdo do texto na forma de um vídeo LIBRAS, o projeto SIGNUM propõe a execução de duas etapas bem distintas. Após o usuário selecionar um texto em seu navegador Web, a primeira etapa consiste no processamento do texto selecionado, gerando um texto em uma interlíngua Portugês-Libras. O segundo passo, a partir do texto já no formato de Interlíngua português-libras, consiste em gerar um vídeo composto pelos termos do texto selecionado, na ordem correta e seguindo as regras gramaticais da linguagem LIBRAS.

1.1 Objetivo geral

No contexto do projeto SIGNUM, este trabalho de conclusão de curso tem por principal objetivo o processamento do texto de uma página Web, ou seja, a primeira etapa deste processo de exibição de um vídeo do texto selecionado no navegador Web. Dessa forma, este trabalho propõe uma metodologia para a conversão de um texto na língua portuguesa para a interlíngua português-LIBRAS.

A conversão do Português para LIBRAS é dificultada por serem línguas parecidas, mas com estruturas diferentes. A solução para esta tradução Português-LIBRAS é o que será apresentado no trabalho. O resultado esperado é que cada termo possa ser relacionado a um sinal em LIBRAS. E, para alcançar esse objetivo, são aplicadas técnicas de *text-mining* ou mineração de texto.

A mineração de texto consiste em um conjunto de técnicas aplicadas para processar um texto onde são realizadas várias funções de busca, análise linguística e categorização, (CHEN, 2001). Para o desenvolvimento deste trabalho foram estudadas ferramentas já existentes que se assemelham com o escopo do trabalho, analisando cada uma delas em seus objetivos, pontos negativos e positivos. Também foram estudadas ferramentas que podem auxiliar na mineração do texto e, a partir das informações obtidas destas ferramentas, realizar a recuperação de texto para a produção da interlíngua Português-LIBRAS.

1.2 Objetivos específicos

Define-se como objetivos específicos:

- Permitir a falantes de LIBRAS a leitura de textos em português em páginas web a partir de um recurso a ser instalado em seu navegador
- Contribuição para a área de mineração de texto, em especial, para a geração de uma interlíngua português-LIBRAS
- Definir regras para a tradução de português para LIBRAS

1.3 Organização do documento

Para iniciar a apresentação do trabalho, no capítulo 2 apresenta-se o porquê do desenvolvimento identificando a sociedade surda no Brasil e no mundo, a sua língua (LIBRAS) e a apresentação do que é mineração de texto e suas técnicas. Em seguida, no capítulo 3, são descritos os trabalhos relacionados, como dicionários LIBRAS online e ferramentas de tradução automática.

O capítulo 4 descreve os materiais e métodos utilizados no desenvolvimento deste trabalho. Neste capítulo são descritas as técnicas de mineração de texto que foram utilizadas e a ferramenta que auxiliou na execução de cada etapa. Apresenta-se alguns conceitos que serão utilizados ao decorrer do trabalho, como o uso de *Tagger* e *Parser*.

O capítulo 5 apresenta os resultados no ambiente em que ele foi proposto, situando-o no projeto SIGNUM e analisando seu comportamento no mesmo. Nesta seção pode ser analisada a arquitetura individual do trabalho na seção 5.1, as regras de tradução LIBRAS na seção 5.2 e como foi implementado na seção 5.3.

Para a finalização, tem-se os capítulos 6 e 7. No capítulo 6 são apresentados os estudos de caso com exemplos de frases traduzidas para a interlíngua gerada pelo trabalho, o funcionamento da extensão desenvolvida no Projeto SIGNUM, e este trabalho atuando na extensão. E, no capítulo 7, as conclusões sobre o trabalho e aspectos apresentados no capítulo 6 e indicação para trabalhos futuros.

2. REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos necessários para o entendimento deste trabalho. Na seção 2.1 é definido o termo tecnologia assistiva e sua origem. Como neste trabalho o foco é o desenvolvimento de uma ferramenta para pessoas com necessidades específicas auditiva e oral que se comunicam por LIBRAS, a sociedade Surda e/ou Muda é apresentada na seção 2.2. Uma descrição mais detalhada sobre a linguagem LIBRAS é encontrada na seção 2.3. Os conceitos sobre mineração de dados, especificamente mineração de texto são detalhados na seção 2.4 e são importantes para o entendimento da implementação deste trabalho.

2.1 TECNOLOGIA ASSISTIVA

Atualmente é observado que a tecnologia apresenta relevância no dia-a-dia das pessoas. Elas estão presentes nas escolas, instituições e comunidade em geral como meio de auxiliar a aprendizagem de pessoas com necessidades específicas. Nesta concepção, a comunicação destas pessoas é crucial para desenvolver suas atividades cotidianas.

Em um sentido amplo, a tecnologia evolui para proporcionar uma vida mais fácil aos seus usuários. A Tecnologia Assistiva (TA) é voltada a auxiliar a independência, qualidade de vida e a inclusão de pessoas com necessidades específicas.

O termo *Assistive Technology*, traduzido para o português como Tecnologia Assistiva, foi criado em 1988 como importante elemento jurídico na legislação norte-americana conhecida como *Public Law 100-407* que foi renovado no mesmo ano como *Assistive Technology Act of 1998 (P.L. 105-394, S.2432)*. Este termo se encontra no *ADA - American with Disabilities Act* (Ato Americanos com Deficiências), que regula os direitos dos cidadãos com deficiência nos Estados Unidos da América.

Segundo o Ato (UNITED STATES, 1988), o termo tecnologia assistiva significa "tecnologia projetada para ser utilizada em um dispositivo de tecnologia assistiva ou serviço de tecnologia assistiva.". Tais dispositivos ou recursos são considerados qualquer item, peça ou sistema que é usado para aumentar, manter ou melhorar as capacidades funcionais de indivíduos com deficiência tais como bengala, cadeira de rodas, brinquedos e roupas adaptadas, computadores, softwares e hardwares especiais. Os serviços de tecnologias assistivas são serviços que auxiliam o uso dos dispositivos de tecnologia assistiva como avaliação da usabilidade de uma tecnologia assistiva para o indivíduo, serviços de compra, venda, manutenção e personalização do dispositivo de TA, assistência e treinamento para um indivíduo com deficiência, membros da família, empregadores entre outros que estão envolvidos na rotina das pessoas com deficiência.

2.1.1 Classificação de Tecnologias Assistivas:

A importância das classificações da tecnologia assistiva servirá para a classificação do trabalho no meio da acessibilidade. Ela também serve para identificar como ferramentas existentes estão fazendo este serviço.

O pedagogo professor facilitador em informática aplicada à educação pelo ProInfo do MEC Josué Geraldo Botura do Carmo (2005) fez uma possível classificação das TA. Entre elas, destacam-se três que são apresentadas nas seções seguintes.

2.1.1.1 Auxílios para a vida diária

Materiais e produtos para auxílio em tarefas rotineiras tais como comer, cozinhar, vestir-se, tomar banho e executar necessidades pessoais e manutenção da casa. A Figura 1 ilustra alguns destes materiais e sua utilização.

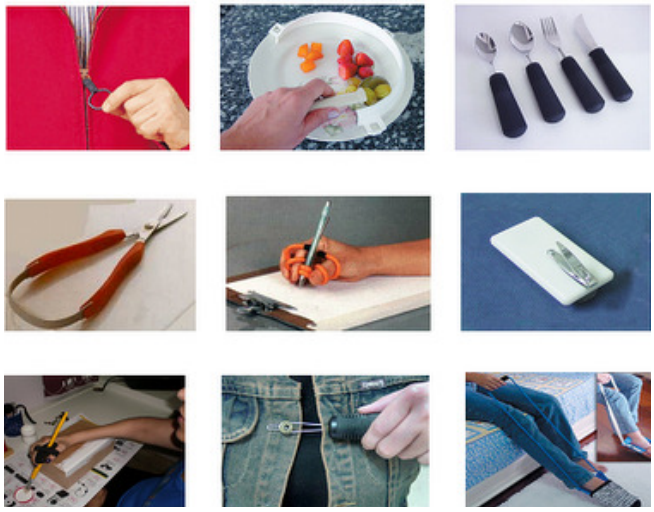

Figura 1 - Materiais de auxílio para vida diária

Fonte: HTTP://acessibilidadenapratica.blogspot.com.br/2011/02/terapia-ocupacional-acessibilidade.html

2.1.1.2 CAA (CSA) Comunicação aumentativa (suplementar) e alternativa

Recursos, eletrônicos ou não, que permitem a comunicação expressiva e receptiva das pessoas sem a fala ou com limitações da mesma. São muito utilizadas as pranchas de comunicação com os símbolos PCS (*Picture Communication Symbols*), apresentados na Figura 2, além de vocalizadores e softwares dedicados para este fim.

A seguir, a Figura 2 apresenta algumas possibilidades de combinação destes símbolos.

Esta combinação é chamada de prancha de comunicação aumentativa.

Figura 2 - Prancha de comunicação aumentativa

FONTE: HTTP://rededucacaosocial.blogspot.com.br/2011/06/o-que-e-um-sistema-de-simbolos-graficos.html

2.1.1.3 Recursos de acessibilidade ao computador

São equipamentos de entrada e saída (como ferramentas de síntese de voz, teclado Braille), auxílios alternativos de acesso (ponteiras de cabeça, de luz), teclados modificados ou alternativos, acionadores, softwares especiais (de reconhecimento de voz, etc.), que permitem às pessoas com deficiência usarem o computador.

Para exemplificar estes recursos, abaixo são apresentados dois tipos de teclados virtuais: na Figura 3 é utilizado no sistema operacional Mac e na Figura 4 no Windows XP pela ferramenta Microsoft Office Word.

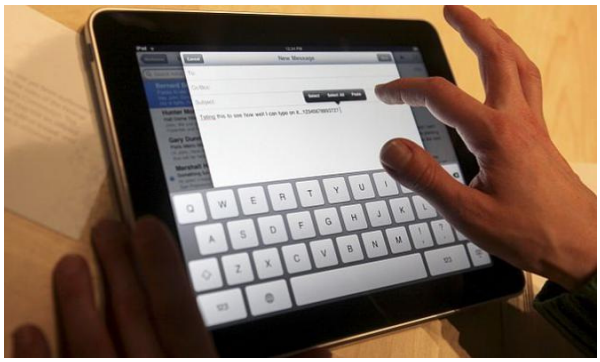

Figura 3 - Teclado virtual no sistema operacional Mac em um ambiente de navegação.

FONTE: http://ioshoy.com/teclado-negro-para-el-ipad-con-el-4-2.html

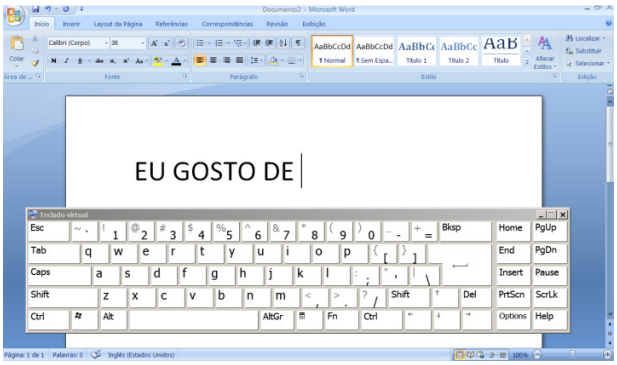

Figura 4 – Teclado virtual no Windows XP em um Microsoft Office Word respectivamente.

FONTE: http://miryampelosi.blogspot.com.br/2011/07/dicas-para-utilizar-o-teclado-virtual.html

2.2 SOCIEDADE SURDA e/ou MUDA

O termo "surdo-mudo" é uma generalização errônea. Pouquíssimos surdos são, de fato, mudos. O que acontece é que, com a capacidade auditiva desfavorecida, a pessoa pode não conseguir pronunciar a palavra por não reconhecer a forma audível da mesma.

Os resultados do Censo Demográfico 2010 apontaram 45.606.048 milhões de pessoas que declararam ter pelo menos uma das deficiências investigadas (deficiências visual, auditiva e motora), correspondendo a 23,9% da população brasileira. 5,1% possui deficiência auditiva, sendo 1,3% está na faixa de 0 a 14 anos e 4,2% de 15 a 64 anos (IBGE, 2010).

Estes dados sugerem que a parcela da população na idade da alfabetização (dados de 0 a 14 anos) pode apresentar dificuldade na comunicação pela presença de surdez. Esta dificuldade pode inferir em dificuldade de aprendizagem, relação social e na capacidade intelectual.

Na população com idade de 15 anos ou mais com deficiência há uma diferença de 8,9% quando se trata da alfabetização das mesmas em contraste com pessoas sem deficiência. Esta taxa poderia ser diminuída com o uso de uma comunicação eficiente entre as pessoas com necessidades específicas, aumentando o grau de entendimento e interesse pelo estudo.

Tais dados apresentados nesta seção refletem no nível de instrução, onde foi encontrada a diferença mais significativa: “*Enquanto 61,1% da população de 15 anos ou mais de idade com deficiência não tinha instrução ou possuía apenas o fundamental incompleto, esse percentual era de 38,2% para as pessoas de 15 anos ou mais que declararam não ter nenhuma das deficiências investigadas”* (CENSO, 2010). Tais dados estão representados na Figura a seguir.

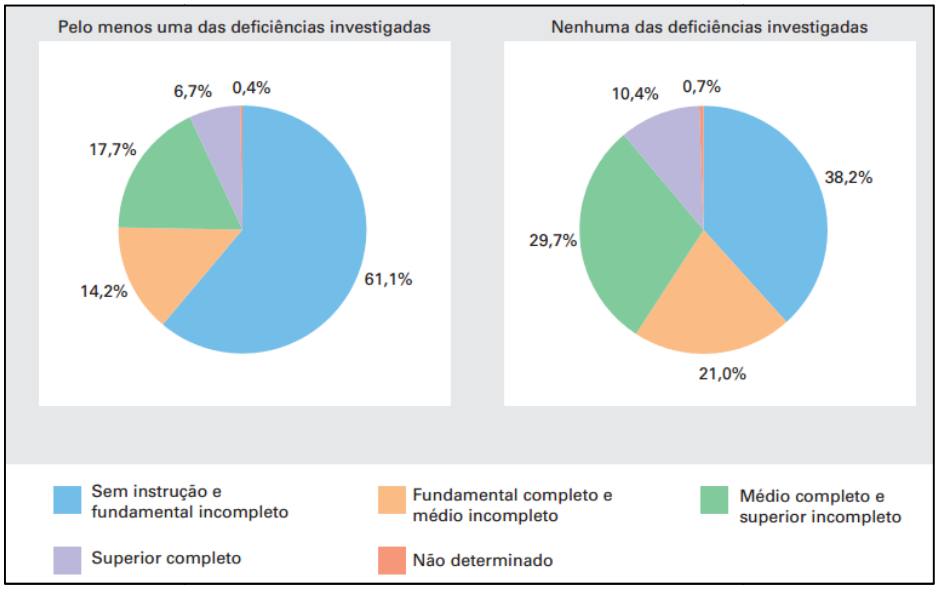

Figura 5 - Distribuição percentual da população de 15 anos ou mais de idade, por existência de pelo menos uma das deficiências investigadas e nível de instrução - Brasil – 2010

Fonte: IBGE, Censo Demográfico 2010

As necessidades específicas abordadas no trabalho, auditivas (surdez) e orais (mudez), são apresentadas a seguir.

2.2.1 Surdez

A audição, do mesmo modo como os outros sentidos, é importante para o desenvolvimento do sujeito, sendo uma privação sensorial. A partir dela é feita a comunicação com a sociedade e dela com o indivíduo, fazendo contato com o mundo.

Para pessoas nesta condição a comunicação é dificultada. Existe uma classificação da surdez que se caracteriza por ser parcial ou total, dependendo do nível de decibéis perdidos. Segundo a classificação do BIAP (*Bureau International d'Audiophonologic*) (BIAP, 2013), tem-se os grau de surdez Leve (perda de 21 e 40 dB), Média (41 e 70 dB), Severa (71 e 90 d B) e Profunda (90 dB). A partir daí é considerado Surdez de 1º Grau (90dB), 2º Grau (entre 90 e 100 dB) e 3º Grau (mais de 100dB) e então a perda total auditiva acima de 120 dB perdidos.

Conforme tais graus de perdas auditivas, (MEC, 2006) refere-se ao encontro na criança que pode ser em seus primeiros anos de vida ou tardiamente. Na surdez leve, a criança reconhece os sons da fala e desenvolve a linguagem oral e por isso é tardiamente descoberto e não é necessário o uso de ferramentas ou aparelho de amplificação. Na surdez moderada, a criança pode demorar um pouco para desenvolver a fala e linguagem, apresentando alterações articulatórias, pois não percebe todos os sons com clareza apresentando dificuldade no

aprendizado da leitura e escrita. Na surdez severa, ter-se-á dificuldades em adquirir a fala e linguagem sem ajuda terceira e poderá adquirir vocabulário do contexto familiar. Neste caso tem-se que usar aparelho de amplificação e acompanhamento especializado.

Quando a surdez é profunda, a criança provavelmente não desenvolverá linguagem oral, mas responderá a sons intensos como bomba, trovão, motor de carro e avião. Por estas características, a criança utiliza a leitura orofacial e é necessitado o uso de aparelho de amplificação ou implante coclear, bem como de acompanhamento especializado. Existe a possibilidade do indivíduo surdo não desenvolver a fala, caracterizando-se como indivíduo mudo. Este assunto será abordado na próxima seção.

2.2.2 Mudez

Além da mudez como consequência da surdez, ela pode ser física podendo estar relacionada, por exemplo, com a garganta, cordas vocais, pulmão ou boca ou pode ser causada por acidentes traumáticos podendo ser recuperada com o tempo.

A mudez devido a possuir perda auditiva é obtida por não ouvir a palavra falada. Muitas vezes o indivíduo possui a capacidade da fala, mas não a desenvolve. Devido a estas características, a comunicação de pessoas que possuem ambas necessidades específicas (auditiva e oral) é ainda mais dificultada. Apesar disto, não se deve considerar quem as possui como incapazes de se comunicar.

2.3 LIBRAS

A Libras é a língua de sinais oficial do Brasil, e é apenas uma das mais de 7.105 existentes no mundo (LEWIS, 2013) e é reconhecida como meio legal de comunicação e expressão no Brasil na lei n° 10.436, de 24 de abril de 2002 por Fernando Henrique Cardoso.

É a língua em que as pessoas nascidas ou que adquiriram recursos de comunicação (fala e/ou escuta) dificultados usam para se comunicar. É uma linguagem do campo visuo-espacial, como citado anteriormente. Para exemplificar, as figuras abaixo mostram como funciona a comunicação por sinais.

As figuras 6 e 7 mostram a sequência de movimentos que representam a palavra indicada. Para falar LIBRAS é importante a configuração e posição das mãos, ponto de articulação (lugar onde incide a mão predominante configurada) o movimento, a orientação e a expressão facial e/ou corporal. Como visto em (FELIPE, 1997), da combinação destes cinco parâmetros, tem-se o sinal com o qual é possível formar palavras e, com elas, frases em um contexto.

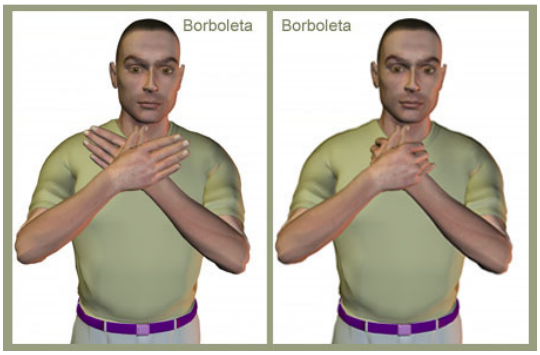

Figura 6 – Exemplo da palavra Borboleta exibida em Libras.
Fonte: www.dicionariolibras.com.br

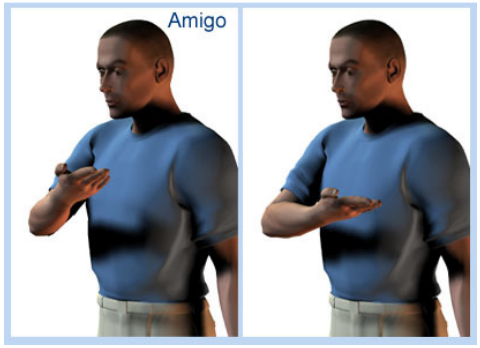

Figura 7 – Exemplo da palavra Amigo exibida em Libras.
Fonte: www.dicionariolibras.com.br

2.4 MINERAÇÃO DE TEXTO OU TEXT MINING

2.4.1 Descoberta de conhecimento em Banco de Dados (KDD) e Mineração de dados

O grande crescimento do volume de dados gera a necessidade da criação de um processamento para seleção de dados úteis e neles realizar padrões descoberta de conhecimento. Da necessidade de descoberta de conhecimento em grandes quantidades de dados foi criada a área de pesquisa chamada de *knowledge discovery in databases* (KDD), traduzido como descoberta do conhecimento em banco de dados.

Segundo Fayyad (1996), KDD é a área que compete o desenvolvimento de métodos e técnicas que dão sentido aos dados. Ele resume como sendo "*a aplicação de métodos de mineração de dados específicos para a descoberta de padrões e de extração*" (Fayyad,1996). A análise do grande volume de dados é demorada se não utilizadas técnicas de mineração de dados, uma etapa do processo de KDD.

A descoberta de conhecimento é composta por principalmente 3 etapas (TAN, 1999):

- o pré-processamento,

- mineração dos dados,

- pós-processamento.

Uma das principais aplicações de mineração de dados é a mineração de texto, o chamado *text mining*. Essa área tem sido alvo de muitas pesquisas devido à necessidade de desenvolvimento de algoritmos e técnicas de pré-processamento específicas para o tratamento de textos. Para Fayyad, *"linguagem natural apresenta oportunidades significativas para a mineração em texto de forma livre, especialmente para anotação automática e indexação antes da classificação de corpora de texto*" (FAYYAD, 1996), o que complementa a importância na proposta deste trabalho: realizar a tradução automática de textos na linguagem natural (português) para LIBRAS com a mineração de texto.

2.4.2 Mineração de texto

A Mineração de texto, ou *text mining*, é uma parte da área de Mineração de Dados. É quando os dados esperados para a mineração são textos, feita em cima de caracteres e/ou strings, e não dados numéricos, quando a mineração é feita em cima de números.

Segundo Tan (1999), mineração de texto pode ser dividida em duas fases: refino de texto que transforma documentos de texto de forma livre em uma forma intermediária escolhida, e aplicação do conhecimento que deduz padrões ou conhecimento da forma intermediária. Para melhor visualização, estas fases são ilustradas na Figura 8. (TAN,1999).

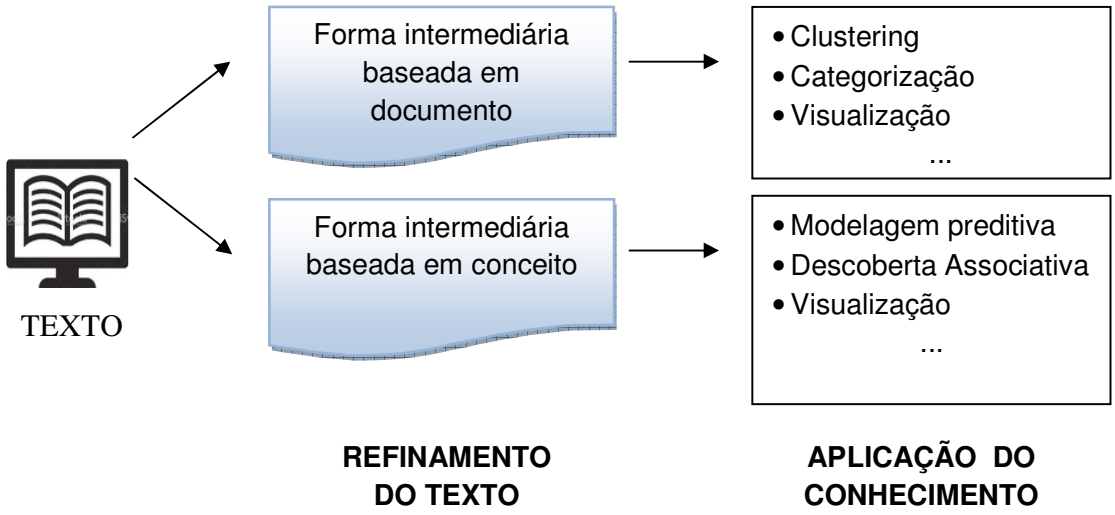

Figura 8 - Identificação das fases de refinamento do texto e aplicação do conhecimento na mineração do texto

As formas intermediárias podem estar semi estruturadas ou estruturadas. A partir destas formas pode ser chegado aos processos de *clustering* (encontro de padrões), categorização, modelagem preditiva, descoberta associativa, visualização entre outros.

2.4.2 Mineração de texto relacionada a recuperação de texto

Com o advento da World Wibe Web (Web), o volume de informações na nuvem sofreu um salto. Este grande volume de informações estão em documentos de texto e arquivos multimídia armazenadas em sua maioria em estruturas HTML e XML. A recuperação de informações (RI) da Web lida com problemas de armazenamento, indexação e recuperação (busca) pelo alto número de páginas web, documentos desestruturados, por não ser limitada a uma linguagem e por comportar pesquisa de forma livre com palavras-chaves segundo (ELMASRI, NAVATHE, 2011).

A Tabela 1 faz uma comparação das principais diferenças entre banco de dados e sistemas de recuperação de informação.

Tabela 1 - Comparação entre banco de dados e sistemas de RI

Banco de Dados	**Sistemas de Recuperação de Dados**
Dados estruturados	Dados desestruturados
Controlados por esquema	Sem esquema fixo
Modelo relacional	Modelo de consulta em forma livre

Uma das características relevantes nesta comparação é a presença ou não de uma estrutura em que os dados são apresentados. Na recuperação de dados, os dados são desestruturados. Ou seja, são informações que não tem um modelo formal pré-definido, baseiam-se no conhecimento da linguagem natural. É com este tipo de dado que o trabalho vai tratar.

Para a interação em sistemas RI, há dois modos: recuperação e navegação. A recuperação é a extração de informações relevantes de um repositório por meio de uma consulta. Já a navegação é a atividade do usuário propriamente dita, o que é relevante para o usuário. Combinando estas interações temos uma busca na web (ELMASRI, NAVATHE, 2011).

Para ilustrar em forma resumida os passos da recuperação de informação, é apresentada a Figura 9 com a estrutura geral da RI.

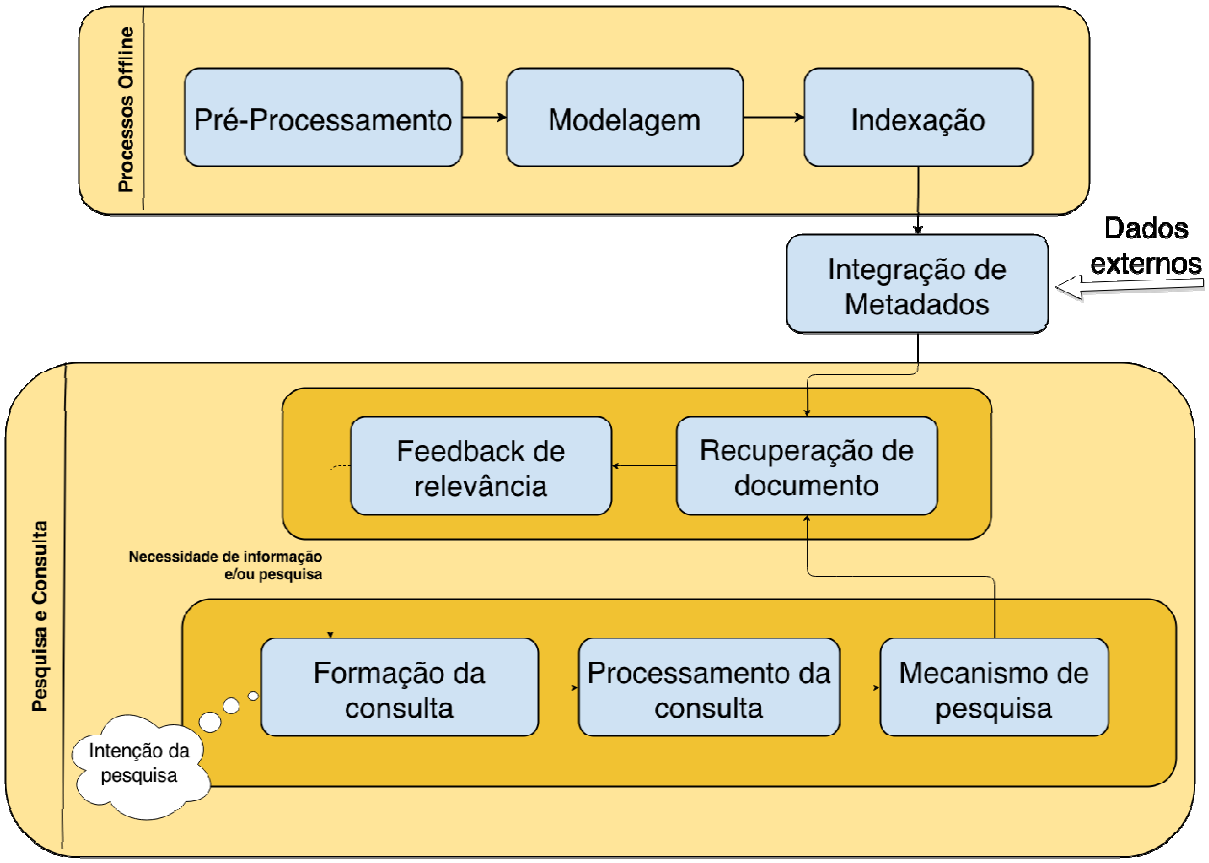

Figura 9 - Estrutura geral de um sistema RI. Adaptada de Elmasri e Navathe (2011)

Na Figura 9 os módulos representam estágios de um sistema de recuperação de informação. As três primeiras etapas representam a preparação para a recuperação e são chamadas de processos off-line. A etapa de integração de metadados representa os dados externos (informações adicionais) deixando ao módulo de recuperação de documento agregar os resultados da indexação e metadados. Com a reunião de todas as informações que o sistema conseguir gerar, faz-se um feedback de relevância onde o resultado pode ser personalizado pelo usuário final e pode ser analisado o padrão dos resultados relevantes.

Com os dados representados e analisados, decide-se recorrer a uma nova pesquisa ou não. Este requerimento é feito aos três módulos referentes a pesquisa e consulta: formação da consulta (a partir de um método adotado pelo programador:palavra-chave, frase entre outras), processamento da consulta (conversão da consulta para formato interno de pesquisa, avaliação e expansão da consulta) e mecanismos de consulta (escolha da estratégia de pesquisa e medida de semelhança).

As duas abordagens para sistemas de RI são a estatística e a semântica. Elas são apresentadas na Tabela 2, fazendo uma comparação entre elas.

Tabela 2 - Comparação entre as abordagens estatística e semântica de um sistema de RI.

Abordagem Estatística	Abordagem Semântica
Análise em trechos	Análise do todo
Palavras são contadas, pesadas, medidas por relevância ou importância	Análise feita a partir dos níveis sintático, léxico e sentencial
Comparação com termos da consulta através de um grau de combinação	Baseada em conhecimento
Usa técnica booleana, espaço de vetor e probabilística	Usa técnicas a partir da semântica

Para o trabalho, adota-se a abordagem semântica. Adotar esta abordagem não significa desconsiderar a abordagem estatística, já que podem e ser usadas em conjunto para melhorar o processo de recuperação. Entretanto, o trabalho não necessita de consecutivas recuperações de informação, não recorrendo por hora à abordagem estatística.

3. TRABALHOS RELACIONADOS

Para verificação da relevância do trabalho foi realizada uma pesquisa sobre trabalhos relacionados ao escopo da proposta do projeto SIGNUM. Durante essa pesquisa foram encontradas diversas ferramentas. No contexto deste trabalho são descritas a seguir as três ferramentas estudadas e uma dissertação de mestrado.

3.1 PoliLibras

O PoliLibras, mais do que uma ferramenta, é um projeto de desenvolvimento de ferramentas para a comunidade surda. Segundo a página oficial do projeto (POLILIBRAS, 2014), a intenção de trabalho futuro era o desenvolvimento de um plug-in para que administradores de páginas pudessem adicionar aos seus sites e a saída que o projeto propõe é uma sequência de animação gráfica (JANUÁRIO, LEITE, KOGA, 2010).

O código é aberto, desenvolvido na linguagem Java. Para a sua execução é necessário possuir plataforma Java instalada, sistema operacional Linux e conexão com a Internet. Na documentação é possível seguir o passo a passo de execução. Para isso é necessário abrir uma janela de terminal e executar o programa. Após executado e em foco da janela do programa deve-se digitar ‘t’, voltar a janela de terminal e digitar a frase que deseja ser traduzida. Em seguida, a orientação é voltar à janela do programa e pressionar a barra de espaço para que a tradução seja mostrada sinal por sinal. As palavras que não constam no dicionário são mostradas com um símbolo que corresponde “palavra não encontrada”, o que é chamado de “coringa”. Este suporta apenas frases no período simples e em ordem direta. Quando a estrutura diferir desta, é mostrada em português sinalizado, ou seja, palavra por palavra.

A animação que traduz o texto introduzido é composta por uma cabeça e duas mãos soltas em um espaço demarcado por coordenadas que podem ser mudadas com clique e arraste de mouse. Na execução do programa, colocada uma frase simples “eu vou para casa” o tradutor não conseguiu traduzir. Traduzindo como “EU CORINGA” e a partir desta tentativa todas outras simples ou apenas com uma palavra foram traduzidas como “CORINGA”, mostrando que o sistema não está funcional. A execução mostrou também que o representante tradutor não apresenta expressão facial, apenas movimentação das mãos.

3.2 Hand Talk

O Hand Talk é um tradutor que abrange várias plataformas: aplicativo móvel nas plataformas android e iOS, plug-in para sites e pode ser contratado totens para eventos. O tradutor também conta com uma representação 3D como saída do tradutor.

Nos testes feitos com o aplicativo HandTalk, ele apresentou-se instável em algumas

plataformas. O teste feito em um *smartphone* com a plataforma android 4.0.5 ao clicar para traduzir ocorria uma falha, voltando a tela de apresentação. Já na plataforma iOS, a partir de um iPad, ele apresentou todas as funcionalidades propostas. O tradutor para páginas web é disponibilizado na página oficial do tradutor (HANDTALK, 2014) e está funcional, mas para testes em outras páginas web o serviço deve ser contratado.

No ambiente web, ele é um serviço contratado e está disponível gratuitamente apenas na sua página. Para os testes feitos na página e no aplicativo iOS, a entrada tem tamanho máximo de 140 caracteres.

3.3 PULØ:

O trabalho é um gerador de interlíngua Português-LIST para LIBRAS, sendo LIST referência a *Libras Script for Translation*. Ele traduz textos em português para LIBRAS de forma semi automática, passando pela revisão e intervenção humana na tradução e correção do texto de entrada.

Como entrada, o sistema PULØ espera a forma mais simplificada do texto português, chamado de português normalizado pelos autores em (MARTINS, 2005). Segundo eles, "*esta entrada é desprovida de elipses, topicalizações, anacolutos, anáforas, ambiguidades léxicas e sintáticas e outros acidentes lógico-gramaticais que pudessem vir a afetar o desempenho da ferramenta*". Foi feito um sistema para gerar a interlíngua LIST utilizando a UNL (*Universal Networking Language*), a partir deste sistema tem-se a tradução direta ou indireta.

A UNL é uma linguagem intermediária entre a língua natural e a do computador e é propriedade da Organização das Nações Unidas (ONU). A resposta dele é passada para um decodificador: o DeCo. Este esquema é proposto pois, segundo os autores, facilitará a tradução a partir de outras línguas pois a LIST foi desenvolvida para ser independente da língua-fonte ou língua-alvo.

O resultado para 12 frases de um diálogo em uma história em quadrinho foi dito satisfatório em (MARTINS, 2005). Entretanto, não foram apresentados resultados específicos e foi dito que "*uma vez que em um corpus maior de sentenças em português deverão surgir outros problemas a serem atacados*". Ou seja, para frases simples houve um resultado válido, mas que se aplicado em frases complexas não apresentará o mesmo resultado.

3.4 Dissertação de mestrado "PRODUÇÃO DE TEXTOS PARALELOS EM LÍNGUA PORTUGUESA E UMA INTERLÍNGUA DE LIBRAS"

A dissertação de mestrado trata da produção de textos paralelos na língua portuguesa e uma interlíngua LIBRAS. Este trabalho tem ênfase criando a arquitetura de um sistema de apoio à anotação de simplificação de texto – parte do projeto PorSimples (SANTOS, 2009).

O projeto PorSimples é um projeto vinculado ao Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo na cidade de São Carlos e tem objetivos da Simplificação Textual do Português para inclusão e Acessibilidade Digital (PEREIRA, 2008). Nele foi desenvolvido um módulo de reescrita em LIBRAS. O módulo realiza operações de substituição dos termos para termos LIBRAS (da interlíngua).

3.5 Quadro comparativo

A seguir é apresentado um quadro comparativo em forma de tabela que compara características chaves das ferramentas analisadas nesta seção e a proposta do projeto SIGNUM. Na tabela "v" significa que a ferramenta possui e/ou desempenha a característica, "o" desempenha em parte e "x" não possui e/ou não desempenha a característica.

Tabela 3 - Comparação entre as ferramentas analisadas e a proposta do projeto SIGNUM

Característica/Ferramenta	PoliLibras	HandTalk	PULØ	Dissertação	SIGNUM
Não é pago	✓	○	✓	✓	✓
Disponível na plataforma web	x	✓	x	x	✓
Disponível em plataformas mobile (android e iOS)	x	✓	x	x	x
Aceita frases complexas (tamanho e estrutura gramatical)	✓	✓	x	✓	✓
Tradução completa (substituição e ordenação dos termos)	x	✓	✓	○	✓

4. MATERIAIS E MÉTODOS

Esse capítulo tem por objetivo descrever os principais materiais e métodos utilizados no desenvolvimento deste trabalho.

4.1 MINERAÇÃO DE TEXTO E FERRAMENTAS UTILIZADAS

A mineração de texto oferece técnicas que podem facilitar o desenvolvimento deste trabalho. As propriedades e técnicas utilizadas são definidas a seguir.

No trabalho são aplicadas técnicas de recuperação de texto, já explicadas anteriormente. No contexto deste trabalho: é através da busca na web que se têm os dados para começar a tradução Português-LIBRAS. O usuário que navega na página web seleciona o texto que deseja traduzir, o sistema faz a recuperação desta seleção e tem-se o material o qual será feita a tradução.

Na Figura 8 apresentada na seção 2.4.2 podemos situar o trabalho na forma intermediária baseando-se no documento, na etapa de categorização. A categorização é utilizada no trabalho para a "etiquetação" de cada termo ou palavra de entrada em sua forma gramatical. Também pode ser enquadrado na forma intermediária baseada em conceito. Ou seja, após a categorização é feita a descoberta associativa, característica desta forma, para produzirmos a interlíngua. É nesta etapa que se define a melhor combinação de cada termo e seu lugar na frase produzida com os termos convertidos para a interlíngua.

Para realizar a categorização, foram pesquisadas ferramentas que podem auxiliar e/ou complementar o trabalho, facilitando e possibilitando a mineração do texto. Algumas ferramentas foram estudadas mas não utilizadas no escopo deste trabalho, porém são importantes para os trabalhos futuros.

4.1.1 CoGrOO

É uma extensão *open source* do LibreOffice (pacote de ferramentas de escritório para contendo editor de texto, planilhas, apresentações entre outros (LIBREOFFICE, 2014)). Tem várias contribuições, inclusive como extensão do Firefox (em Java).

O CoGrOO é um corretor gramatical (e não ortográfico). Ou seja, não corrige palavras escritas de forma errônea, mas sim sentenças que não possuem forma gramatical correta. Exemplos de tipos de erros gramaticais que o CoGrOO corrige: colocação pronominal, concordância nominal, concordância entre sujeito e verbo, concordância verbal, uso de crase, regência nominal, regência verbal e erros comuns da língua portuguesa escrita (COGROO,2013). A ferramenta disponibiliza junto com o código um dicionário da classificação das palavras que ele abrange.

4.1.2 Simplifica

A ferramenta Simplifica faz parte do projeto "PorTexto, simplificando o português" (SIMPLIFICA, 2014) e foi desenvolvida em Ruby on Rails. Consiste em uma página web onde pode ser inserido o texto. Após ser escolhido o tipo de simplificação (léxica ou sintática) o texto é simplificado. É uma ferramenta muito utilizada para pessoas com dificuldade de entendimento de textos complexos. Segundo a Agência USP de Notícias (2013), a ferramenta auxilia o acesso à informação de pessoas com alfabetização deficiente ou com problemas de cognição.

Essa ferramenta também permite que o usuário escolha dicionários utilizados. Para a simplificação léxica são disponibilizados os dicionários de Sinônimos (Papel www.linguateca.pt/PAPEL, e Tep 2.0 - www.nilc.icmc.usp.br/tep2) e Dicionários de Palavras Simples (Dicionário do PorSimples e Dicionário Televisão).

Para a simplificação sintática têm-se as opções de nível de simplificação:

- Simplificação natural: para pessoas que cursaram o quinto ano do ensino fundamental;
- Simplificação forte: para pessoas que cursaram entre o segundo e o quarto ano do ensino fundamental ou com distúrbios cognitivos;
- Simplificação personalizada onde pode-se escolher o que será analisado entre aposto, voz passiva, orações coordenadas, orações subordinadas, cláusulas relativas e adjuntos adverbiais e cláusulas reduzidas.

Atualmente essa ferramenta está em um processo de migração de servidor e por esse motivo não está funcional online.

4.2 TAGGER:

4.2.1 Definição

Taggers são ferramentas que podem ser utilizadas para etiquetação de palavras, sendo utilizadas para acrescentar informação a ela. São muito utilizadas no Processamento de Língua Natural (PLN). No trabalho, será utilizado para apresentação da classificação de cada palavra. As seções procedidas apresentam ferramentas que desempenham este papel.

4.2.2.1 MxPost

É um etiquetador morfossintático. Atribui a classe de palavra a cada uma das unidades de um texto/corpus. Inicialmente foi produzido para a língua inglesa para resolver ambiguidades da linguagem natural com a precisão usando uma única técnica de modelagem estatística com base

no princípio da entropia máxima (RATNAPARKHI, 1998). Esta entropia máxima significa o máximo de informação possível a ser processada.

O objetivo do projeto era resolver problemas de detecção limite frase, etiquetação (*part-of-speech tags*), ligação entre frases, análise da linguagem natural e categorização de texto. O objetivo foi alcançado com sucesso utilizando a técnica de entropia máxima e modelos probabilísticos e em 2000 o trabalho foi feito para o PLN portuguesa, possuindo uma precisão de 96,98% de precisão (FINATTO, 2011).

4.2.2.2 POSTagger

O POSTagger, ou na versão portuguesa chamado de Anotador Categorial, é uma das ferramentas constituintes do LX-Center (LX-CENTER, 2014). Foi desenvolido na Universidade de Lisboa pelo NLX Group (*Natural Language and Speech Group*), possui versão on-line e off-line e como resultado apresenta após cada palavra o seu conjunto de *tags*/etiquetas. No Anexo 1 deste trabalho são apresentadas as tabelas com as etiquetas reconhecidas e produzidas por este tagger.

Para os testes feitos com esta ferramenta, ela apresentou uma cobertura total (acrescentando etiquetas a todas as palavras). A precisão da ferramenta foi pesquisada na própria documentação e artigos dos autores, como em (BRANCO, 2004), onde foi testado e apresentou a precisão de 96,87% de acerto para esta ferramenta.

4.2.3 Quadro Comparativo

A Tabela 4 é uma comparação entre as duas ferramentas de tagger apresentadas anteriormente. Nela, as características analisadas são: alta precisão de acertos, se possui versão online e se possui versão off-line.

Tabela 4 - Comparação entre as ferramentas de *tagger* estudadas

Característica/Ferramenta	**MXPost**	**POSTagger**
Alta precisão de acerto	✓	✓
Possui versão off-line	✓	✓
Possui versão on-line	x	✓

4.3 PARSER:

O termo inglês, *parser*, significa analisador. Em PLN a função do *parser* é analisar a estrutura gramatical e classificar a relação entre as palavras ao longo de uma oração. A seguir são apresentadas as ferramentas estudadas que desempenham esta função.

4.3.1 PALAVRAS

O anotador morfossintático PALAVRAS é também um parser automático, ou seja, é caracterizado por ser uma ferramenta de tagger e parser. Seu processamento semântico realiza a categorização semântica com o significado de cada item lexical feita por traços semânticos e não por definições de dicionários (TOMAZELA, 2010).

Esta ferramenta começou a ser desenvolvida por Eckhard Bick como projeto de doutorado na Universidade de Århus, Dinamarca em 2000. Conta com regras gramaticais para a análise morfológica e sintática de qualquer texto (BICK, 2000). Após foi incorporado ao projeto VISL (*Visual Interactive Language Learning*) que atualmente atua em 28 línguas (VISL, 2014) e é uma ferramenta com licença paga.

Nesta análise de Tomazela (2010), é dito que para o modelo de classificação, "são considerados somente os substantivos, entidades nomeadas e alguns adjetivos, para os quais é possível atribuir um valor semântico" (TOMAZELA, 2010). Este modelo é adotado para identificar similaridade, ou traços semânticos entre os termos.

Quando apresentado um texto com termos ambíguos, o sistema é capaz de verificar as palavras ambíguas e a partir delas verificar qual o sentido nela na frase. Assim, é classificada de forma correta.

Bick diz em sua dissertação que "*as taxas de acerto com exatidão sobre texto livre são de mais de 99% para a morfologia/PoS e 96-97% para a sintaxe (...) e para sistemas probabilísticos, pairam em torno da marca de 97% para correção PoS-tagging*" (BICK, 2000, p. 438). Por ter tais índices de acerto tão altos é considerado uma das melhores ferramentas para estas funções.

4.3.2 LXParser

O LX-Parser é um parser estatístico para sentenças do português. Faz parte do conjunto de ferramentas do LX-Center, desenvolvido pela Universidade de Lisboa pelo NLX Group.

Possui a versão online e off-line para download que requer a máquina virtual Java 5 instalada no ambiente de execução. Na versão off-line, o texto de entrada deve ter sido "tokenizado" pelo LX-Tokenizer que adiciona referências lexicais necessárias a cada termo para o funcionamento do *parser* (LX-CENTER, 2013).

4.3.3 Quadro Comparativo

As características analisadas para a comparação entre as ferramentas de *parser* são a verificação de ambiguidade, se possui versão online, se possui versão off-line e se a ferramenta é livre ou paga. Esta comparação é apresentada na Tabela 5.

Tabela 5 - Comparação entre as ferramentas de *parser* estudadas

Característica/Ferramenta	**PALAVRAS**	**LXParser**
Verifica palavras ambíguas	✓	o
Possui versão off-line	✓	✓
Possui versão on-line	✓	✓
Ferramenta livre	x	✓

4.4 (PARSER & TAGGER) LX-Center

O LX-Center, que já foi citado anteriormente, é um centro de serviços linguísticos web. Ele possui uma gama de ferramentas de processamento da linguagem natural, como divisão da sentença, tokenização, lematização nominal, análise morfológica nominal, conjugação verbal, POS-tagging, entre outras (FINATTO, 2011). Estas funcionalidades são fornecidas por um ou mais serviços do LX, como dito em (BRANCO, 2009) e são encontrados na página web da ferramenta (LX-CENTER, 2013) para serem utilizados gratuitamente na versão on-line e off-line, disponível para *download*.

Por ser um centro de serviços disponibilizados on-line, sem custos e possuir uma boa taxa de acertos segundo a pesquisa, as ferramentas a serem utilizadas neste trabalho pertencem ao LX-Center. Além disso, existem outras ferramentas deste centro que poderão ser usadas futuramente e facilitam na utilização e comunicação entre as ferramentas se todas possuírem o mesmo formato de resposta.

4.5 LINGUAGEM DE PROGRAMAÇÃO PHP

O PHP é uma linguagem de script muito utilizada no desenvolvimento de aplicações web. O código PHP é executado no servidor, que executa a rotina criada e retorna a resposta ao cliente (PHP,2013).

É uma linguagem open source que disponibiliza uma gama de funções e estruturas a serem utilizadas. Neste trabalho, a linguagem é relevante para o fácil tratamento de *strings* (texto) e lidar com leitura de páginas web com marcações HTML (linguagem utilizada para

desenvolvimento *web*). Segundo a página oficial, onde se encontra a documentação, “o PHP é extremamente útil em recursos de processamento de texto, do POSIX Estendido ou expressões regulares Perl até como interpretador para documentos XML”.

Ao utilizar as ferramentas online do LX-Center, a resposta encontra-se na página e para isso deve ser feito uma varredura procurando a resposta desejada. Esta procura torna-se muito mais fácil se utilizada uma linguagem que reconhece as marcações de linguagens Web. É o caso do PHP.

Por ser uma linguagem web, ele garante uma vasta amplitude de funcionamento em sistemas operacionais e servidores web (PHP,2013). Isto garante uma liberdade de escolha do ambiente de trabalho, permitindo também a utilização de programação estrutural ou orientada a objeto (POO), introduzido por completo na versão PHP 5.

5. RESULTADOS: PROPOSTA DE METODOLOGIA PARA A GERAÇÃO DE INTERLINGUA PORTUGUES-LIBRAS

A partir da identificação do problema da leitura de textos português por pessoas falantes LIBRAS, foi proposto o projeto SIGNUM: um tradutor LIBRAS para executar em navegadores web. No contexto do projeto SIGNUM, este trabalho tem por objetivo a geração da interlíngua português-LIBRAS onde cada termo simboliza um gesto LIBRAS a ser exibido no formato de um vídeo no navegador Web.

A partir deste objetivo foi pesquisado o que há desenvolvido para a comunidade surda brasileira e apresentado nas seções de trabalhos relacionados anteriormente. Estes trabalhos afirmaram a relevância do projeto SIGNUM desenvolvido e, para alcançar o objetivo proposto, foram estudadas ferramentas de mineração de texto que poderiam ser utilizadas no trabalho e optou-se pela utilização da linguagem de programação web PHP para a integração das ferramentas estudadas e desenvolvimento dos algoritmos necessários a todo o processo de *text mining*.

Com estas etapas concluídas discutiu-se uma arquitetura a ser adotada para o trabalho. A arquitetura, a implementação e como o trabalho foi desenvolvido são apresentados nas seções a seguir.

5.1 ARQUITETURA

Para a realização do projeto SIGNUM foi construída uma arquitetura de comunicação entre os serviços de: seleção de texto pelo usuário, processamento deste texto e visualização da tradução pelo link da interlíngua com os vídeos que a representa. Esta arquitetura é apresentada na Figura 10.

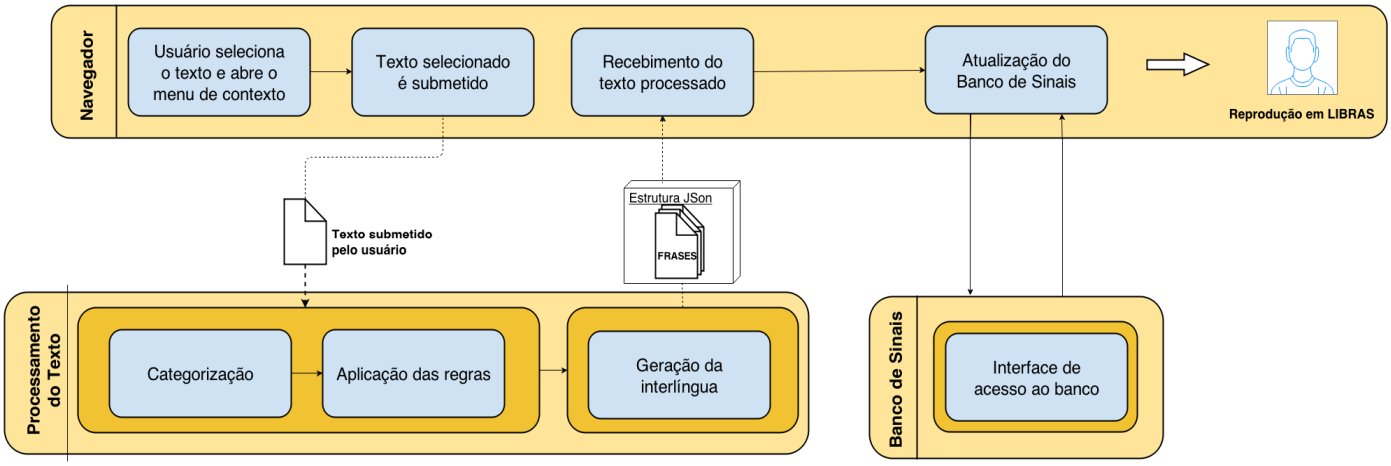

Figura 10 - Arquitetura Geral do Projeto SIGNUM

Módulo "Usuário seleciona o texto e abre o menu de contexto": O funcionamento da ferramenta parte da seleção de um texto em uma página da internet pelo usuário. Clicando com o botão direito do mouse sobre ele tem-se a opção da tradução do mesmo pelo SIGNUM.

Módulo "Texto selecionado é submetido": Neste módulo o texto selecionado é enviado para o bloco de processamento do texto, módulo onde este trabalho se encontra. Após o processamento, é devolvido para o módulo navegador a interlíngua gerada. Cada termo desta interlíngua está referenciado em um banco de dados com sua devida representação visual (gesto em LIBRAS). Depois de encontrada esta referência, é mostrado o vídeo representando o texto selecionado em uma janela pequena aberta no navegador.

Módulo "Recebimento do texto processado": O texto processado é devolvido para o módulo das etapas feitas no navegador.

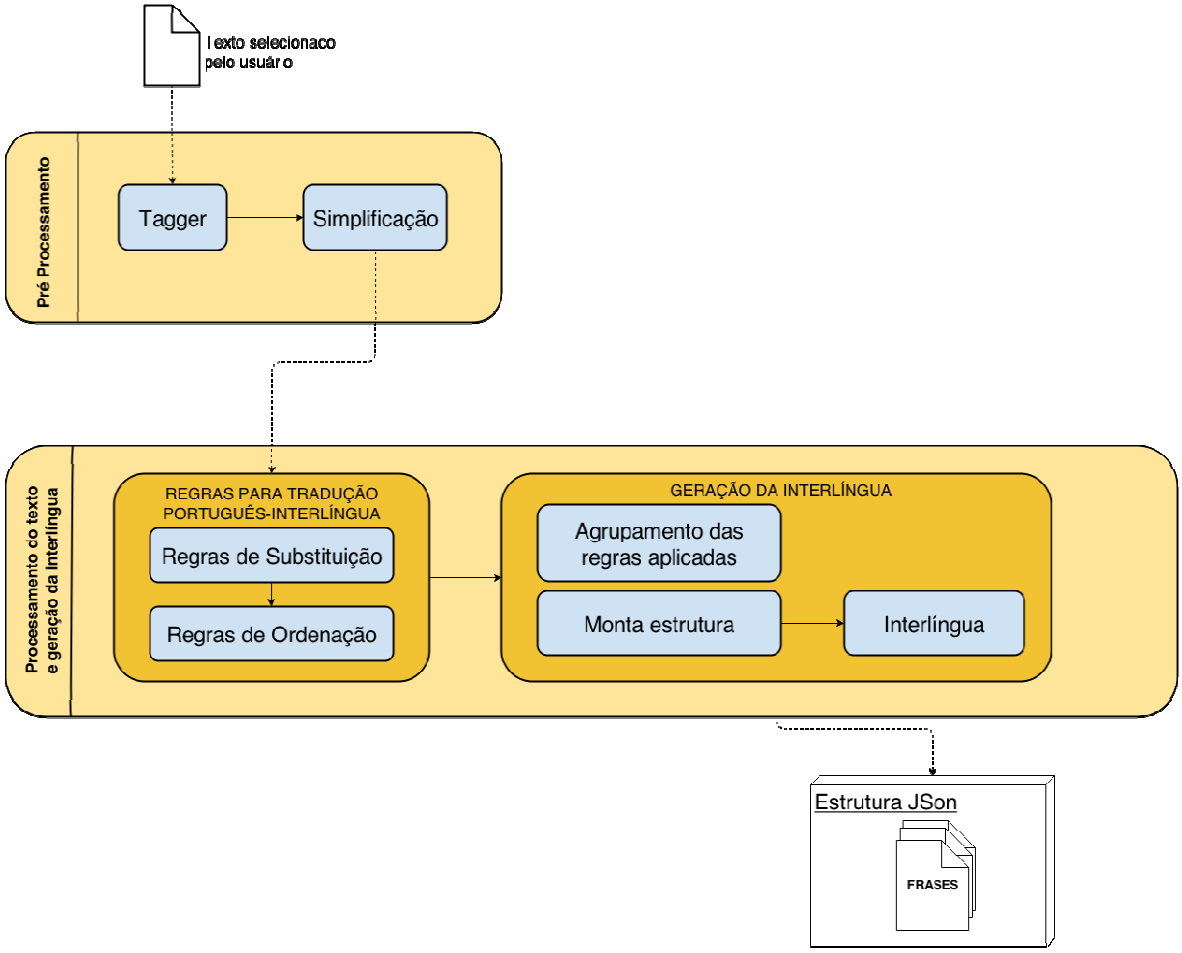

Figura 11 - Detalhamento do módulo de processamento de texto

Módulo "Atualização do Banco de Sinais": Cada termo representa um vídeo e neste módulo eles são ligados. Se o termo não existe no banco, ele é atualizado. A atualização é feita com o campo da palavra, mas sem o vídeo. Este pode ser incorporado ao banco futuramente.

Módulo "Reprodução em LIBRAS": Os gestos em LIBRAS do texto selecionado são apresentados ao usuário na forma de um vídeo único em uma janela do navegador.

Como dito anteriormente, este trabalho é o desenvolvimento do módulo de processamento do texto. Este módulo é detalhado em etapas mais específicas conforme mostra a Figura 11.

Feita a recuperação do texto, ele é encaminhado para um pré-processamento. Os módulos deste bloco são:

Módulo "Tagger": É desempenhado pela ferramenta online LX-Tagger, que categoriza cada palavra do texto morfologicamente e sintaticamente. É a partir destas *tags* que é realizada a maior parte do processamento.

Módulo "Simplificação": a esta etapa compete a separação do texto em frases e após a separação dos termos de cada frase para uma estrutura que possa ser melhor visualizada e percorrida no processamento do texto. Como trabalho futuro, será utilizada a ferramenta Simplifica (descrita na seção 4.1.2) para a divisão de um texto complexo em frases simples. O simplifica será incorporado à arquitetura assim que a ferramenta estiver disponível novamente.

Após o pré-processamento do texto, a estrutura do mesmo está pronta para as próximas etapas de geração da interlíngua Português-Libras. O bloco de processamento do texto divide-se em duas partes: aplicação das regras e geração da interlíngua.

Módulo "Regras para tradução português-interlíngua": As regras de tradução são aplicadas neste módulo. Para a tradução, são aplicados dois tipos de regras: substituição, onde o termo em português é representado por outro termo em LIBRAS e ordenação, onde os termos são ordenados no formato reconhecido em LIBRAS.

Módulo "Geração da interlíngua": Após a aplicação das regras, é preciso de ajustes finais onde são aplicadas as denominadas regras gerais. Neste módulo é feito um refinamento e os termos prontos são colocados na estrutura para a entrega da interlíngua ao módulo feito no navegador.

As regras aplicadas para a geração da interlíngua são descritas a seguir na seção 5.2 deste trabalho e como foi desenvolvida a implementação na seção 5.3.

5.2 DEFINIÇÃO DE REGRAS DE TRADUÇÃO PORTUGUÊS- LIBRAS

Um projeto interdisciplinar requer a dependência de especialistas das áreas relacionados ao objetivo do trabalho para a realização eficiente deste. Como não há um registro das regras de tradução português-LIBRAS, foi necessário organizá-las a partir de estudos na literatura e conversas com especialistas na área. Elas foram revisadas por uma tradutora e intérprete de LIBRAS.

As regras de tradução português-LIBRAS são difíceis de serem identificadas por LIBRAS acontecer no espaço. Por esta razão, a descoberta e listagem destas regras é a parte primordial do trabalho.

Primeiramente foi dividido em três tipos de regras: regras de substituição, regras de ordenação e regras gerais. As regras de substituição estão relacionadas à palavra em si e como ela é descrita em LIBRAS. Já a regra de ordenação está ligada a sintaxe, ou seja, as *tags*. Com esta etapa é possível reordenar estes elementos conforme regras implementadas. E regras gerais são aplicadas para os ajustes finos na frase para compor a interlíngua.

5.2.1 Regras de substituição

Estas regras identificam as palavras que possuem outra representação em LIBRAS ou que não existem e podem ser substituídas para que o sentido da palavra permaneça. São exemplos destas as chamadas palavras compostas.

As palavras compostas em LIBRAS ocorrem quando a palavra na língua portuguesa é representada pela combinação de mais palavras em LIBRAS (FELIPE, 1997). A tabela 6 apresenta as palavras compostas identificadas até então para o trabalho. A ligação das palavras é expressa por "^".

Tabela 6 - Palavras compostas LIBRAS

Português	**Interlíngua LIBRAS**
Zebra	CAVALO^LISTRAS
Açougueiro	HOMEM^VENDER^CARNE
Faqueiro	CAIXA^GUARDAR^COLHER^FACA^GARFO
calmante	PÍLULA^CALMA
Analgésico	PÍLULA^DOR DE CABEÇA
Ginecologista	MÉDIC@^SEXO
Oftalmologista	MÉDIC@^OLHO
Pediatra	MÉDIC@^CRIANÇA
Cardiologista	MÉDIC@^CORAÇÃO
Frutas	MAÇÃ^VARIADOS
Colorido	COR^VARIADOS
Alimentos	COMER^VARIADOS
Animais	LEÃO^VARIADOS
Escola	CASA^ESTUDAR

Em LIBRAS, a representação de número e gênero para adjetivos não é necessária pois o adjetivo vem junto com o que é adjetivado. Nesses casos, a palavra a ser adjetivada carrega estas

informações. Então, a representação de adjetivos é feita com a substituição do caractere de representação do gênero e número “o”, “a”, “os”, “as” pelo caractere “@” (FELIPE, 1997).

Temos como exemplo o adjetivo bonito. Quando aparece BONIT@, pode significar: bonito, bonita, bonitos ou bonitas. Se é necessária uma caracterização de gênero para o adjetivo, é possível acrescentar a palavra HOMEM ou MULHER após o adjetivo. Entretanto, como dito anteriormente, não é necessário.

Estas palavras que possuem gênero e número normalmente só são expressas quando são substantivos (nomes). Assim, quando forem nomes podem ser expressas com as palavras HOMEM ou MULHER para gênero e com o sinal de MUITO ou DIVERSOS para quantidade. Um exemplo é o termo “árvores” que pode ser expresso como MUITO ÁRVORE.

Quando a palavra expressa um sentimento, é difícil expressar em LIBRAS e por isso algumas são substituídas por palavras que possam expressá-la. É o caso de ansiedade que é expresso com a palavra depressa.

5.2.2 Regras de ordenação

As regras de ordenação foram as mais difíceis de serem identificadas pois em LIBRAS as frases podem ser ditas de várias formas, assim como em qualquer outra língua. Este foi um dos maiores desafios deste trabalho: encontrar a melhor forma de ordenar uma frase em LIBRAS.

Segundo a linguista e intérprete, é essencial detalhar primeiramente o espaço onde a ação ocorreu e após o que ou quem fez a ação. É como se primeiramente desenhássemos o cenário, após os personagens e então o que ocorreu. Desta forma, foi seguida a ordem: objeto, sujeito verbo (OSV).

Quando a frase não possui “lugar” na frase, ou seja, não existe as palavras “em” “no” ou “na”, o sujeito passa a ser o nome ao qual o verbo se refere. Como no exemplo:

Português	Juliana pegou a cadeira
Interlíngua	CADEIRA JULIANA PASSADO PEGAR

No exemplo nota-se que o verbo é acompanhado do seu tempo verbal. Esta regra é apresentada na seção regras gerais.

5.2.3 Regras gerais

As regras gerais são regras de adição ou exclusão de palavras para o contexto da LIBRAS. Estas regras geralmente representam a ligação entre os termos e são extremamente importantes por serem regras básicas foram descritas através das frases de exemplo da estrutura sintática da LIBRAS (STROBEL, FERNANDES, 1998).

A primeira regra básica é a não existência de preposições. Na LIBRAS, o verbo ou, se existir, o adjetivo carregam todas as indicações e sentido de gênero, número e grau da frase. Então, não é preciso fazer esta ligação, desconsiderando todas as preposições.

Outra regra de exclusão é o não aparecimento dos verbos "ser" e "estar" e suas conjugações. Com a indicação do lugar ou com adjetivos não se faz necessário estes verbos na frase LIBRAS. Fernandes explica que "é como se os verbos ficassem na frase quase sempre no infinitivo (...) é marcado sintaticamente através de advérbios de tempo que indicam se a ação está ocorrendo no presente, passado ou futuro".

Nas regras de adição de palavras, uma das principais é a adição do tempo verbal antes do verbo. Como explicado anteriormente, o verbo carrega indicações sejam elas de tempo ou de ação. A indicação de ação direcional é feita por marcas manuais. Estas marcas são os gestos feitos com outras partes do corpo que não as mãos. Elas podem ser a direção do olhar, expressão facial.

Quando a frase é uma frase interrogativa, o pronome interrogativo pode ser colocado no final da frase. Ou seja, para a frase "Porque faltar aula?" ficaria "AULA FALTAR PORQUE". Pronomes interrogativos vêm sempre com a expressão facial de dúvida, indicando a pergunta.

Na presença de palavras ou expressões de intensidade, há duas formas de expressar. Ela pode ser dita com a expressão facial característica ou pode-se repetir a palavra. Como no exemplo a seguir.

Português	Eu comi sem parar
Interlíngua	EU PASSADO COMER COMER COMER

No exemplo mostra a segunda representação: repetição da palavra para identificar a intensidade. Para isso, repete-se três vezes a palavra a qual a intensidade se refere.

Nomes próprios em LIBRAS são soletrados. Normalmente, cada pessoa possui um gesto que significa o seu nome. Mas, quando apresentado a primeira vez o nome é soletrado e após feito o gesto em LIBRAS que representará o nome nas próximas vezes que este for referenciado.

Outras palavras são soletradas em LIBRAS, mas não se encontrou o motivo geral para todas. Algumas delas são "reais", quando se refere à moeda, "acho", "quem" e "nunca".

5.3 IMPLEMENTAÇÃO

A implementação do trabalho foi feita na linguagem de programação PHP e o "anotador morfossintático", ferramenta online distribuída pelo LX-Center. Ela segue o modelo proposto mostrado na Figura 11. Seguindo a figura, são apresentados os módulos desenvolvidos.

O texto é recebido por uma variável compartilhada entre os módulos pais "navegador" e "processamento de texto".

Texto de exemplo:

Esta é uma frase de exemplo. Utilizamos ela para testar o trabalho.

Após a recuperação deste texto, ele é enviado para a ferramenta POSTagger para a categorização e anotação dos termos. A categorização feita pela ferramenta segue o padrão "palavra/TAGS". Então, com estas TAGS pode fazer o trabalho. Elas podem ser encontradas da seção "Anexos" no final deste trabalho. Uma frase de saída do POSTagger é representada a seguir:

```
<p><s> Esta/DEM#fs é/SER/V#pi-3s uma/UM#fs frase/FRASE/CN#fs de/PREP exemplo/EXEMPLO/CN#ms .*//PNT </s>
<s> Utilizamos/UTILIZAR/V#pi-1p ela/PRS#fs3 para/PREP testar/TESTAR/V#INF-nInf o/DA#ms trabalho/TRABALHO/CN#ms ./PNT </s></p>
```

A marca indica início de parágrafo, assim como o final do mesmo. A marca <s> e </s> é o mesmo para frases. Pode ser observado que no modelo "palavra/TAG" apresenta várias *tags* para os termos.

No módulo Tagger as frases do texto são separadas em uma matriz , onde cada linha da matriz corresponde a uma frase. Nas colunas de cada linha (que corresponde a cada frase) tem-se os termos de cada frase e sua respectiva TAG. Ou seja, primeiramente o texto separado por frases, na seguinte forma:

```
Array
([0] => esta/DEM#fs é/?/V#? uma/UM#fs frase/FRASE/CN#fs de/PREP exemplo/EXEMPLO/CN#ms .*/
 [1] => utilizamos/UTILIZAR/V#pi-1p ela/PRS#fs3 para/PREP testar/TESTAR/V#INF-nInf o/DA#ms trabalho/TRABALHO/CN#ms .

)
```

E então, cada frase em palavras e suas palavras divididas em palavras e *tags*. Esta estrutura pode ser vista no Anexo 3.

As regras de substituição (no caso de palavras compostas) e regras gerais já podem ser aplicadas. Estas consistem em adicionar tempo do verbo e substituí-lo por sua forma infinitiva e colocar quando existir o pronome de interrogação ao final da frase se for uma pergunta. São feitas a seguir do método simplifica. Após o módulo de simplificação e substituição, a aplicação das regras de ordenação pode ser realizada. O módulo de ordenação segue as regras definidas anteriormente.

A implementação considera que a frase de entrada do processamento esteja na forma mais simplificada possível (perto do formato SVO) e foi desenvolvido para a combinação com a ferramenta Simplifica. Esta ferramenta foi desenvolvida na Universidade de São Paulo, sede em São Carlos para simplificar textos lexicalmente e/ou sintaticamente. Entretanto, durante o período de desenvolvimento deste, a ferramenta não estava funcional.

A função principal é a que recebe o texto selecionado pelo usuário no navegador e o encaminha para todas as etapas até a geração da interlíngua. Esta função é apresentada no pseudocódigo resumido a seguir. As funções do PHP estão em itálico e as funções criadas no trabalho em negrito.

Pseudocódigo resumido da função principal

```
Função principal (texto){
início
    texto = texto_selecionado;
    (…)
    link="http://lxcenter.di.fc.ul.pt/(...)".urlencode($texto)."&submit=Annotate";
    sFile = file_get_contents($link);

    /*Separa frases da saída do Lx a cada pontuação*/
    vetorFrases = separaFrasePorPontuação('/PNT', $texto);

    /*Separa os termos de cada frase*/
    foreach ($vetorFrases as $frase){
        vetPalavras[n] = explode(' ', frase);
        n++;
    }
    /*Separa palavra das tags*/
    listaTag= separaTag(vetorPalavras, tamanhoVetor, qtasPalavras);
    /*Pre-processamento*/
    textoPreProcessado = substitui(listaTag, $amanhoVetor);
    /*Regras de ordenacao*/
    ordenado = ordenar(textoPreProcessado, tamanhoVetor);
    /*Refinamento, regras gerais e geração da interlingua*/
    interlingua = gerarInterlingua(ordenado, tamanhoVetor, QtasPalavras);
    interlingua = json_encode(interlingua);
fim
}
```

A função “substitui” é onde é feito o pré-processamento já especificado anteriormente. Nela estão as palavras compostas a serem substituídas e os verbos a serem complementados com seu tempo verbal. Esta função é apresentada no código resumido a seguir.

Pseudocódigo resumido da função substitui

```
Função substitui(palavras, qtasFrases){
início
    for (j=0; j < qtasFrases; j++){
        for (i=0; i<tamanho; i++){
                procura = palavras[j][i][0];

                SWITCH (procura){
                    CASE "zebra":
                            palavras[j][i][0]="CAVALO^LISTRAS";
                            break;
                        (...)
                    /*case para toda as palavras compostas como abaixo*/
                    CASE PALAVRA :
                            palavras [j][i][0]= PALAVRA^PARA^SUBSTITUIR;
                            break;

                }

        }
    }
        preProcessado = tratarVerbo(palavras, qtasFrases);
        retorna preProcessado;
fim

}
```

Na função “tratarVerbo”, são acrescentadas palavras para que em LIBRAS tenha sentido. Nesta função, é acrescido o tempo verbal após o verbo “PASSADO” ou “FUTURO” após a indicação do verbo.

A ordenação dos termos pode ser feita de várias formas. Não há uma forma genérica que esteja completamente correta para todos os casos, entretanto foi sugerido pela linguista algumas formas de ordenação mais comuns e que o falante LIBRAS tem entendimento do que é falado.

A LIBRAS, sendo uma língua visuo-espacial, precisa que primeiramente o cenário seja descrito, após os personagens e então a ação. Seguindo este formato, que pode ser considerado OSV, a maioria dos casos a frase será válida.

A função que implementa esta funcionalidade é a “ordenar”. Esta é apresentada resumidamente no Anexo 2.

Outra função importante a ser apresentada é a função gerarInterlingua. Até aqui as preposições ditas nas regras gerais que não existe em LIBRAS não foram desconsideradas. É nesta função em que as palavras são selecionadas para fazerem parte ou não da interlíngua. Por esta razão, neste momento que as preposições são desconsideradas, pois até então as mesmas são

utilizadas para auxiliar na ordenação dos termos. Esta função também é responsável por realizar a extração das *tags* na estrutura, mantendo apenas os termos da interlíngua.

Código função gerarInterlíngua

```
função gerarInterlingua(vetorCompleto, qtasFrases, qtasPalavras){
início
    needle1 = "/PREP";
    needle2 = "/DA";
    needle3 = "/CL";
    for (j=0; $j<qtasFrases; j++){
          for (i=0; $i<qtasPalavras; i++){
                    pocuraoPara= procuraEm(vetorCompleto[j][i][1], $needle1);
                    procuraAO= procuraEm(vetorCompleto[j][i][1], $needle2);
                    procuraSE= procuraEm (vetorCompleto[j][i][1], $needle3);

                    if (isset(vetorCompleto[j][i][0])){//está preenchido
                              if (procuraPara === false){ //se nao for preposicao 'para'
                                        if (procuraAO === false){ //se nao for preposicao 'o' 'a'
                                                  if (procuraSE ===false){
                                                  vetorTermos[j][i]=vetorCompleto[j][i][0];
                                                  }}}}
    }}
    retorna vetorTermos;
fim
}
```

6. ESTUDO DE CASO

6.1 Exemplos de frases na interlíngua Português-Libras

Com as funções descritas no capítulo anterior implementadas, foram escolhidas frases extraídas do site da Rede Globo de Televisão (GLOBO, 2014) onde pode ser vista a aplicação das regras. Cada uma exemplifica a ação de pelo menos uma das regras implementadas. São manchetes de notícias, pois precisam ser curtas e simples para exemplificar as regras de forma mais clara e objetiva.

6.1.1 Exemplo 1

O primeiro exemplo mostra a ação das regras presentes no pré-processamento, ou seja, as regras onde o termo em português é representado pela composição de palavras em LIBRAS e o acréscimo do tempo verbal da frase após o verbo. Neste exemplo, a frase em português é a frase extraída diretamente da notícia e a frase em Interlíngua é a frase resultante da execução das etapas de processamento de texto desenvolvidas neste trabalho.

Português	Listras das zebras servem para espantar insetos, afirma pesquisa.[1]
Interlíngua:	INSETOS LISTRAS CAVALO^LISTRAS SERVIR ESPANTAR. PESQUISA AFIRMAR.

FONTE: http://g1.globo.com/natureza/noticia/2012/02/listras-das-zebras-servem-para-espantar-insetos-afirma-pesquisa.html

Na frase apresentada, vê-se a substituição da palavra zebra, uma palavra composta em LIBRAS, pela sua representação CAVALO^LISTRAS. Também é possível identificar a ordenação, colocando os nomes (objeto e sujeito) e após a ação representada pelos verbos.

Na frase, o verbo “espantar” já encontra-se em sua forma infinitiva e os outros verbos no presente. Por isso as palavras PASSADO ou FUTURO não apareceram. Para reproduzir este evento foi escolhido o próximo exemplo.

6.1.2 Exemplo 2

Português:	Cantor desfilou pela Viradouro, no segundo dia de desfiles do Grupo Série A
Interlíngua:	viradouro cantor PASSADO DESFILAR. segundo dia desfiles grupo.

FONTE: http://globotv.globo.com/globocom/ego/v/voar-voar-subir-subir-biafra-desfila-pela-viradouro-e-canta-trecho-de-sucesso/3184763/

Para o segundo exemplo, é possível ver o acréscimo do tempo verbal após o verbo: “PASSADO DESFILAR” para o verbo em português “desfilou”. Nela vê-se também a separação da frase pela presença da vírgula e a ordenação OSV.

6.2 Exemplo de texto na interlíngua Português-Libras

Esse estudo de caso consiste na aplicação do sistema desenvolvido em um texto completo, ou seja, mais de uma frase. Este foi aplicado em uma notícia do site da Globo apresentada a seguir.

Inscrições para cursos gratuitos de inglês, espanhol, português e Libras
Matrícula inicia nessa terça-feira (11). Ao todo, são oferecidas 160 vagas em Petrolina, PE.
As inscrições para aulas de Língua Inglesa, Língua Espanhola, Língua Portuguesa e Língua Brasileira de Sinais (Libras), oferecidas gratuitamente pelo Núcleo Municipal de Línguas (Numel) de Petrolina, Sertão pernambucano, começam nesta terça-feira (11) e encerram na quinta-feira (13). O início das aulas é na próxima segunda (17). Estão sendo disponibilizadas 40 vagas para cada curso, que tem duração de no mínimo um e no máximo dois anos. “O duração varia dependendo da quantidade de aulas que o aluno tenha por semana”, explicou a coordenadora do Numel, Stella Márcia de Alencar. Apenas o curso de Língua Portuguesa tem duração de 6 meses. As aulas acontecerão nos turnos manhã e tarde. De acordo com Stella, a instituição está tentando patrocínio para conseguir que os materiais sejam gratuitos aos alunos. “Caso não consigamos, será cobrado um valor pequeno para custear as apostilas”, disse.

FONTE: http://g1.globo.com/pe/petrolina-regiao/noticia/2014/02/inscricoes-para-cursos-gratuitos-de-ingles-espanhol-portugues-e-libras.html

Foi dividido entre os parágrafos, pela restrição de tamanho do POSTagger, para a entrada do sistema. O resultado é a aplicação das regras de substituição e ordenação dos termos. O resultado apresentado a seguir. A interlíngua está apresentada na divisão que o sistema faz para a geração. Esta divisão é baseada na pontuação do texto.

[inscrições cursos gratuitos inglês][espanhol][português e libras]
[matrícula PASSADO INICIAR essa terça-feira][todo][são oferecidas 160 petrolina][pe .*]
[inscrições aulas língua inglesa] [língua espanhola] [língua portuguesa] [língua brasileira de sinais] [libras] [,*][oferecidas gratuitamente núcleo municipal línguas] [numel] [Petrolina ,*] [sertão pernambucano] [? Esta terça-feira] [PASSADO ENCERRAR quinta-feira][,*][início aulas próxima segunda] [17] [estão SER disponibilizadas 40 false cada curso ,*][que PASSADO TER duração mínimo um e máximo dois anos .*] [\u201co duração PASSADO DEPENDER quantidade aulas que aluno TER semana \u201d ,*],[numel coordenadora PASSADO EXPLICAR ,*] [PASSADO STELLAR Márcia ALENCAR .*][apenas curso língua portuguesa PASSADO TER duração 6 meses .*] [aulas acontecerão turnos manhã e tarde] [acordo Stella ,*][instituições estão TENTAR patrocínio CONSEGUIR que materiais SER gratuitos alunos .*][caso não CONSEGUIR ,*][será COBRAR um false pequeno CUSTEAR apostilas \u201d ,*][PASSADO DIZER"]

Pelo estudo de caso, pode ser visto que o sistema apresentou em geral o comportamento previsto para a geração da interlíngua. As regras de substituição foram totalmente atendidas e os verbos foram todos colocados em sua forma no infinitivo acompanhados de seu tempo verbal.

Entretanto, algumas falhas foram verificadas:

- Palavras não esperadas como “false” ao decorrer do texto;

-Alguns símbolos não são reconhecidos. Eles apresentam-se na saída do sistema como “\u201co” ou “\u201d”, no caso das aspas;

- Alguns verbos não foram identificados como verbos pelo *tagger,* apresentados como “?”, pois o sistema substitui pelo infinitivo do verbo que consta na TAG e lá está “?” ou no caso do verbo “acontecerão” que foi etiquetado pelo *tagger* como nome próprio.

-As regras de ordenação funcionaram para algumas frases curtas como “numel coordenadora PASSADO EXPLICAR ,*” mas falhou em frases como “instituições estão TENTAR patrocínio CONSEGUIR que materiais SER gratuitos alunos .*”.

7. CONCLUSÃO E CONSIDERAÇÕES FINAIS

O trabalho apresentou como resultado uma interlíngua português-LIBRAS de forma a ser aceita pelos falantes LIBRAS, apresentando melhores resultados em frases simples e curtas. Este problema pretende-se resolver com a ferramenta Simplifica, onde as frases serão reescritas de maneira menos complexa. Entretanto, alguns ajustes ainda precisam ser feitos a engrandecer o trabalho.

Foram apresentadas regras de tradução português-LIBRAS e estas implementadas na linguagem PHP. Para o trabalho ficar de uma maneira ótima, mais regras devem ser encontradas e implementadas. A dificuldade está na semelhança entre a língua portuguesa e LIBRAS e mesmo assim na independência delas. Com mais regras implementadas, a interlíngua ficará mais entendível aos falantes LIBRAS.

A contribuição para a área de mineração de texto foi feita a partir da implementação das regras de substituição e ordenação. Com esta pretende-se expandir para um melhor funcionamento da ferramenta.

Mesmo sendo uma ferramenta web, o trabalho pode ser melhorado utilizando a ferramenta offline para a função de *tagger*. Assim, mesmo sem conexão de rede, a interlíngua pode ser gerada.

Com o desenvolvimento deste, percebeu-se o quanto diferentes áreas podem ser combinadas com a computação e o quão promissor é a continuação da pesquisa e desenvolvimento de ferramentas que auxiliem o cotidiano de pessoas com necessidades específicas.

A partir dos estudos bibliográficos e discussões com o grupo, percebeu-se que a ferramenta terá grande valia na vida de pessoas com deficiências específicas auditivas e pessoas que gostariam de aprender uma nova língua. Á pessoa com surdez é entregue uma ferramenta onde ela poderá facilmente traduzir os textos em páginas web para a sua língua e, para falantes de outras línguas, pode ser usado para o aprendizado em uma nova: LIBRAS.

Referências

Agência USP de Notícias: Softwares facilitarão a compreensão de textos na internet. Disponível em http://www.usp.br/agen/?p=8157, 14 de novembro 2013.

BIAP: International bureau for Audiophonology . Disponível em http://www.biap.org/, 22 de outubro de 2013.

BICK, Eckhard. The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus - Dinamarca , 2000. Disponível em http://beta.visl.sdu.dk/postscript/PLP20-amilo.ps, 13 de novembro de 2013.

BRANCO, António; SILVA, João. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva (orgs.), Proceedings of the 4th International Conference on Language Resources and Evaluation , Paris, 2004. p.507-510.

CARMO, Josué Geralto Botura do. Tecnologia Assistiva. 2005. Disponível em http://www.educacaoliteratura.com/index%20140.htm, 12 de abril de 2013.

CHEN, Hsinchun. Knowledge Managment Systemss: A text mining perspective. Knowledge Computing Corporation Department of Management Information Systems Eller College of Business and Public Administration. The University of Arizona. Tucson-Arizona, 2001.

COGROO. CoGrOO - Corretor Gramatical acoplável ao LibreOffice. Disponível em http://cogroo.sourceforge.net/, 4 de dezembro de 2013.

ELMASRI, Ramez; NAVATHE, Shamkant B. Sistemas de Banco de Dados, 6ª edição. Editora Pearson do Brasil, 2011.

GLOBO. G1 O Portal de Notícias da Globo. Disponível em http://g1.globo.com/, 26 de Fevereiro de 2014.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic; UTHURUSAMY, Ramasamy. Advances in Knowledge Discovery and Data Mining, Menlo Park, CA, MIT Press, Boston, MA, 1996.

FELIPE, Tanya. Introdução à Gramática da LIBRAS . Artigo publicado pela SEESP, In: Giuseppe Rinaldi et al.Educação Especial Deficiência Auditiva, Série Atualidades Pedagógicas, Brasília, 1997.

FINATTO, Maria José B. Sobre a Eficácia do MXPOST Etiquetador morfossintático para o português do Brasil. UFRGS, Porto Alegre-RS, 2011.

JANUÁRIO, Guilherme Carvalho; LEITE, Leonardo Alexandre Ferreira; KOGA, Marcelo Li.

POLI-LIBRAS: Um Tradutor de Português para LIBRAS. São Paulo, 2010.

HAND TALK. HandTalk: tradutor de sites automático para Libras. Disponível em http://www.handtalk.me/ , 4 de fevereiro de 2014.

IBGE. Censo Demográfico: Características gerais da população, religião e pessoas com definiência. Rio de Janeiro, p.71-89, 2010.

LEWIS, M. Paul; SIMONS, Gary F.; FENNIG, Charles D. Ethnologue: Línguas do Mundo, edição XVII SIL Internacional. Dallas-Texas, 2013. Disponível em http://www.ethnologue.com, 8 de fevereiro de 2014.

LIBREOFFICE. Disponível em http://pt-br.libreoffice.org/libreoffice/, 3 de março de 2014.

LX-CENTER. Language Resources and Technology for Portuguese. Disponível em http://lxcenter.di.fc.ul.pt/, 21 de Novembro de 2013.

MARTINS, Ronaldo; PELIZZONI, Jorge; HASEGAWA, Ricardo. PULØ - Para um sistema de tradução semi-automática português-libras. Anais do XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo-RS, 2005.

MEC. Saberes e Práticas da Inclusão: Desenvolvendo competências para o atendimento às necessidades educacionais especiais de alunos surdos. Brasília, 2006.

MOURA, Débora Rodrigues. O uso da LIBRAS no Ensino de Leitura de Português como segunda língua para Surdos: Um estudo de caso em uma perspectiva bilíngüe. São Paulo-SP, 2008.

PEREIRA, T. F.; ALUISIO, S. M. Editor de Anotação de Simplificação: Construção. Technical Report NILC-TR_08_12. São Carlos-SP, 2008.

PHP. PHP: Manual do PHP. Disponível em http://php.net/manual/pt_BR/ , 22 de Novembro de 2013.

POLILIBRAS. http://www.polilibras.com.br/, disponível em 29 de Janeiro de 2014.

PORSIMPLES. Simplifica - Sistema para simplificação de textos. Disponível em http://www.nilc.icmc.usp.br/porsimples/simplifica/sobre.php em 8 de Janeiro de 2014.

RATNAPARKHI, Adwait. Maximum Entropy Models for Natural Language Ambiguity Resolution. Dissertação da Universidade da Pensilvânia. Pensilvânia-EUA, 1998.

SANTOS, Guilherme Spolavori dos. Produção de Textos Paralelos em Língua Portuguesa e uma interlíngua LIBRAS. PUCRS, Porto Alegre, 2009.

SIMPLIFICA. Disponível em http://www.nilc.icmc.usp.br/porsimples/simplifica/ , 8 de janeiro de 2014.

STROBEL, Karin Lilian, FERNANDES, Sueli. Aspectos Linguísticos da LIBRAS: Língua Brasileira de Sinais. Curitiba, 1998.

TAN, Ah-Hwee. Text Mining: The state of the art and challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced. Kent Ridge Digital Labs. Singapura, 1999.

TOMAZELA, Élen Cátia; BARROS, Cláudia Dias de; RINO, Lucia Helena Machado. Avaliação da anotação semântica do PALAVRAS e sua pós-edição manual para o Corpus Summ-it. Linguamática. ISSN: 1647–0818. Volume 2. Número 3. Páginas 29-42. São Carlos-SP, 2010.

UNITED STATES (105th Congress of the United States), *Public Law 105-394: Assistive Technology Act, 1988.* Disponível em http://en.wikisource.org/wiki/Assistive_Technology_Act_of_1998#Sec._3.

VYGOTSKI, Lev Semenovitch. A formação social da mente. *Psicologia* 153.1989.

Anexos

Anexo 1: TABELAS de TAG's reconhecidas pela ferramenta LX-Tagger, disponível em http://lxcenter.di.fc.ul.pt/services/en/LXServicesSuite.html

Tag	**Categoria**	**Examples**
ADJ	Adjetivos	bom, brilhante, eficaz, …
ADV	Advérbios	hoje, já, sim, felizmente, …
CARD	Cardinais	zero, dez, cem, mil, …
CJ	Conjunções	e, ou, tal como, …
CL	Cliticos	o, lhe, se, …
CN	Nomes Comuns	computador, cidade, ideia, …
DA	Artigos definidos	o, os, …
DEM	Demonstrativos	este, esses, aquele, …
DFR	Denominadores de Frações	meio, terço, décimo, %, …
DGTR	Números Romanos	VI, LX, MMIII, MCMXCIX, …
DGT	Digitos	0, 1, 42, 12345, 67890, …
DM	Marcador de Discurso	olá, …
EADR	Endereços Eletrônico	http://www.di.fc.ul.pt, …
EOE	Final de Enumeração	etc
EXC	Exclamativos	ah, ei, etc.
GER	Gerundios	sendo, afirmando, vivendo, …
GERAUX	Gerundios "ter"/"haver" em tempos compostos	tendo, havendo …
IA	Artigos Indefinidos	uns, umas, …
IND	Indefinidos	tudo, alguém, ninguém, …
INF	Infinitivos	ser, afirmar, viver, …

INFAUX	Infinitivos "ter"/"haver" e tempos compostos	ter, haver …
INT	Interrogativos	quem, como, quando, …
ITJ	Interjeição	bolas, caramba, …
LTR	Letras	a, b, c, …
MGT	Magnitude de Classes	unidade, dezena, dúzia, resma, …
MTH	Meses	Janeiro, Dezembro, …
NP	Nomes de Frases	idem, …
ORD	Ordinais	primeiro, centésimo, penúltimo, …
PADR	Parte de Endereço	Rua, av., rot., …
PNM	Parte de Nome	Lisboa, António, João, …
PNT	Pontuação	., ?, (, …
POSS	Possessivos	meu, teu, seu, …
PPA	Particípio passado sem tempo composto	afirmados, vivida, …
PP	Frases Preposicionais	algures, …
PPT	Particípio passado em tempo composto	sido, afirmado, vivido, …
PREP	Preposições	de, para, em redor de, …
PRS	Personais	eu, tu, ele, …
QNT	Quantificadores	todos, muitos, nenhum, …
REL	Relativos	que, cujo, tal que, …
STT	Títulos Sociais	Presidente, drª., prof., …
SYB	Simbolos	@, #, &, …
TERMN	Opção de terminação	(s), (as), …
UM	"um" ou "uma"	um, uma

UNIT	Unidades de medidas abreviadas	kg., km., …
VAUX	Finitivo "ter" ou "haver" em tempos compostos	temos, haveriam, …
V	Verbos (outros além de PPA, PPT, INF or GER)	falou, falaria, …
WD	Dias da semana	segunda, terça-feira, sábado, …
LADV1…LADVn	Advérbios compostos	de facto, em suma, um pouco, …
LCJ1…LCJn	Conjunções compostas	assim como, já que, …
LDEM1…LDEMn	Demonstrativos compostos	o mesmo, …
LDFR1…LDFRn	Denominadores de fração compostos	por cento
LDM1…LDMn	Marcadores de discussão compostos	pois não, até logo, …
LITJ1…LITJn	Interjeição composta	meu Deus
LPRS1…LPRSn	Personais compostos	a gente, si mesmo, V. Exa., …
LPREP1…LPREPn	Preposições compostas	através de, a partir de, …
LQD1…LQDn	Quantificadores compostos	uns quantos, …
LREL1…LRELn	Relativos compostos	tal como, …

Outras TAG'S

Tag	**Descrição**
m	Masculino
f	Feminino
s	Singular
p	Plural
dim	Diminutivo
sup	Superlativo

comp	Comparativo
1	Primeira Pessoa
2	Segunda Pessoa
3	Terceira Pessoa
pi	Presente do Indicativo
ppi	Pretérito Perfeito do Indicativo
ii	Pretérito Imperfeito do Indicativo
mpi	Pretérito Mais que Perfeito do Indicativo
fi	Futuro do Indicativo
c	Condicional
pc	Presente do Conjuntivo
ic	Pretérito Imperfeito do Conjuntivo
fc	Futuro do Conjuntivo
imp	Imperativo

Anexo 2: Pseudocódigo resumido da função ordenar

Código resumido da função ordenar

```
função ordenar($words_tagged, $qtasFrases){
início
    qtasPalavras = count($words_tagged);
    for (j=0; j<qtasFrases; j++){
        for (i=0; i<qtasPalavras; i++){
                /*Possui preposição*/
                if(procura_em(words_tagged[j][i][1], "/PREP")){
                                prep_pos[j][count_prep] = i;
                                count_prep++;
                }
                 /*o mesmo para as tags que são utilizadas na ordenação*/
                    (...)
        }
}
/*TRATAR O QUE APARECE EM PRIMEIRO LUGAR*/
/*se tiver lugar, ou seja NO ou NA na frase = EM "prep" e depois "da"
então, tem lugar e este aparece primeiro descrevendo o cenário*/

/*for(p=0; p<qtas_prep; p++){
        for(da=0; $da<count(da_pos); da++){
                if (da_pos[da]-prep_pos[p]==1){
                        lugar=true;
                        pos_da_lugar = $da;
                        pos_prep_lugar = p;
                }
        }
}

/*se tiver lugar, ele é o primeiro a ser mostrado*/
if(lugar){
        interlingua[0] = words_tagged[da_pos[pos_da_lugar]+1][0];
}
else{
        //Se não tem lugar explícito, começa a tratar os nomes
        //tratar nomes
        if (count==1){ //se tem um nome
                posicao = nouns_pos[0]; //busca a posição em que ele se encontra
                interlingua[0] = words_tagged[$posicao][0]; //coloca como primeiro
        }
        else{
                if (count>2){ //se tiver mais que dois nomes
                        inicio = count(interlingua);
                        if (nouns_pos[1]-nouns_pos[0]==1){
```

```
                                interlingua[inicio]=words_tagged[$nouns_pos[0]][0];
                                interlingua[inicio+1]=words_tagged[$nouns_pos[1]][0];
                        }
                        else{
                                interlingua[inicio]=words_tagged[nouns_pos[1]][0];
                                interlingua[inicio+1]=words_tagged[nouns_pos[0]][0];
                        }
                }
        }
}

if(count_v==1){
interlingua[pos+1]=words_tagged[verb_pos[0][0]];
}
        /*percorre as tags procurando nomes(CN)*/
        for (i=0; i<count(words_tagged); i++){
                if(strpos(word_tagged[1][i], "CN")!==false){
                        /*salva posicao onde está
                        nouns_pos[count] = i;
                        count++;
                }
                else if(word_tagges[1][i], "PRS")!==false){
                        /*salva posicao onde esta*/
                        nouns_pos[count] = i;
                        count++;
                }
        }
        nouns_quantity = count(nouns_pos);
        }
    retorna words_tagged;
fim
}
```

Anexo 3: Estrutura do texto divida por frases e esta pelas palavras e *tags* que a compõem

```
Array
(
   [0] => Array
      (
         [0] => Array
            (
               [0] => esta
               [1] => /DEM#fs
            )
         [1] => Array
            (
               [0] => ?
               [1] => /?
            )

         [2] => Array
            (
               [0] => uma
               [1] => /UM#fs
            )

         [3] => Array
            (
               [0] => frase
               [1] => /FRASE/CN#fs
            )

         [4] => Array
            (
               [0] => de
               [1] => /PREP
            )

         [5] => Array
            (
               [0] => exemplo
               [1] => /EXEMPLO/CN#ms
            )

         [6] => Array
            (
               [0] => .*
               [1] => /
            )
      )
```