SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 16.02.2004

Assinatura: Ana-Paula Jampai Junc

Alinhamento lexical de textos paralelos português-inglês

Aline Maria da Paz Silva

Orientadora: Profa. Dra. Maria das Graças Volpe Nunes

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências: Área – Ciências de Computação e Matemática Computacional.

USP – São Carlos
Fevereiro/2004

A Comissão Julgadora:

Profa. Dra. Maria das Graças Volpe Nunes

Profa. Dra. Sandra Maria Aluisio

Profa. Dra. Ariadne Maria Brito Rizzoni Carvalho

Ariadne M. B. R. Carvalho

A Deus que me capacita para a realização das minhas tarefas. Aos meus pais queridos e minha irmã amada que com amor e paciência enxugaram as minhas lágrimas e me incentivaram a prosseguir.

Agradecimentos

Meus mais sinceros agradecimentos...

A Deus que tem me sustentado e me guiado a cada dia durante todo este período de aprendizado, que tem sido meu refúgio e fortaleza nos momentos difíceis de tribulação.

Aos meus pais Everaldo e Angelita e à minha irmã Eveline que foram meus maiores motivadores nesta etapa, acreditando na minha capacidade e me fornecendo todo o amor e os recursos necessários para que eu pudesse concluir este mestrado.

À minha amiga Maria Fernanda Teline, que esteve ao meu lado ao longo destes dois anos compartilhando das dificuldades e incertezas, e à sua família (Tio José, Tia Izabel, Pamela e Emerson), que me acolheu como membro da família e me ajudou a seguir em frente nos meus momentos de desânimo.

Aos meus amigos, familiares e irmãos em Cristo que estiveram orando por mim, intercedendo com amor pelo meu sucesso.

À minha orientadora Graça Nunes que me indicou o caminho, esclarecendo minhas dúvidas, sendo paciente e me acalmando nos momentos certos.

A todos os meus colegas do NILC, em especial Helena, Jorge Pelizzone, Ivair e Marcela, que contribuíram diretamente para o desenvolvimento deste trabalho.

A CAPES pelo suporte financeiro.

Sumário

LISTA DE FIGURAS.....	iii
LISTA DE TABELAS.....	iv
RESUMO.....	v
ABSTRACT.....	vi
CAPÍTULO 1.....	1
INTRODUÇÃO.....	1
1.1. MOTIVAÇÃO E OBJETIVOS.....	2
1.2. ORGANIZAÇÃO.....	4
CAPÍTULO 2.....	5
ALINHAMENTO LEXICAL DE TEXTOS PARALELOS.....	5
2.1. MÉTODOS DE ALINHAMENTO LEXICAL.....	6
2.1.1. <i>Métodos Empíricos</i>	7
Método <i>SIMR</i>	7
Uma Avaliação do <i>SIMR</i> no Alinhamento Português-Inglês.....	11
2.1.2. <i>Métodos Híbridos</i>	12
Método <i>LWA</i>	13
2.2. TÉCNICAS DE AVALIAÇÃO.....	18
CAPÍTULO 3.....	23
ALINHAMENTO LEXICAL PORTUGUÊS-INGLÊS.....	23
3.1. ESCOLHA DOS MÉTODOS.....	23
3.2. CONSTRUÇÃO DE RECURSOS LINGÜÍSTICOS.....	25
3.2.1. <i>Cópus de Teste</i>	26
3.2.2. <i>Cópus de Referência</i>	27
3.2.3. <i>Listas de Unidades de Classe Fechada</i>	28
3.2.4. <i>Listas de Unidades Multipalavras</i>	28
CAPÍTULO 4.....	31
EXTRAÇÃO AUTOMÁTICA DE MULTIPALAVRAS.....	31
4.1. OS CÓRPUIS UTILIZADOS PARA A EXTRAÇÃO DE MULTIPALAVRAS.....	32
4.2. TÉCNICAS DE EXTRAÇÃO DE UNIDADES MULTIPALAVRAS.....	33
4.2.1. <i>Esperança Mútua</i>	33
Modificações no método da Esperança Mútua.....	35
4.2.2. <i>N-gram Statistic Package (NSP)</i>	36
4.3. RESULTADOS OBTIDOS.....	38
4.3.1. <i>Listas Originais</i>	38
4.3.2. <i>Eliminando Unidades Não Relevantes</i>	39
4.3.3. <i>Listas Finais</i>	41
4.4. AVALIAÇÃO DAS LISTAS OBTIDAS.....	41
4.4.1. <i>Etapa 1 - Antes da Eliminação (Listas Originais)</i>	42
4.4.2. <i>Etapa 2 - Depois da Eliminação (Listas Processadas)</i>	45
CAPÍTULO 5.....	49
IMPLEMENTAÇÃO E AVALIAÇÃO DOS ALINHADORES LEXICAIS.....	49

5.1. IMPLEMENTAÇÃO DOS ALINHADORES LEXICAIS.....	49
5.1.1. <i>Alinhador SIMR</i>	49
5.1.2. <i>Alinhador LWA</i>	52
5.2. AVALIAÇÃO DOS ALINHADORES LEXICAIS	54
5.2.1. <i>Alinhador SIMR</i>	54
5.2.2. <i>Alinhador LWA</i>	56
Identificando Fatores que Levam a Baixos Valores de Revocação e Precisão.....	58
Etapa 1 - Influência das listas extraídas automaticamente no processo de alinhamento.....	58
Etapa 2 - Influência das traduções muito livres	59
Etapa 3 - Tamanho do corpus utilizado.....	59
Etapa 4 - Influência da etiquetagem POS no alinhamento.....	61
CAPÍTULO 6	65
CONCLUSÕES	65
REFERÊNCIAS BIBLIOGRÁFICAS	67
APÊNDICE 1: UNIDADES MULTIPALAVRAS EXTRAÍDAS PELAS TÉCNICAS IMPLEMENTADAS	71
1. ESPERANÇA MÚTUA	71
2. PACOTE NSP	72
APÊNDICE 2: ALINHAMENTOS CORRETOS PROPOSTOS PELOS ALINHADORES <i>LWA</i> E <i>SIMR</i>	73
1. ALINHADOR <i>SIMR</i>	73
2. ALINHADOR <i>LWA</i>	74

Lista de Figuras

Figura 1 - Espaço do Bitexto (Melamed, 2000)	8
Figura 2 - Expansão do retângulo de busca (Melamed, 2000)	9
Figura 3 - Tratamento de Expressões de Classe Aberta e de Classe Fechada	16
Figura 4 - Trecho do corpus de teste	20
Figura 5 - Técnicas de Alinhamento	24
Figura 6 - Exemplo de um fragmento de um par de textos paralelos do corpus de teste	27
Figura 7 - Regras de formação de tokens (Teline, Manfrim & Aluísio, 2003)	37
Figura 8 - Comparação da quantidade de unidades geradas pelas técnicas de extração	42
Figura 9 - Comparação entre a quantidade de unidades eliminadas das listas geradas pelas técnicas de extração	43
Figura 10 - Diferença entre as listas geradas pelos métodos extratores	45
Figura 11 - Comparação entre a quantidade de unidades obtidas após o processo de eliminação	46
Figura 12 - Diferença entre as listas obtidas após o processo de eliminação	47
Figura 13 - Contraste entre os valores de revocação, precisão e medida-F das tarefas de alinhamento utilizando as listas extraídas automaticamente	57

Lista de Tabelas

Tabela 1 - Lista de pares de tradução	15
Tabela 2 - Quantidade de unidades das listas geradas pelas técnicas de extração selecionadas	39
Tabela 3 - Quantidade de unidades das listas obtidas após o processo de eliminação	41
Tabela 4 - Quantidade de unidades eliminadas das listas originais	43
Tabela 5 - Quantidade de unidades coincidentes nas listas originais	44
Tabela 6 - Quantidade de unidades da diferença entre as listas geradas pelos métodos extratores	44
Tabela 7 - Quantidade de unidades coincidentes nas listas processadas	46
Tabela 8 - Quantidade de unidades no conjunto diferença entre as listas processadas	47
Tabela 9 - Fragmento de bitexto visto pelo <i>SIMR</i>	50
Tabela 10 - Mapeamentos produzidos pelo <i>SIMR</i>	51
Tabela 11 - Exemplo de alinhamentos propostos pelo <i>SIMR</i>	51
Tabela 12 - Exemplo de bitexto usado como entrada para o <i>LWA</i>	52
Tabela 13 - Escores de associação produzidos pelo <i>LWA</i>	53
Tabela 14 - Exemplos de alinhamentos propostos pelo <i>LWA</i>	54
Tabela 15 - Valores de revocação, precisão e medida-F obtidos na realização da tarefa de alinhamento	55
Tabela 16 - Valores de revocação, precisão e medida-F considerando alinhamentos parcialmente corretos	55
Tabela 17 - Total de alinhamentos propostos em cada uma das tarefas de alinhamento	57
Tabela 18 - Valores de revocação, precisão e medida-F obtidos nas tarefas de alinhamento utilizando listas extraídas automaticamente	57
Tabela 19 - Valores de revocação, precisão e medida-F obtidos utilizando listas de unidades multipalavras extraídas manualmente	58
Tabela 20 - Valores de revocação, precisão e medida-F obtidos utilizando um subconjunto do corpus <i>PE</i> e as listas de unidades multipalavras extraídas manualmente	59
Tabela 21 - Valores de revocação, precisão e medida-F obtidos utilizando o novo corpus (contendo 87 pares de textos paralelos)	60
Tabela 22 - Valores de revocação, precisão e medida-F obtidos utilizando o subconjunto do corpus <i>PE</i> com as etiquetas morfológicas corrigidas	62
Tabela 23 - Valores de revocação, precisão e medida-F considerando alinhamentos parcialmente corretos	62

Resumo

Num cenário mundial de rápida expansão das relações interculturais e da transmissão de conhecimento científico e tecnológico, a língua muitas vezes representa uma barreira para a comunicação. Ferramentas computacionais, como os tradutores automáticos, podem acelerar e facilitar a comunicação escrita e oral entre povos de línguas e culturas diversas. Algoritmos e técnicas de alinhamento de textos paralelos têm recebido muita atenção nesse cenário. Neste projeto foram implementadas algumas técnicas de alinhamento lexical de textos paralelos para um corpus em português do Brasil e inglês, que foram avaliadas segundo as métricas de revocação e precisão, e seus resultados foram analisados.

Abstract

In a scenario of fast expansion of intercultural relations and transmission of scientific and technological knowledge, the language can represent a communication barrier. Software applications, such as machine translation systems, can accelerate and facilitate written and verbal communication between people of different languages and cultures. Algorithms and techniques of parallel texts alignment has been received a lot of attention in this context. In this project some techniques of parallel texts lexical alignment for a Brazilian Portuguese and English corpus were implemented and evaluated according to recall and precision metrics, and the results were analyzed.

Capítulo 1

Introdução

A idéia de sistemas automáticos para tradução de textos de uma língua para outra é quase tão antiga quanto a idéia de sistemas de computador. Sistemas de tradução têm desempenhado um papel muito importante, principalmente no cenário atual de globalização em que as diferentes línguas existentes dificultam a comunicação, restringindo a troca de conhecimento entre os diferentes povos.

Nos últimos anos, uma das áreas do Processamento de Língua Natural (PLN) que mais tem se desenvolvido na construção de sistemas que auxiliem na superação dessa barreira lingüística é a de alinhamento de textos paralelos. O alinhamento de textos possibilita abordagens para tradução automática parcial ou completa de grandes corpú de texto.

Os textos paralelos são definidos na literatura como textos acompanhados de sua tradução em uma ou várias línguas, sendo também conhecidos como bitextos – texto fonte e texto alvo - quando envolvem apenas duas línguas no processo de alinhamento. Eles devem ser diferenciados dos chamados textos comparáveis, pois estes são textos pertencentes ao mesmo domínio escritos em línguas diferentes, mas não são necessariamente traduções uns dos outros.

A primeira descrição de um método de alinhamento foi feita por Martin Kay e Martin Röscheisen em 1987 (Kay & Röscheisen, 1993), reportando um método empírico que considerava apenas a distribuição das palavras no texto para realizar o processo de alinhamento entre os textos. Mais recentemente, tem sido possível obter grandes corpú digitalizados de textos acompanhados de suas traduções, como, por exemplo, o Canadian Hansard, que é formado pela transcrição, em inglês e francês, de procedimentos do parlamento Canadense, proporcionando a realização de estudos estatísticos em grande escala para a criação de métodos com resultados cada vez mais precisos.

No sentido geral, o alinhamento consiste em tentar encontrar as possíveis correspondências existentes entre as unidades que compõem o texto fonte e o texto alvo, qualquer que seja o nível da unidade utilizada. A classificação dos procedimentos de alinhamento tem como base a definição da unidade do texto a ser alinhada, podendo ser

parágrafos, sentenças, palavras e unidades multipalavras. Essa classificação determina o nível em que se dá o alinhamento, denominado nível de resolução.

Os primeiros trabalhos de alinhamento focavam a unidade sentencial. Embora os resultados obtidos com esse tipo de alinhamento fossem muito bons, descobriu-se que alinhamento sentencial não era uma opção razoável para bitextos com ruídos (Caseli, 2003), ou seja, textos que possuem erros gramaticais, de tradução ou pontuação (Church, 1993). Assim, alguns estudos foram desenvolvidos e mostraram que a unidade de texto que gera um bom alinhamento é a palavra, pois situando as posições das palavras no bitexto, as heurísticas dos algoritmos de alinhamento de sentenças podem ser exploradas igualmente no nível de palavras, como é o caso da heurística de cognatos (unidades com grafias semelhantes) que funciona melhor para palavras. A esse tipo de alinhamento dá-se o nome de alinhamento lexical.

1.1. Motivação e Objetivos

Uma das grandes motivações para o estudo de metodologias e técnicas de alinhamento de textos paralelos é que o conhecimento produzido pelo alinhamento lexical – dicionários e glossários bilíngües, listas de traduções, gramáticas de tradução, entre outros – pode ser utilizado em aplicações como a tradução automática, a recuperação de informação interlínguas (*cross-language retrieval*), a extração de terminologia e o aprendizado de idiomas. Além disso, ressalta-se a importância de investigar os métodos da literatura aplicados pela primeira vez ao português do Brasil.

Sendo assim, o objetivo deste trabalho é investigar, implementar e avaliar algumas técnicas de alinhamento lexical de textos paralelos português-inglês.

A metodologia utilizada neste trabalho seguiu as seguintes etapas: a) o levantamento bibliográfico de técnicas e metodologias de alinhamento lexical de textos paralelos, que geralmente envolvem diferentes pares de línguas; b) a implementação de técnicas de extração automática de unidades multipalavras; c) a implementação de uma técnica pertencente a cada possível tipo de abordagem existente; d) a construção de um corpus de referência; e) a avaliação das técnicas de alinhamento implementadas de acordo com critérios de avaliação reportados na literatura.

Os métodos de alinhamento lexical estão separados em três abordagens distintas: empírica, lingüística e híbrida. Os métodos empíricos são aqueles que utilizam informações puramente estatísticas, tais como frequência e co-ocorrência, para realizar o alinhamento. Eles são independentes do par de línguas envolvido, pois não utilizam qualquer tipo de informação lingüística. Os métodos lingüísticos são aqueles que utilizam especificamente recursos lingüísticos das línguas utilizadas nos corpú, tais como léxicos, glossários e listas de palavras âncoras, sendo portanto dependentes das línguas envolvidas no processo. Finalmente, os métodos híbridos são aqueles que unem as abordagens empírica e lingüística, utilizando informações estatísticas e lingüísticas combinadas. Neste trabalho deverá ser implementada e avaliada uma técnica pertencente a cada uma dessas abordagens.

O corpú utilizado neste trabalho será o corpú *PE*: um corpú formado por resumos e *abstracts* de teses, artigos, dissertações e monografias de qualificação, da área de Ciência da Computação, desenvolvidos no ICMC-USP-SC (Martins, Caseli & Nunes, 2001). Esses trabalhos que compõem o corpú pertencem a subdomínios variados da área de Computação, como: banco de dados, computação de alto desempenho, computação gráfica e processamento de imagens, engenharia de software, hipermídia, inteligência computacional, matemática computacional, sistemas digitais, sistemas distribuídos e programação concorrente. Esse corpú é um dos produtos do projeto PESA (Portuguese-English Sentence Alignment), desenvolvido no NILC¹, que visa investigar, implementar e avaliar diversos métodos de alinhamento sentencial de textos paralelos. Os textos pertencentes ao corpú *PE* utilizados neste trabalho passaram por uma fase, na qual os textos foram pré-editados por um tradutor humano, para remoção dos ruídos derivados da tradução (ambigüidades, equívocos e erros). A escolha do corpú foi feita devido à disponibilidade de material digitalizado. O corpú *PE* foi (corretamente) alinhado manualmente com auxílio da ferramenta TagAlign (Caseli & Nunes, 2002), e foi usado como corpú de referência, isto é, os resultados dos algoritmos implementados foram comparados com ele.

Como resultados deste trabalho destacam-se a implementação, avaliação e análise dos principais métodos de alinhamento lexical aplicados a bitextos envolvendo o português brasileiro.

¹ Outras informações sobre o PESA podem ser obtidas em <http://www.nilc.icmc.sc.usp.br/projects/pesa.htm>

1.2. Organização

Este trabalho está organizado da seguinte forma. No capítulo 2 são mostrados os diferentes métodos existentes para o alinhamento lexical de textos paralelos, são detalhadas também as técnicas selecionadas para implementação e as medidas que são utilizadas para avaliar os resultados obtidos por tais técnicas.

No capítulo 3 são detalhadas as estratégias adotadas para o desenvolvimento do projeto de alinhamento lexical português-inglês. Além disso, são listados todos os recursos lingüísticos necessários para a implementação dos métodos de alinhamento.

No capítulo 4 é mostrada a definição de unidades multipalavras, sendo apresentadas duas técnicas para realizar o processo de extração automática de unidades multipalavras, o resultado deste processo de extração e uma avaliação do processo.

Em seguida, no capítulo 5, é detalhado como foram implementados e o funcionamento dos alinhadores lexicais. É também apresentada a avaliação de desempenho dos alinhadores, bem como são analisados os resultados obtidos.

No capítulo 6 são mostradas as conclusões, apresentando os pontos mais importantes do desenvolvimento de cada tarefa e a análise final dos resultados obtidos, além de sugestões de futuros trabalhos.

Este trabalho contém ainda dois apêndices. O Apêndice 1 apresenta alguns exemplos de unidades multipalavras que foram obtidas durante a tarefa de extração automática. O Apêndice 2 mostra alguns exemplos de alinhamentos corretos, propostos pelos métodos de alinhamento lexical investigados neste trabalho.

Capítulo 2

Alinhamento Lexical de Textos Paralelos

No final da década de 1950, iniciaram-se as primeiras tentativas de se utilizar textos paralelos na tradução automática. Contudo, a utilização destes textos paralelos ainda era bastante limitada devido a dificuldades como a baixa capacidade de armazenamento de dados e a escassez de material digitalizado.

Assim, foi somente na década de 80 que o alinhamento de textos paralelos começou realmente a se desenvolver. Uma das alavancas para esse desenvolvimento foi a publicação do método de Kay e Röscheisen em 1987, apresentado no Capítulo 1. Esse método realiza o alinhamento de pontos correspondentes entre uma língua e outra, não levando em consideração informações específicas a respeito das línguas envolvidas no processo.

Historicamente, os textos paralelos mais antigos foram encontrados no início do século 19 na cidade de Rosetta, Egito. A pedra de Rosetta, como é conhecida, contém inscrições paralelas em três diferentes sistemas de escrita: Grego e duas formas de egípcio antigo (demótica e hieróglifos). Um linguísta e egiptologista chamado Jean-Francois Champollion lançou a hipótese de que aquelas inscrições tratavam-se de textos paralelos e após vários anos de estudo e pesquisa, ele foi capaz de decifrar as inscrições em hieróglifos. (Smadja, McKeown & Hatzivassiloglou, 1996).

Para realizar o alinhamento entre os textos paralelos podem-se considerar diferentes aspectos, tais como informações estatísticas ou informações lingüísticas das línguas envolvidas no processo de alinhamento. De modo geral, tal processo baseia-se na determinação dos pontos com base nos quais o alinhamento poderá ser feito, e na filtragem desses pontos, isto é, na eliminação daqueles que apresentam pouca probabilidade de representar, no texto alvo, a tradução correspondente do texto fonte.

De forma matemática, o alinhamento pode ser compreendido como descrito a seguir.

Dados dois textos, A e B, tomados como conjuntos de unidades lingüísticas, que no alinhamento lexical referem-se às palavras encontradas nos textos, têm-se:

$$A = \{a_1, a_2, \dots, a_m\} \text{ e } B = \{b_1, b_2, \dots, b_n\}$$

em que os símbolos a_i e b_j , para $i = 1, 2, \dots, m$ e $j = 1, 2, \dots, n$, representam as m palavras do texto A e as n palavras do texto B. Dessa forma, o alinhamento entre eles é um alinhamento binário X_{AB} definido como um subconjunto de $A \times B$

$$X_{AB} = \{(a_1, b_1), (a_2, b_2), (a_2, b_3), \dots\}$$

A interpretação de X_{AB} é a seguinte: $(a_i, b_j) \in X_{AB}$, se e somente se, existir alguma equivalência de tradução, total ou parcial, entre a_i e b_j (Simard, 2000).

2.1. Métodos de Alinhamento Lexical

O alinhamento lexical pode ser dividido em duas fases, ao menos teoricamente (Verónis, 2000): (1) identificar as palavras nas duas metades do bitexto, e (2) estabelecer a correspondência entre as palavras identificadas. Na prática, a modularização dessas tarefas não é tão simples porque unidades na língua alvo dependem da língua fonte, por exemplo, a expressão francesa “*demand de brevet*” deve ser considerada como uma expressão única quando alinhada com o alemão “*Patentanmeldung*”, mas quando alinhada com o italiano “*domanda di brevetto*”, as palavras devem ser alinhadas uma a uma.

Métodos estatísticos foram propostos para realização desse processo (Lafon, 1984; Church & Hanks, 1990; Smadja & McKeown, 1990 apud Verónis 2000). A vantagem da abordagem estatística é a robustez, pois pode-se eliminar a dependência de contexto das palavras. Porém, esses métodos muitas vezes apresentam baixos índices de desempenho devido à não normalidade das unidades léxicas, ou seja, a ordem das palavras nas sentenças varia de acordo com a língua utilizada, tornando difícil escolher quais medidas estatísticas usar.

Como consequência, o uso de conhecimento lingüístico foi estimulado pela abordagem lingüística. Os métodos dessa categoria baseiam-se em expressões regulares ou em gramáticas (Jacquemin, 1991; Bourigault, 1992; Smadja, 1993; Daille, 1994 apud Verónis, 2000). Apesar de essa abordagem ter alcançado algum grau de sucesso, a utilização de conhecimento lingüístico é cara e específica para as línguas envolvidas. Para o alinhamento lexical, especificamente, métodos puramente lingüísticos não são viáveis, pois

seria necessário construir um léxico contendo todas as palavras do texto fonte e suas traduções. O processo de alinhamento consistiria então de uma tradução palavra a palavra, o que, sabemos, não é eficaz, dadas as diferenças das línguas envolvidas. Por exemplo, muitas palavras únicas no inglês são traduzidas por expressões compostas em português, e vice-versa. Desta forma, métodos puramente lingüísticos não são computacionalmente viáveis, pois são muito custosos e dependentes da língua, tornando-se não eficazes (Véronis, 2000).

Assim, os pesquisadores procuram desenvolver melhorias para os métodos estatísticos, mostrando um notável progresso recentemente (Melamed, 1997a; Wu, 1997; Hiemstra, 1998; Gaussier, 1998 apud Verónis, 2000). Segundo Ahrenberg, Andersson & Merkel (2000), a abordagem mais promissora para o alinhamento de palavras é a híbrida, que combina o baixo custo e a independência das línguas envolvidas dos métodos empíricos e o alto grau de desempenho dos métodos lingüísticos.

Diante destes fatores, decidiu-se, para este trabalho, investigar apenas métodos empíricos ou híbridos, deixando de fora eventuais métodos puramente lingüísticos.

2.1.1. Métodos Empíricos

Métodos empíricos são aqueles que utilizam apenas informações estatísticas, tais como frequência, co-ocorrência, distribuição das palavras no texto e técnicas de cognatos (palavras com grafias similares) para realizar o processo de alinhamento. Este tipo de método é considerado independente das línguas envolvidas, pois não utiliza nenhum tipo de informação lingüística específica das mesmas.

O primeiro método de alinhamento de textos paralelos de Kay e Röscheisen (93), citado no capítulo 1, é um exemplo de método empírico.

A seguir, um método mais recente, proposto por Melamed (1997b, 2000) e Melamed, Al-Adhaileh e Kong (2001), e baseado em reconhecimento de padrões, será descrito em detalhes.

Método *SIMR*

O algoritmo utilizado é o *Smooth Injective Map Recognizer (SIMR)*, que é um algoritmo genérico de reconhecimento de padrão, particularmente bem utilizado para

mapear as correspondências existentes em bitextos. Cada um destes bitextos é visto pelo *SIMR* como um espaço de bitexto retangular, no qual o texto alvo e o texto fonte representam os eixos respectivamente como mostra a Figura 1. O canto inferior esquerdo do retângulo corresponde ao início dos dois textos e é chamado de origem do espaço de bitexto. Já o canto superior direito, que representa o final dos textos, é chamado de término. A linha que liga a origem ao término é a diagonal principal, cuja inclinação representa a inclinação do bitexto.

Figura 1 - Espaço do Bitexto (Melamed, 2000)

Cada espaço de bitexto contém *pontos de correspondência verdadeiros* (PCVs), que correspondem à coordenada no espaço de bitexto em que uma palavra na posição p no eixo x e uma palavra na posição q no eixo y são traduções uma da outra. Dessa forma, o *SIMR* busca produzir mapeamentos de bitexto que se aproximem ao máximo dos PCVs.

Para construir tais mapeamentos, o *SIMR* procura por grupos de PCVs (chamaremos estes grupos também de cadeias) com arranjos mais ou menos linear no espaço de bitexto, executando alternadamente a fase de geração de pontos e a fase de reconhecimento de cadeias. A busca é realizada em uma pequena região retangular do espaço do bitexto denominada retângulo de busca. Este retângulo de busca é usado para limitar a busca por correspondência nos segmentos do bitexto, sendo que o *SIMR* procura por todas as correspondências possíveis dentro deste retângulo de busca, gerando as cadeias ou grupos de Pontos de Correspondência Verdadeiros. Um bom lugar para iniciar tal busca é o início do bitexto, já que a origem é sempre um PCV (Melamed, 1997b).

Caso nenhuma cadeia seja encontrada, o retângulo de busca é expandido proporcionalmente e o ciclo de geração-reconhecimento é repetido até que ao menos uma cadeia seja encontrada. Por outro lado, se mais de uma cadeia é encontrada no mesmo ciclo, o *SIMR* aceita a cadeia cujos pontos são menos dispersos em torno de sua linha de mínimos quadrados (pontos com arranjo mais ou menos linear). Cada vez que uma cadeia é aceita, uma nova região do espaço de bitexto é selecionada para a busca pela próxima cadeia. Assim, a idéia geral do processo é que o primeiro retângulo de busca seja ancorado na origem e os retângulos de busca subsequentes sejam ancorados no canto superior direito da cadeia previamente encontrada, como mostra a Figura 2.

Figura 2 - Expansão do retângulo de busca (Melamed, 2000).

Na fase de geração, o *SIMR* gera todos os pontos de correspondência, dentro de um retângulo de busca, que satisfazem uma determinada condição. Essa condição é uma heurística que permite decidir se um dado par de palavras corresponde a um provável par de traduções mútuas, e pode ser baseada em informações sobre cognatos ou em listas bilíngües de palavras.

Na fase seguinte, fase de reconhecimento, o *SIMR* encontra os grupos de pontos adequados entre os pontos gerados. A maioria dos grupos possui as seguintes propriedades:

1. Linearidade: PCVs tendem a se alinhar linearmente.

2. Inclinação Constante: a inclinação de um grupo de PCVs é raramente muito diferente da inclinação do bitexto.
3. Injetividade: não há dois pontos em um grupo de PCVs que tenham a mesma coordenada x e y.

O *SIMR* explora estas propriedades para decidir quais grupos de pontos do retângulo de busca podem ser grupos de PCVs, levando em conta três parâmetros: tamanho da cadeia, dispersão máxima do ponto e desvio máximo do ângulo. O tamanho da cadeia é simplesmente o número de pontos que ela contém, sendo consideradas apenas as cadeias de tamanho especificado cujos pontos são injetivos. Por sua vez, a linearidade das cadeias é testada medindo-se a raiz da distância média dos pontos da cadeia, a partir da linha de mínimos quadrados dessa cadeia. Caso essa distância ultrapasse o limiar da dispersão máxima do ponto, a cadeia é então rejeitada. No próximo passo, o ângulo da linha de mínimos quadrados de cada cadeia é comparado ao arco-tangente da inclinação do bitexto. Caso a diferença obtida ultrapasse o limiar do desvio máximo do ângulo, a cadeia é rejeitada.

Um outro filtro utilizado pelo *SIMR* é baseado no nível de ambigüidade máxima do ponto. Para cada ponto $p = (x, y)$ deve-se calcular o nível de ambigüidade (NA) através da fórmula:

$$NA(p) = X + Y - 2$$

na qual X é o número de pontos na coluna x e Y é o número de pontos na linha y, dentro do retângulo de busca. Se p é o único ponto em sua linha e coluna, seu nível de ambigüidade é zero. Por outro lado, pontos com nível de ambigüidade muito alto são ignorados. É importante notar que o nível de ambigüidade de um dado ponto pode mudar quando o retângulo de busca se expande ou se move. Como exemplo de nível de ambigüidade pode-se considerar que no texto fonte se encontre a unidade “governo” e o texto alvo contenha as unidades “governor” e “government”. Assim, para o ponto que intercepta a unidade “governo” e a unidade “government” tem-se um nível de ambigüidade igual a 1, pois na coluna x temos 2 pontos (as unidades “governor” e “government” que possuem o mesmo valor de LCSR em relação a unidade “governo”) e na linha y temos apenas 1 ponto.

Para reduzir ainda mais o espaço de busca, o *SIMR* utiliza a propriedade de linearidade para delimitar o tamanho da cadeia, já que cadeias muito pequenas podem apresentar um arranjo linear por coincidência, não sendo, portanto, confiáveis, e cadeias muito grandes possuem pontos muito esparsos para serem considerados lineares. Assim, o tamanho é fixado em k , $6 \leq k \leq 11$, seu valor exato é otimizado juntamente com os outros parâmetros.

A propriedade de inclinação constante também produz uma restrição: devem ser consideradas apenas cadeias paralelas à diagonal principal. Cadeias paralelas são aquelas cujos pontos possuem o mesmo deslocamento em relação à diagonal principal.

Uma avaliação do *SIMR* mostrada em Melamed (2000), usando o *cópus* de bitexto inglês-francês do projeto ARCADE², faz uma comparação do *SIMR* com o *Word_align* de Dagan, Church & Gale (1993). Em 55% dos casos, existe um erro de, no máximo, 2 caracteres, isto é, menos do que uma palavra na saída do *Word_align*; em 73% o erro é de, no máximo, 6 caracteres; e em 84% o erro é de 14 caracteres, equivalente a 3 palavras. Para o *SIMR* os números são: em 93% dos casos, erro máximo de 2 caracteres; em 97% dos casos, erro máximo de 6 caracteres; e em 98% dos casos, erro máximo de 14 caracteres, mostrando bons índices de desempenho para o *SIMR*.

Uma Avaliação do *SIMR* no Alinhamento Português-Inglês

Para o propósito de identificação de falhas e desenvolvimento de melhorias, foi feita uma avaliação do método empírico, *SIMR*. Para tal avaliação foi utilizado para fins de teste um *cópus* livre de erros composto por 11.306 palavras em português e 10.186 palavras em inglês, sendo produzidos como resultado do processo 1.037 alinhamentos. A partir deste conjunto T de 1.037 alinhamentos, foi selecionado aleatoriamente um subconjunto S_T composto por 125 alinhamentos para a realização da avaliação.

² O projeto ARCADE é um dos vários projetos de avaliação de processamento de línguas naturais e de fala, iniciados e financiados por universidades francófonas (AUPELF-UREF). Um dos principais objetivos do projeto ARCADE era contribuir para o desenvolvimento de uma metodologia de avaliação de sistemas de alinhamento de textos paralelos, para melhor compreender as dificuldades dessas tecnologias e, assim, melhorá-las.

O tamanho do subconjunto foi estabelecido a partir de uma medida estatística que permite selecionar um subconjunto considerando-se a porcentagem da ocorrência de falhas no conjunto total, conforme a fórmula a seguir:

$$|S_T| * (\% falha) = 5$$

Como resultados da avaliação feita no subconjunto ST, foram encontrados 96% de alinhamentos corretos. Os erros de alinhamento identificados devem-se a dois fatores principais. Primeiramente, o *SIMR* considera apenas cognatos (palavras com grafia semelhante), assim, por exemplo, o método produz um alinhamento entre o artigo 'a' em português e o artigo 'a' em inglês, quando deveria alinhar com 'the'.

Um outro problema gerado pela utilização deste método empírico é que os alinhamentos são feitos palavra a palavra, produzindo apenas alinhamentos 1:1. Todos os demais tipos de alinhamento (1:N, N:1, N:N com $N > 1$) são ignorados. Assim, o *SIMR* produz um alinhamento incorreto entre as unidades 'frequência:frequently', quando o alinhamento correto seria 'com frequência:frequently'.

Uma alternativa encontrada para tentar minimizar os problemas causados pelo método *SIMR*, procurando gerar alinhamentos mais precisos, é o método híbrido, conhecido como *LWA*, que será apresentado e detalhado nas próximas seções.

2.1.2. Métodos Híbridos

Os métodos híbridos realizam o alinhamento a partir da combinação das abordagens empírica e lingüística, ou seja, eles unem as técnicas estatísticas e os recursos lingüísticos com o objetivo de alcançar altos níveis de desempenho. Segundo Ahrenberg, Andersson & Merkel (2000), a abordagem mais promissora para o alinhamento de palavras é a híbrida, que agrupa em um único método, o baixo custo e a independência das línguas envolvidas decorrente dos métodos empíricos, e o alto grau de desempenho decorrente dos métodos lingüísticos.

O método *LWA*, que será detalhado a seguir, foi proposto por Ahrenberg, Andersson e Merker (1998, 2000, 2002), baseando-se na distribuição e co-ocorrência de palavras nos segmentos dos bitextos e na etiquetagem morfológica.

Método *LWA*

O *LWA* (Linköping Word Aligner) utiliza, para realizar o alinhamento, medidas estatísticas de co-ocorrência combinadas com 3 módulos de conhecimento superficial para: (i) categorização de palavras, (ii) ordem da palavra, e (iii) reconhecimento de frase.

A idéia básica utilizada pelo *LWA* é que unidades correspondentes em um bitexto têm uma probabilidade maior de co-ocorrência em segmentos do bitexto que unidades não correspondentes. Sendo assim, calcula-se um determinado escore de associação para cada unidade considerando-se co-ocorrências em segmentos e ocorrências gerais no bitexto. Uma das medidas que pode ser utilizada para o cálculo deste escore de associação é o *T-score*, que é obtido pela fórmula:

$$T - score = \frac{prob(V_F, V_A) - prob(V_F)prob(V_A)}{\sqrt{\frac{1}{N} prob(V_F, V_A)}}$$

onde,

$$prob(V_F, V_A) = \frac{occur(V_F, V_A)}{N}$$

ou seja, $prob(V_F, V_A)$ é a probabilidade do par de *tokens* (V_F - token do texto fonte, V_A - token do texto alvo) ocorrer nos seguimentos do bitexto (co-ocorrência), em relação a quantidade total de *tokens* presentes no corpus.

$$prob(V_F) = \frac{occur(V_F)}{N}$$

De forma semelhante, $prob(V_A)$ é a probabilidade do *token* do texto fonte (V_F) ocorrer isoladamente nos seguimentos do bitexto, em relação a quantidade total de *tokens* presentes no corpus.

$$prob(V_A) = \frac{occur(V_A)}{N}$$

sendo que $prob(V_A)$ é a probabilidade do *token* do texto alvo (V_A) ocorrer isoladamente nos seguimentos do bitexto, em relação a quantidade total de *tokens* presentes no corpus.

A unidade candidata do texto alvo obtida através do maior escore de associação necessita ainda satisfazer duas condições: (a) que o escore seja maior que um dado limiar, e (b) que a frequência global do par seja suficientemente alta.

O processo de geração de pares de tradução é iterativo, sendo repetido até que nenhum outro par possa ser gerado, ou até que um certo número de operações já tenha sido completado. A cada iteração o bitexto é reduzido a partir da remoção dos pares que foram encontrados.

O algoritmo básico é melhorado pela adição de 3 módulos principais que dão um direcionamento lingüístico ao processo de alinhamento, podendo ser invocados para melhorar a performance do sistema. O primeiro módulo é o de unidades de classe fechada, que separa as unidades (*tokens*) em categorias. As unidades dos textos são separadas em duas classes: unidades de classe fechada (pronomes, artigos, preposições e conjunções) e de classe aberta (substantivos, adjetivos, verbos, advérbios, etc.), estabelecendo-se a restrição de que as unidades só podem ser alinhadas com unidades da mesma classe. As unidades de classe fechada precisam ser listadas explicitamente e classificadas com o correspondente *part of speech* (POS) para que se possa excluir pares candidatos de categorias não correspondentes.

Um algoritmo para tratar as unidades de classe aberta e de classe fechada é apresentado em (Ahrenberg, Andersson & Merkel, 2000). A cada iteração, as operações executadas são:

- (i) Para cada unidade (unidade simples ou unidade multipalavra) de classe aberta no texto fonte, ordenam-se as unidades de classe aberta nas sentenças correspondentes do texto alvo com base em seus escores de associação de unidades. Esta operação resulta em uma lista de pares de tradução envolvendo unidades de classe aberta. Para exemplificar esta operação, considere-se uma sentença P no texto fonte formada pelas unidades de classe aberta, $P = \{p_1, p_2, p_3, p_4\}$, e uma sentença E do texto alvo composta por unidades de classe aberta, $E = \{e_1, e_2, e_3, e_4\}$, utiliza-se o *T-score* para calcular os escores de associação entre as unidades do texto fonte e as

unidades do texto alvo. Desta forma, o *LWA* gera uma lista com valores de associação crescentes como mostrada na Tabela 1.

Tabela 1 - Lista de pares de tradução.

Unidade do Texto Fonte	Candidatas do Texto Alvo
p_1	e_1, e_2
p_2	e_2, e_1, e_3
p_3	e_4, e_2, e_3
p_4	e_3, e_2, e_4

- (ii) A operação anterior é repetida, desta vez para unidades de classe fechada. A diferença é que apenas unidades da mesma categoria são consideradas.
- (iii) Unidades de classe aberta que constituem uma sentença geram pares de tradução com as unidades de classe aberta de suas sentenças correspondentes.
- (iv) Quando todas as unidades do texto fonte tiverem sido testadas desta maneira, um certo número de pares de tradução é obtido e removido do bitexto. Têm-se então menos pares de candidatos a serem considerados, reduzindo as frequências e modificando algumas correspondências. O bitexto reduzido é usado como entrada para a próxima iteração.

O diagrama da Figura 3 resume as atividades realizadas por este algoritmo e os resultados obtidos.

Figura 3 - Tratamento de unidades de Classe Aberta e de Classe Fechada.

O segundo módulo a ser considerado é o de unidades multipalavras (UMPs). Unidades multipalavras são grupos de palavras que expressam idéias e conceitos que não podem ser explicados ou definidos por meio de uma única palavra. Incluem-se aí, entre outros, as locuções (por exemplo, “guerra fria”, “levar adiante”) e verbos preposicionados (por exemplo, “turn on”). As UMPs representam um grande desafio para o alinhamento lexical e, no caso de métodos puramente empíricos, elas são a maior causa de insucesso. Daí a necessidade deste módulo, dependente das línguas envolvidas, pelo paradigma híbrido. As unidades utilizadas pelo módulo de UMPs são geradas numa fase de pré-processamento e armazenadas numa tabela de unidades multipalavras. Os T-score para pares candidatos envolvendo unidades multipalavras são calculados semelhantemente aos de palavras simples. Quando a informação utilizada para calcular os escores de associação estiver na forma de pesos de posição, considera-se o peso de posição da primeira palavra.

O último dos três módulos é o de peso de posição, que assume que unidades candidatas do texto fonte possuem tradução na mesma posição relativa à própria posição. O algoritmo pode utilizar um entre dois parâmetros para implementar essa suposição:

- (i) Janela de correspondência: limita a busca no segmento do texto alvo, por exemplo, são examinadas, no texto alvo, no máximo n unidades antes e n unidades depois da posição da unidade no texto fonte.
- (ii) Sequência de pesos de posição: pesos, fornecidos pelo usuário, são distribuídos para as unidades do texto alvo de acordo com suas posições em

relação às unidades dadas do texto fonte. Um escore é calculado como a soma de pesos de posição para as instâncias de um determinado par, podendo ser utilizado no cálculo do *T-score* no lugar do escore de co-ocorrência.

Além desses módulos, alguns parâmetros globais podem ser especificados para auxiliar no processo de alinhamento. O teste de cognato é usado para ajudar na escolha entre vários candidatos que têm o mesmo valor de escore de associação de palavras. A lista de traduções candidatas, para uma determinada palavra do texto fonte, é inspecionada em ordem decrescente e aquelas que satisfazem uma função de cognato são deslocadas para o topo da lista. Essa função de cognato pode ser o *Longest Common Subsequence Ratio* (*LCSR*) ou um teste mais simples empregado por (Simard, Foster & Isabelle, 1992 apud Ahrenberg, Andersson & Merkel, 2000). Em (Melamed, 1999) o *LCSR* de dois *tokens* é definido como a razão do comprimento de sua maior subsequência comum (*LCS*), não necessariamente uma subsequência contígua, e do tamanho do *token* mais longo. O cálculo do *LCSR* é dado pela seguinte fórmula:

$$LCSR(A, B) = \frac{length[LCS(A, B)]}{\max[length(A), length(B)]}$$

Assim, se forem consideradas as palavras ‘governo’ (7 caracteres) e ‘government’ (10 caracteres), verifica-se que possuem 6 caracteres na mesma ordem. Desta forma, $LCSR(\text{governo}, \text{government}) = 6/10$. Um outro exemplo, considerando as palavras ‘morfológico’ (11 caracteres) e ‘morphological’ (13 caracteres), o *LCSR* obtido para estas duas palavras é $8/13$.

O teste de palavra única realiza uma busca no bitexto à procura de palavras que não estejam ligadas a uma tradução. Havendo palavras únicas em ambos os lados do bitexto, elas serão ligadas como traduções mútuas. Caso deseje-se aumentar a precisão desse teste, pode-se combiná-lo com o teste de cognato.

O teste de duplicata verifica a ocorrência de pares de sentenças duplicadas. Em traduções com alto grau de recorrência os *tokens* dos pares de sentenças são tratados como uma instância única.

O teste de alternância consiste em tentar determinar o maior número de alinhamentos das palavras do texto alvo com as do texto fonte, uma vez que todos os alinhamentos possíveis das palavras do texto fonte com o alvo já foram encontradas.

Outros parâmetros que também precisam ser considerados e definidos são: o Limiar de Frequência, que determina a menor frequência usada no cálculo de associação de palavras; o Limiar do Escore de Associação de Palavras, que determina o menor limiar usado para o escore de associação; e o Número de Iterações, que determina o número de vezes que o processo deve ser executado.

O *LWA* apresentou uma precisão de 83,9 e 96,7% em uma avaliação realizada com dois corpuses de gêneros diferentes escritos em inglês e sueco (Ahrenberg et al., 2000). Em outra avaliação envolvendo o *LWA*, no projeto ARCADE, com outros quatro sistemas de alinhamento de textos paralelos no nível de palavras, CEA, LILA, *LWA*, RALI e XEROX, para um corpus de inglês e francês (contendo aproximadamente 1,1 milhão de palavras), a melhor precisão obtida foi de 75%, do sistema XEROX (Verónis & Langlais, 2000). Sendo o francês mais próximo ao português do que o sueco, esse limiar de 75% de precisão pode ser considerado como objetivo para o trabalho aqui apresentado. Embora este número não seja muito expressivo, é importante ressaltar que o alinhamento lexical apresenta dificuldades de tratamento consideradas bastante sofisticadas, mesmo que se aproveite do conhecimento prévio adquirido no alinhamento sentencial.

No Capítulo 5 será exemplificado o funcionamento dos métodos de alinhamento lexical aqui descritos.

2.2 Técnicas de Avaliação

É necessário avaliar o desempenho dos métodos de alinhamento a partir dos resultados obtidos no processo de alinhamento. Para a avaliação dos resultados dos alinhamentos é necessária a utilização de algumas métricas padrões, bem como a utilização de um corpus de referência previamente construído, por um tradutor humano, manualmente ou com auxílio de ferramentas automáticas. Tal corpus servirá de parâmetro para a comparação com os resultados obtidos.

As métricas padrões utilizadas para medir a performance de sistemas alinhadores são: revocação, precisão e medida-F³. Revocação determina a quantidade de alinhamentos corretos encontrados dentre todos os possíveis. Precisão determina a quantidade de alinhamentos corretos dentre todos os que foram encontrados. Medida-F é uma medida de frequência que combina revocação e precisão numa única métrica.

Tais métricas podem ser definidas da seguinte forma:

$$\text{revocação} = \frac{\text{Número de Alinhamentos Corretos}}{\text{Número de Alinhamentos de Referência}}$$

$$\text{precisão} = \frac{\text{Número de Alinhamentos Corretos}}{\text{Número de Alinhamentos Propostos}}$$

$$\text{medida-F} = 2 \frac{\text{revocação} \times \text{precisão}}{\text{revocação} + \text{precisão}}$$

Assim, revocação corresponde à cobertura do método de alinhamento; quanto mais alinhamentos corretos forem propostos, maior será revocação, sendo $0 \leq \text{revocação} \leq 1$. Já precisão corresponde à capacidade de encontrar alinhamentos corretos; quanto maior a precisão maior será a consistência do método, sendo $0 \leq \text{precisão} \leq 1$. Medida-F, conseqüentemente, corresponde à capacidade que o método tem de produzir todos os alinhamentos corretos possíveis, sendo $0 \leq \text{medida-F} \leq 1$.

Métodos capazes de gerar todos os alinhamentos identificados no cópulo de referência possuem revocação igual a 1. Métodos nos quais todos os alinhamentos encontrados estão corretos possuem precisão igual a 1. É importante observar que o ideal é que tanto revocação quanto precisão sejam iguais a 1, nesse caso, a medida-F também é igual a 1, garantindo que todos os alinhamentos possíveis são encontrados e que todos eles são alinhamentos corretos.

³ As métricas precisão, revocação e medida-F são usadas neste texto com suas denominações em português, mas elas são geralmente encontradas na literatura como *precision*, *recall* e *F-measure*, respectivamente.

Um problema a ser considerado é que as medidas de revocação e precisão apresentadas acima não são apropriadas quando os alinhamentos não são um para um, ou seja, quando unidades multipalavras devem ser consideradas. Estas medidas precisam ser adaptadas para lidar com alinhamentos parcialmente corretos.

Alinhamentos parcialmente corretos são aqueles para os quais o método estabelece correspondências em que faltam partes, ou seja, se o alinhamento correto seria um alinhamento 1:2 e o método consegue apenas estabelecer um alinhamento 1:1, este será considerado um alinhamento parcialmente correto (desde que este alinhamento 1:1 esteja contido no alinhamento correto).

Para exemplificar o conceito de alinhamento parcialmente correto, será utilizado um trecho de um dos textos do corpus de teste, ilustrado na Figura 4.

```
<text lang=pt id=quali3R><p><s><w>Este</w> <w>trabalho</w>
<w>propõe</w> <w>uma</w> <w>modelagem</w> <w>lingüística</w>
<w>dos</w> <w>itens</w> <w>lexicais</w> <w>do</w>
<w>português</w> <w>do</w> <w>Brasil</w>, <w>uma</w>
<w>modelagem</w> <w>relacional</w> <w>e</w> <w>sua</w>
<w>implementação</w> <w>na</w> <w>forma</w> <w>de</w> <w>uma</w>
<w>Base</w> <w>de</w> <w>Dados</w> <w>Lexicais</w>.</s><s>...</s>
</p></text>

<text lang=en id=quali3A><p><s><w>This</w> <w>dissertation</w>
<w>proposes</w> <w>a</w> <w>linguistic</w> <w>modeling</w>
<w>of</w> <w>lexical</w> <w>items</w> <w>of</w> <w>Brazilian</w>
<w>Portuguese</w>, <w>a</w> <w>relational</w> <w>modeling</w>
<w>and</w> <w>its</w> <w>implementation</w> <w>in</w> <w>the</w>
<w>form</w> <w>of</w> <w>a</w> <w>Lexical</w>
<w>Database</w>.</s><s>...</s></p></text>
```

Figura 4 - Trecho do corpus de teste.

Considerando, na Figura 4, as unidades “Base de Dados”, da sentença em português brasileiro, e “Database”, da sentença em inglês, o alinhamento correto seria um alinhamento 3:1, estabelecendo-se a correspondência entre a unidade multipalavras “Base de Dados” e a unidade simples “Database”. Caso o alinhamento proposto pelo método seja algo do tipo: “Base” corresponde a “Database” (1:1) ou “Dados” corresponde a “Database” (1:1), este alinhamento é considerado um alinhamento parcialmente correto, já que falta uma parte do alinhamento, mas a correspondência proposta faz parte do alinhamento correto.

Uma alternativa para essas medidas é proposta em (Ahrenberg et al., 2000). As novas métricas, propostas para resolver o problema substituindo revocação e precisão, são:

$$Q = \frac{C_{src} + C_{trg}}{\max(S_{src}, G_{src}) + \max(S_{trg}, G_{trg})}$$

sendo:

$$revocação = \frac{\sum Q}{n(I) + n(P) + n(C) + n(M)}$$

$$precisão = \frac{\sum Q}{n(I) + n(P) + n(C)}$$

onde,

C_{src} – número de *overlapping tokens* do texto fonte em alinhamentos (parcialmente) corretos, $C_{src} = 0$ para alinhamentos incorretos.

C_{trg} – número de *overlapping tokens* do texto alvo em alinhamentos (parcialmente) corretos, $C_{trg} = 0$ para alinhamentos incorretos.

S_{src} – número de *tokens* do texto fonte propostos pelo método.

S_{trg} – número de *tokens* do texto alvo propostos pelo método.

G_{src} – número de *tokens* do texto fonte no alinhamento do corpus de referência.

G_{trg} – número de *tokens* do texto alvo no alinhamento do corpus de referência.

$n(I)$ – número de alinhamentos incorretos.

$n(P)$ – número de alinhamentos parcialmente corretos.

$n(C)$ – número de alinhamentos corretos.

$n(M)$ – número de alinhamentos que não são encontrados pelo método.

Pode-se notar, portanto, que a revocação e a precisão consideram agora todos os possíveis alinhamentos, inclusive aqueles que são denominados parcialmente corretos. Revocação fornece a razão entre o número de *tokens* nos alinhamentos propostos que coincidem com os do alinhamento de referência e o número total de alinhamentos propostos, incluindo o número de alinhamentos não encontrados pelo método. De forma semelhante, a precisão fornece a razão entre o número de *tokens* nos alinhamentos propostos que coincidem com os do alinhamento de referência e o número total de alinhamentos propostos, só que nesta medida os alinhamentos não encontrados pelo método não são considerados.

A partir da utilização dessas métricas, a diferença entre o número de alinhamentos do corpus de referência e o número de alinhamentos propostos é minimizada, pois são considerados os alinhamentos parcialmente corretos.

Capítulo 3

Alinhamento Lexical Português-Inglês

O objetivo deste trabalho é investigar, programar e avaliar algumas técnicas de alinhamento lexical de textos paralelos português-inglês.

O par de línguas escolhido para este trabalho, o português brasileiro e o inglês, foi determinado pelos seguintes fatores: há apenas um número muito reduzido de trabalhos de alinhamento envolvendo o português, em especial, o português do Brasil; o inglês assume o papel de língua franca em diversos setores da atividade humana (científico, negócios, Web, entre outros). Um outro ponto importante que motivou tal escolha foi a disponibilidade de material envolvendo as duas línguas, o *corpus PE*.

Nas próximas seções será apresentada a estratégia de escolha e construção dos recursos lingüísticos necessários para a investigação aqui descrita.

3.1 Escolha dos Métodos

Com base nos paradigmas e técnicas de alinhamento lexical de textos paralelos já apresentados no Capítulo 2, foram construídos dois protótipos de alinhadores baseando-se em uma técnica pertencente a cada um dos diferentes paradigmas. A seguir serão listadas as candidatas selecionadas para implementação.

Entre os métodos empíricos, o escolhido para análise e implementação foi o *SIMR* (Melamed 1997b, 2000; Melamed, Al-Adhaileh & Kong, 2001), já detalhado na seção 2.1.1. Esse método baseia-se em reconhecimento de padrão, sendo utilizado para mapear correspondências existentes entre as unidades dos textos fonte e alvo.

A escolha desse método foi motivada pelo fato de que ele foi utilizado como base para implementação no projeto PESA. Assim, pretende-se analisar a contribuição do método *SIMR* quando aplicado ao alinhamento lexical em comparação com a sua contribuição para o alinhamento sentencial.

Na abordagem lingüística, como mencionado na seção 2.1, não foi encontrado nenhum método de alinhamento lexical de textos paralelos. Portanto, esta abordagem não contribuirá com nenhum representante para a implementação dos protótipos.

Por fim, o método escolhido na abordagem híbrida para ser analisado e implementado foi o *LWA*. O método *LWA* (Ahrenberg, Andersson & Merkel, 1998; 2000; 2002) baseia-se na distribuição e co-ocorrência das palavras nos segmentos do texto fonte e do texto alvo, parte empírica do método, e também na utilização de módulos lingüísticos que usam recursos como etiquetagem morfológica e listas de unidades de classe fechadas e listas de unidades multipalavras, parte lingüística.

O *LWA* foi escolhido para implementação, pois apresenta uma precisão satisfatória, como mostrado na seção 2.1.2. e por ser um método bastante referenciado na literatura consultada durante o levantamento bibliográfico.

Um dos pontos importantes que deve ser considerado no alinhamento lexical é que muitas vezes as unidades nas duas metades do texto não são alinhadas uma a uma. Estes são os casos em que ocorrem correspondências entre uma única unidade no texto fonte e unidades multipalavras no texto alvo, ou vice-versa. Considerando o par de línguas proposto para este trabalho, português do Brasil e inglês, se a palavra “maquiagem”, por exemplo, for encontrada no texto fonte ela deve ser alinhada com a unidade multipalavra “make up”. O método *SIMR* não considera as unidades multipalavras no processo de alinhamento, realizando apenas um alinhamento palavra a palavra. Por outro lado, o método *LWA* considera unidades multipalavras presentes nos textos, desde que elas estejam incluídas na lista de unidades multipalavras que será detalhada nas próximas seções.

O diagrama da Figura 5 fornece uma idéia geral da organização dos métodos mostrando as técnicas escolhidas para implementação, separadas de acordo com a abordagem de alinhamento lexical de textos paralelos a que pertencem.

Figura 5 - Técnicas de Alinhamento.

Para que estes métodos possam ser implementados, é necessário que alguns recursos lingüísticos sejam previamente construídos. Tais recursos, bem como o seu processo de construção, serão mostrados a seguir.

3.2 Construção de Recursos Lingüísticos

A implementação dos métodos selecionados no capítulo anterior depende da construção de alguns recursos lingüísticos, tais como: os córpis de teste, os córpis de referência, o córpis etiquetado morfológicamente, a lista de unidades de classe fechada e a lista de unidades multipalavras.

Os córpis de teste são constituídos dos textos que são usados como entrada para os métodos de alinhamento. Já os córpis de referência são formados por textos alinhados, manualmente ou com auxílio de uma ferramenta automática, por um tradutor humano. Estes córpis são usados como parâmetro na comparação com os textos alinhados retornados pelos métodos de alinhamento. É importante enfatizar que o conteúdo dos córpis de teste e dos córpis de referência é o mesmo, sendo diferenciados apenas pelo fato de que os córpis de referência possuem algumas marcações que indicam o alinhamento das palavras.

Os córpis, tanto de teste quanto de referência, são compostos por 65 pares de texto português-inglês (Martins, Caseli & Nunes, 2001). Os córpis utilizados são formados pelos 65 pares de textos com correções feitas por um tradutor humano, com a finalidade de extrair dos textos ambigüidades, equívocos e erros gramaticais e/ou de tradução para o inglês.

Serão utilizados dois córpis identificados pelas siglas (Caseli, 2002):

- CPT – Córpis pré-editado de teste, formado pelos 65 pares de textos com alterações para eliminar ambigüidades, equívocos e erros gramaticais e/ou de tradução para o inglês, mas sem marcações de alinhamento lexical.
- CPR – Córpis pré-editado de referência, formado pelos 65 pares de textos com alterações, assim como no CPT, além das marcações que indicam o alinhamento entre as palavras. Essas marcações foram inseridas semi-automaticamente.

Para que esses *cópus* possam ser utilizados pelo o método *LWA* é necessário que eles sejam etiquetados morfologicamente. O *cópus* de teste (CPT) definido acima foi etiquetado, com o auxílio do etiquetador MXPOST⁴, que foi treinado para o português a partir de um *cópus* de 100.000 palavras, obtendo uma precisão geral de 91,99% (Aires & Aluisio, 2001). Em relação ao inglês, o MXPOST apresenta uma precisão de 96,6% (Ratnaparkhi, 1996). Sendo assim, foi gerado um novo *cópus* referenciado pela sigla:

- CPTe – *Cópus* pré-editado de teste etiquetado morfologicamente. Esse *cópus* corresponde ao CPT após o processo de etiquetagem morfológica.

Além desses recursos, foi necessária também a construção de algumas listas: listas de unidades de classe fechada e listas de unidades multipalavras.

As listas de unidades de classe fechada são compostas por artigos, preposições, pronomes e conjunções, devendo haver uma lista para a língua fonte, o português brasileiro, e outra para a língua alvo, o inglês.

As listas de unidades multipalavras armazenam as unidades multipalavras que devem ser consideradas durante o processo de alinhamento. Assim como nas listas acima, deve haver uma lista de unidades multipalavras para a língua fonte e uma para a língua alvo.

3.2.1 *Cópus* de Teste

Os *cópus* de teste utilizados neste trabalho foram construídos a partir do *cópus* PE: formado por resumos e *abstracts* de teses, artigos, dissertações e monografias de qualificação, da área de Ciência da Computação, desenvolvidas no ICMC-USP-SC (Martins, Caseli & Nunes, 2001).

Todos os *cópus* foram codificados de acordo com as especificações da *Text Encoding Initiative* (TEI)⁵. Assim, a linguagem escolhida para tal codificação foi a XML (*Extensible Markup Language*). Tal escolha baseou-se no fato de XML ser uma extensão da SGML (*Standard Generalized Markup Language*), sendo independente das línguas

⁴ Disponível em: <http://nilc.icmc.usp.br/nilc/tools/nilctaggers.html>

⁵ Em: <http://www.tei-c.org>

envolvidas, um dos aspectos mais importantes do alinhamento de textos paralelos (Caseli, 2002).

Os textos paralelos presentes no corpus CPT possuem marcações de fronteiras, sentenças e palavras como mostra a Figura 6.

```
<text lang=pt id=quali3R>p><s><w>Este</w> <w>trabalho</w>
<w>propõe</w> <w>uma</w> <w>modelagem</w> <w>lingüística</w>
<w>dos</w> <w>itens</w> <w>lexicais</w> <w>do</w>
<w>português</w> <w>do</w> <w>Brasil</w>, <w>uma</w>
<w>modelagem</w> <w>relacional</w> <w>e</w> <w>sua</w>
<w>implementação</w> <w>na</w> <w>forma</w> <w>de</w>
<w>uma</w> <w>Base</w> <w>de</w> <w>Dados</w>
<w>Lexicais</w>.</s><s>...</s>/p>/text>

<text lang=en id=quali3A>p><s><w>This</w> <w>dissertation</w>
<w>proposes</w> <w>a</w> <w>linguistic</w> <w>modeling</w>
<w>of</w> <w>lexical</w> <w>items</w> <w>of</w>
<w>Brazilian</w> <w>Portuguese</w>, <w>a</w> <w>relational</w>
<w>modeling</w> <w>and</w> <w>its</w> <w>implementation</w>
<w>in</w> <w>the</w> <w>form</w> <w>of</w> <w>a</w>
<w>Lexical</w> <w>Database</w>.</s><s>...</s>/p>/text>
```

Figura 6 - Exemplo de um fragmento de um par de textos paralelos do corpus de teste.

Na próxima subseção serão detalhados os corpus com os quais serão comparados os resultados obtidos a partir dos métodos de alinhamento, quando os corpus de teste são fornecidos como entrada.

3.2.2 Corpus de Referência

O corpus alinhado em nível de palavras, corpus de referência, foi construído semi – automaticamente (com o auxílio da ferramenta TagAlign). Este corpus, gerado a partir do corpus de teste, servirá como base para a comparação dos resultados retornados pelos métodos de alinhamento lexical que serão implementados.

Como já mencionado nesta seção, a única diferença entre o corpus de teste e o corpus de referência são as marcações de correspondência entre as palavras que estes últimos possuem.

3.2.3 Listas de Unidades de Classe Fechada

Essas listas são usadas em um dos módulos de conhecimento linguístico do método escolhido como representante da abordagem híbrida. Nas listas de unidades de classe fechada devem estar inseridos os pronomes, os artigos, as conjunções e as preposições. Deve haver uma lista de unidades de classe fechada para cada uma das línguas envolvidas no processo de alinhamento, ou seja, no caso deste trabalho, foi construída uma lista para o português do Brasil e outra para o inglês.

A lista de unidades de classe fechada para o português brasileiro foi obtida da base de dados lexical Diadorim⁶, do NILC.

Já a lista de unidades de classe fechada para o inglês foi obtida em conjunto com o algoritmo do *SIMR*, fornecido por I. Dan Melamed⁷.

3.2.4 Listas de Unidades Multipalavras

As listas de unidades multipalavras, assim como as listas de unidades de classe fechada, são utilizadas em um dos módulos de conhecimento linguístico do método *LWA* (Ahrenberg, Andersson & Merkel, 1998; 2000; 2002).

É importante enfatizar que, a princípio, são consideradas apenas as unidades multipalavras contidas na lista de unidades multipalavras. Caso haja alguma unidade multipalavra nos segmentos do bitexto que não esteja inserida nesta lista, o alinhamento será feito palavra a palavra, não considerando a unidade multipalavra e gerando algumas vezes alinhamentos incorretos. Assim, por exemplo, se a unidade multipalavra “turn off” for encontrada no texto alvo, texto em inglês, e não estiver incluída na lista de unidades multipalavras, ela será alinhada palavra a palavra com as unidades do texto fonte. Será necessária a construção de duas listas de unidades multipalavras, uma para o português brasileiro e outra para o inglês.

Deseja-se, neste trabalho, extrair automaticamente as unidades multipalavras presentes no corpus para verificar qual a alteração de desempenho dos métodos alinhadores provocada pela utilização do processo de extração automática. Para isto são utilizadas duas técnicas de extração automática de unidades multipalavras: a *Esperança Mútua* e o Pacote

⁶ Maiores informações: <http://nilc.icmc.usp.br/nilc/tools/intermed.htm>

⁷ <http://www.cs.nyu.edu/~melamed/>

NSP. A definição, a implementação e a avaliação de tais técnicas de extração serão apresentadas e detalhadas no Capítulo 4.

Capítulo 4

Extração Automática de Multipalavras

Para que se possa traduzir e analisar textos em uma determinada língua é necessário levar em consideração o fato de que tais textos não são formados apenas por unidades simples (palavras), mas eles possuem também unidades mais complexas: unidades multipalavras – grupos de palavras que quando combinadas possuem um significado único. Diante disto, algumas ferramentas computacionais têm sido desenvolvidas para identificar e extrair unidades multipalavras a partir de um *cópus* eletrônico.

Uma das maiores dificuldades do processo de alinhamento lexical de textos paralelos é o tratamento das unidades multipalavras. Isto se deve ao fato de que os textos paralelos não podem ser somente alinhados palavra a palavra; grupos de palavras (unidade multipalavra), que têm significado como uma unidade única, devem ser tratados como unidades lexicais para efeito de alinhamento.

Dessa forma, é necessário encontrar meios de se identificar no *cópus* as unidades que não podem ser alinhadas palavra a palavra, tais como “base de dados”, as quais devem ser consideradas como um grupo de palavras. Uma das maneiras de se tratar tais unidades é utilizando-se listas de unidades multipalavras que são consultadas durante o alinhamento para auxiliar na identificação das unidades multipalavras contidas nos textos a serem alinhados. Essas listas podem ser construídas manualmente ou automaticamente a partir de um *cópus*. Neste trabalho, optamos por utilizar listas pré-construídas de forma automática. Para isso, duas técnicas foram investigadas e aplicadas: a *Esperança Mútua* e o pacote de extração de *n-gramas* chamado *NSP*⁸.

A tarefa de extração automática de unidades multipalavras foi realizada tanto em *cópus* do português quanto em *cópus* do inglês.

Os *cópus* e as técnicas que serão utilizados para a tarefa de extração automática das unidades multipalavras estão descritos nas seções a seguir.

⁸ Disponível em: <http://www.d.umn.edu/~tdeperse/code.html>

4.1. Os Crpus utilizados para a Extrao de Multipalavras

Para a tarefa de extrao de unidades multipalavras necessrio utilizar crpus diferentes daqueles que so usados como entrada para as tcnicas de alinhamento lexical, para no influenciar o desempenho dos prottipos alinhadores. Sendo assim, foram construdos alguns crpus para cada uma das lnguas envolvidas no processo de alinhamento, portugus brasileiro e ingls.

Os crpus construdos para esta tarefa so formados por um crpus especfico e um crpus geral. O crpus especfico constitudo de textos tcnicos da rea de Cincia da Computao. A razo da utilizao deste crpus a necessidade de se obter unidades multipalavras especficas deste domnio de conhecimento, uma vez que os textos a serem alinhados pertencem a essa rea especfica. J o crpus geral composto por textos jornalsticos para que se possam obter as unidades multipalavras mais comuns possveis (presentes em textos de qualquer domnio). Os crpus construdos foram codificados analogamente aos crpus utilizados no processo de alinhamento, possuindo marcaes de incio e fim de texto, sentena e palavra. Alm disso, tanto os crpus do portugus brasileiro quanto os crpus do ingls foram etiquetados com um etiquetador do tipo MXPOST para que pudessem ser analisados de acordo com suas classes gramaticais, ajudando a eliminar possveis composies de multipalavras incoerentes.

Para o portugus brasileiro, os crpus construdos esto definidos da seguinte forma:

1. Crpus EP – Crpus Especfico do portugus: Composto por 82 arquivos contendo introdues de artigos, monografias, dissertaes na rea de Cincia da Computao desenvolvidos no ICMC-USP-SO CARLOS. A quantidade total de ocorrncia de palavras neste crpus de 809.708 (sendo considerados tambm nmeros).
2. Crpus GP – Crpus Geral do portugus: Composto por 93 arquivos contendo textos jornalsticos obtidos do Jornal do Brasil e da Folha de So Paulo, extrados do crpus NILC⁹. A quantidade total de ocorrncia de palavras contidas neste crpus de 967.219 (sendo considerados tambm nmeros).

⁹ Maiores informaes em: <http://nilc.icmc.usp.br/nilc/tools/corpora.htm>

Analogamente, para o inglês, os corpuses construídos estão definidos da seguinte forma:

1. Corpus EI – Corpus Específico do inglês: Composto por 100 arquivos contendo textos da área de Ciência da Computação obtidos do *ACM Journals*¹⁰. A quantidade total de palavras no corpus é 704.915 (sendo considerados também números).
2. Corpus EG – Corpus Geral do inglês: Composto por 58 arquivos contendo textos jornalísticos obtidos do *New York Times*. A quantidade total de palavras no corpus é 742.212 (sendo considerados também números).

4.2. Técnicas de Extração de Unidades

Multipalavras

Foram utilizados dois extratores distintos para a realização da tarefa de obtenção automática das unidades multipalavras. Um dos extratores foi implementado utilizando-se uma técnica baseada no cálculo da *Esperança Mútua* (Dias & Kaalep, 2002). O outro extrator utilizado foi o pacote *NSP (N-gram Statistic Package)*. A seguir é apresentada uma descrição detalhada de cada um destes extratores de multipalavras.

4.2.1. Esperança Mútua

Esta técnica baseia-se no fato de que unidades multipalavras são agrupamentos de palavras que têm probabilidade elevada de ocorrerem de forma combinada nos segmentos do texto. Dessa forma, a *Esperança Mútua (EM)* é usada para determinar o grau de coesão entre as palavras contidas em um *n-grama*. Para calcular esse grau de coesão entre as palavras, a medida *Esperança Mútua* utiliza uma outra medida denominada *Esperança Normalizada (EN)*.

A *EN* é definida como a esperança de ocorrência de uma palavra em uma dada posição, sendo conhecidas as ocorrências das demais *n-1* palavras e suas respectivas posições. Como exemplo, considere-se a unidade multipalavra “engenharia de software”. O objetivo da *EN* é determinar o grau de esperança desse *trigrama*, levando em consideração a esperança da unidade “software” aparecer depois de “engenharia de”, como também a

¹⁰ Disponível em: <http://www.informatik.uni-trier.de/~ley/db/journals/>

esperança de ocorrência da unidade “de” ligando as unidades “engenharia” e “software” e finalmente a esperança da unidade “engenharia” ocorrer antes de “de software”. Quanto mais coeso um grupo de palavras é, maior será sua esperança normalizada.

A fórmula utilizada para o cálculo da esperança normalizada é descrita pela razão entre a probabilidade de ocorrência de um *n-grama* e a média das probabilidades de ocorrência dos *n-1-gramas* que ele contém, como mostrado a seguir:

$$EN = \frac{prob(n - grama)}{\frac{1}{n} \sum prob(n - 1 - grama)}$$

A probabilidade, neste caso, é determinada pela razão entre a frequência de ocorrência do *n-grama* em questão e o número total de possíveis *n-gramas* que ocorrem no *cópus*.

Uma vez calculada a *EN* para os *n-gramas*, é possível determinar qual deles tem maior chance de ser uma unidade multipalavra calculando-se a *EM* do *n-grama*. De acordo com (Daille, 1995 apud Dias & Kaalep, 2002) se forem considerados dois *n-gramas* com mesma esperança normalizada, o mais provável de ser uma unidade multipalavra será aquele que possuir maior frequência relativa. Assim:

$$EM = prob(n - grama) \times EN(n - grama)$$

Dessa forma, a *Esperança Mútua* entre as palavras que compõem um *n-grama* é baseada na esperança normalizada e na frequência relativa das palavras no *cópus*.

Depois que os valores de *EM* são calculados para todos os *n-gramas* candidatos e também para todos *n-1-gramas* que os compõem, é necessário determinar qual deles é realmente uma unidade multipalavra (o *n-grama* ou um dos *n-1-gramas* que o compõem). Para isso, utiliza-se um algoritmo conhecido como GenLocalMaxs (Dias & Kaalep, 2002).

Esse algoritmo elege as unidades multipalavras a partir do conjunto de *n-gramas* candidatas, baseado em duas suposições: 1) quanto mais coeso um grupo de palavras (*n-grama*) é, maior será o valor de *EM* associado a ele; 2) unidades multipalavras são grupos

de palavras altamente associados. De acordo com essas suposições, é possível deduzir que: um *n-grama* é considerado uma unidade multipalavra se o grau de coesão entre as palavras que o compõem é maior ou igual ao grau de coesão de todos os subgrupos (*n-1-gramas*) que ele contém, e é maior que o grau de coesão de todos os supergrupos (*n+1-gramas*) que o contém. Como no caso do exemplo anterior, considerando o *trigrama* “engenharia de software”, ele será considerado uma unidade multipalavra se o seu valor de *EM* for maior que o valor de *EM* para os *bigramas* “engenharia de” e “de software” e também seja maior que o valor de *EM* para qualquer *tetragrama* que o contém.

Modificações no método da Esperança Mútua

Os corpus descritos na seção 4.1 foram submetidos à técnica de extração de unidades multipalavras apresentada e, como resultado final, foram obtidas quatro listas de unidades multipalavras, uma para cada corpus construído.

Numa primeira fase, foi utilizada uma implementação seguindo exatamente a descrição do método apresentada nesta subseção. Os quatro corpus construídos foram fornecidos como entrada para o protótipo implementado. Dessa forma, foram obtidas quatro listas de unidades multipalavras, duas para o português do Brasil e duas para o inglês.

As listas foram analisadas e foi possível observar que muitas unidades incoerentes eram geradas, por exemplo, as unidades “para de”, “a de”, “uma a”, entre outras. Diante deste problema, decidiu-se utilizar, combinado a esta técnica de extração, um filtro capaz de minimizar a ocorrência destas unidades consideradas “sem sentido”. O filtro é o mesmo utilizado na técnica de extração de unidades multipalavras baseada no grau de entropia entre as palavras descrito em (Merkel & Andersson, 2000). Tal filtro, que é aplicado durante a fase de construção da unidade multipalavra, consiste da utilização de três listas de palavras que não podem iniciar, ser parte de, ou finalizar uma unidade multipalavra. Tais listas são definidas como:

- ✓ uma lista de palavras que não podem iniciar uma unidade multipalavra (por exemplo, artigos, preposições, números ou pontuação);
- ✓ uma lista de palavras que não podem fazer parte de uma unidade multipalavra (por exemplo, números ou pontuação); e

- ✓ uma lista de palavras que não podem terminar uma unidade multipalavra (por exemplo, artigos, números ou pontuação).

Um novo protótipo foi implementado, nesta nova fase já com o filtro para eliminar unidades multipalavras “sem sentido”, e quatro novas listas foram obtidas. As novas listas ainda apresentavam unidades incoerentes, mas em um número bem mais reduzido. Os resultados são discutidos na seção 4.3.

4.2.2. N-gram Statistic Package (NSP)

O *NSP* é um pacote desenvolvido por Ted Pedersen que é composto por um conjunto de programas que buscam identificar e analisar os *n-gramas* contidos em um cópulo. Para este processo são utilizados dois programas implementados em Perl, o *count.pl* e o *statistics.pl*.

O programa *count.pl* toma um arquivo texto como entrada, formado pelo cópulo do qual devem ser extraídos os *n-gramas*, e gera uma lista contendo todos os *n-gramas* contidos no arquivo e suas respectivas frequências. O arquivo de saída possui na primeira linha a quantidade total de *n-gramas* encontrados e as linhas subsequentes possuem os *n-gramas* acompanhados de suas respectivas frequências de ocorrência no cópulo, obedecendo ao seguinte padrão:

```
token1<>token2<>...<>tokenN<>freq.n-grama freq.token1 freq.token2 ...freq.tokenN
```

Os *tokens* do *n-grama* são separados e delimitados pelo símbolo “<>”. Assim, o *bigrama* “por exemplo” iria aparecer na saída como “por<>exemplo<>” seguido de sua frequência de ocorrência no cópulo e da frequência dos *unigramas* (“por” e “exemplo”) que o compõem. Da mesma forma, o *trigrama* “engenharia de software” aparecerá como “engenharia<>de<>software<>” também seguido de sua frequência e das frequências dos *n-1-gramas* que o compõem.

Como observado nos exemplos acima, este programa permite realizar a extração de *n-gramas* de tamanho *N*, sendo que *N* é o número de *tokens* que se deseja obter em cada *n-grama*. Por exemplo, para obtenção de *bigramas*, *N* = 2, para *trigramas*, *N* = 3, e assim por diante. Tal tamanho deve ser definido pelo usuário.

Para a delimitação dos *tokens* foi utilizada uma regra de formação baseada na tabela ASCII estendida (formato reconhecido pelo pacote). Tal regra de formação é mostrada na Figura 7.

/[a-zA-Z-]	representa caracteres alfabéticos que podem apresentar hífen
[\\w\\xb0]	representa o “°” (grau)
[\\w\\xc0-\\xc5]	representa a letra maiúscula com as acentuações possíveis
[\\w\\xc7-\\xcf]	representa a “ç”, as letras “e” e “i” com acentuações (maiúsculos)
[\\w\\xd1-\\xd6]	representa o “ñ” e a letra o com acentuações (maiúsculos)
[\\w\\xd9-\\xdc]	representa a letra “u” maiúscula com acentuações
[\\w\\xdf-\\xe5]	representa a letra “ß” e a letra “a” minúscula com acentuações
[\\w\\xe7-\\xef]	representa o “ç”, as letras “e” e “i” com acentuações (minúsculos)
[\\w\\xf1-\\xf6]	representa o “ñ” e a letra o com acentuações (minúsculos)
[\\w\\xf9-\\xfc)]+ /	representa a letra “u” minúscula com acentuações

Figura 7 - Regras de formação de tokens (Teline, Manfrim & Aluísio, 2003).

O outro programa, *statistics.pl*, utiliza como entrada a lista gerada pelo programa anterior e computa um escore para cada um dos *n-gramas* de acordo com uma medida estatística. Entre as medidas possíveis estão o coeficiente de *Dice*, a *log-likelihood*, a informação mútua, entre outras. O *statistic.pl* toma as listas dos *n-gramas* com suas respectivas freqüências e os fornece aos pacotes da medida selecionada para que, com base nestas freqüências, sejam calculados os escores referentes aos *n-gramas* encontrados. Para o caso específico deste trabalho, a medida selecionada para o cálculo dos escores foi a *log-likelihood*. Para a escolha desta medida, foi realizada uma fase de teste para comparar o desempenho das medidas disponíveis - *log-likelihood* e informação mútua - que podem ser utilizadas para extrair tanto *bigramas* quanto *trigramas*, e ambas obtiveram os mesmos resultados. Isso pode ser explicado pelo fato de que a medida informação mútua é uma medida muito próxima da medida *log-likelihood* (essencialmente, somente diferenciada por um fator de escala)¹¹, sendo que a medida *log-likelihood* é boa tanto para *n-gramas* contínuos quanto para esparsos, diferentemente da informação mútua, que tem um bom desempenho em dados contínuos.

A medida *log-likelihood* indica quanto uma determinada ocorrência de um *n-grama* é mais provável de ser uma unidade multipalavra do que uma outra ocorrência. Esta medida leva em consideração a ocorrência dos tokens que formam um *n-grama* e o seu complemento, ou seja, a não ocorrência destes tokens. Sua fórmula é definida como:

¹¹Conforme confirmado pelo autor Ted Pedersen (<http://www.d.umn.edu/~tpederse/nsp.html>)

$$\log\text{-likelihood} = -2 \log \frac{[P_X P_Y P_X P_Y]^{f_T}}{[P_{XY} P_{\overline{XY}}]^{f_X} [P_{\overline{XY}} P_{XY}]^{f_Y}}$$

onde f_X é a frequência de ocorrência de uma palavra X ; f_Y é a frequência de ocorrência de uma palavra Y ; P_X é a probabilidade de ocorrência de uma palavra X ; P_Y é a probabilidade de ocorrência de uma palavra Y . A variável XY representa um *bigrama* formado pelos tokens X e Y . A barra sobreposta indica o complemento de uma variável.

Uma vez que os escores foram calculados segundo a medida *log-likelihood*, o programa *statistic.pl* gera um arquivo de saída, no qual a primeira linha é a quantidade de *n-gramas* gerados e nas linhas subsequentes estão os *n-gramas* ordenados segundo um *rank* que é determinado pelo valor do escore encontrado, seguindo o padrão:

```
token1<>token2<> ... <>tokenN<>rank score freq.n-grama freq.token1token2...
freq.token1tokenN freq.token1 freq.token2 ... freq.tokenN
```

Para estabelecer o *rank* usa-se o seguinte critério: o maior escore recebe *rank* igual a 1, o segundo maior escore recebe igual a 2, o terceiro maior escore recebe *rank* igual a 3 e assim por diante.

Todo este processo é realizado de maneira semelhante para os *trigramas*, de forma que o escore é calculado pela medida *log-likelihood* estendida diretamente para os *trigramas*. A única modificação é que ao invés de serem considerados apenas dois tokens, passam a ser considerados três tokens.

4.3. Resultados obtidos

4.3.1. Listas Originais

A partir dos métodos anteriormente descritos foram obtidas oito listas de unidades multipalavras. Quatro destas listas foram obtidas pela técnica baseada em *Esperança Mútua* aplicada ao corpus EP, ao corpus GP, ao corpus EI e ao corpus GI, respectivamente. De

forma semelhante, as outras quatro listas foram extraídas pelo pacote *NSP* aplicado aos mesmos *cópus*.

A Tabela 2 mostra o contraste entre a quantidade de unidades contidas nas listas obtidas utilizando-se a técnica de *Esperança Mútua* e o pacote *NSP* para realizar o processo de extração de unidades multipalavras. É essencial lembrar que todas as listas geradas, independentemente de qual extrator está sendo utilizado, são dependentes dos *cópus* que estão sendo utilizados, sendo a abrangência das listas tão amplas quanto a abrangência dos *cópus* em questão.

Tabela 2 - Quantidade de unidades das listas geradas pelas técnicas de extração selecionadas.

	Cópus EP	Cópus GP	Cópus EI	Cópus GI
<i>Esperança Mútua</i>	12.197	23.196	8.694	17.226
<i>NSP</i>	90.503	185.650	30.516	73.938

Analisando as listas geradas foi possível observar que muitas das unidades geradas não eram unidades relevantes (um exemplo de uma unidade considerada relevante é o bigrama “guerra fria”, outros exemplos podem ser encontrados no Apêndice 1), ou seja, eram unidades que não deveriam necessariamente ser consideradas como unidades multipalavras, como por exemplo, as unidades “alto desempenho”, “feminino e masculino”, “O desempenho”. Com base neste fato, percebe-se que era necessário realizar uma etapa de eliminação de tais unidades, permitindo que só fizessem parte da lista apenas as unidades que realmente satisfizessem as propriedades de uma multipalavra.

Todo o processo e os critérios estabelecidos para a tarefa de eliminação das unidades consideradas não relevantes estão descritos na subseção a seguir.

4.3.2. Eliminando Unidades Não Relevantes

Em (Schone & Jurafsky, 2001), unidades multipalavras são definidas como uma sequência de palavras vizinhas conectadas cujo significado ou conotação não pode ser obtido a partir do significado ou conotação de seus componentes. Desta forma, é possível estabelecer alguns critérios que devem ser satisfeitos para que um determinado *n-grama* possa ser considerado uma unidade multipalavra. Tais critérios são: a não-composicionalidade, a não-substitutibilidade e a não-modificabilidade.

A não-composicionalidade é a propriedade que garante que o significado de um *n-grama* só pode ser obtido a partir do conjunto de palavras que o compõem como um todo, não podendo ser obtido a partir da decomposição de seus componentes. O *bigrama* “alto desempenho” é um exemplo de uma unidade que não satisfaz o critério da não-composicionalidade, uma vez que o seu significado pode ser obtido a partir da composição do significado das palavras que o formam. Em contrapartida, o *bigrama* “guerra fria” é um exemplo de uma unidade que satisfaz o critério da não-composicionalidade, pois o seu significado só pode ser abstraído do grupo de palavras como um todo.

A não-substituibilidade garante que nenhum dos componentes do *n-grama* pode ser substituído por seus sinônimos, pois tal substituição não mais transmite o significado original. Por exemplo, o *n-grama* “compact disk” não necessariamente implica “densily-packed disk”; pode-se tratar de uma mídia musical, nesse caso ele é considerado não-substituível. Por outro lado, um exemplo de unidade que não satisfaz a propriedade de não-substituibilidade é o *bigrama* “preço alto”, pois se a palavra “alto” for trocada por “elevado” (“preço elevado”) no *bigrama* em questão, o significado ainda é mantido.

Por fim, a não-modificabilidade estabelece o critério de que a estrutura de uma unidade multipalavra não pode ser modificada sem que seu significado seja alterado. Considerando mais uma vez a unidade “compact disk”, se mudarmos sua estrutura para “disk that is compact” seu significado não é preservado, satisfazendo o critério da não-modificabilidade. Em contraste, o *bigrama* “garota bonita” não perde seu significado se for modificado para “bonita garota” não obedecendo desta forma o critério da não-modificabilidade.

Além de satisfazer as propriedades acima descritas, foi estabelecida ainda uma condição. Nos *corpus* em português brasileiro seriam consideradas apenas as unidades geradas em português; analogamente, dos *corpus* em inglês seriam consideradas apenas as unidades geradas em inglês. Esta condição foi estabelecida, pois algumas vezes eram geradas *n-gramas* em inglês nas listas geradas a partir dos *corpus* em português, o que não era interessante para o objetivo deste trabalho.

As eliminações foram realizadas manualmente por agentes humanos, numa fase de pós-processamento (depois da geração das listas), levando em consideração todos os critérios pré-estabelecidos acima.

4.3.3. Listas Finais

Obedecendo aos critérios descritos, após o processo de eliminação das unidades consideradas não relevantes foram obtidas oito novas listas, equivalentes às listas originalmente obtidas pelos corpúsculos de entrada (corpúsculo EP, corpúsculo GP, corpúsculo EI e corpúsculo GI) submetidos às técnicas de extração (*Esperança Mútua* e *NSP*), algumas das unidades contidas nestas listas são mostradas no Apêndice 1.

A Tabela 3 mostra como são compostas, em unidades, as listas finais obtidas após o processo de eliminação. Pode-se notar pela Tabela 3 que a quantidade de unidades obtidas nas listas geradas pelo pacote *NSP* foi aproximadamente duas vezes maior que a quantidade obtida pela técnica baseada em *Esperança Mútua*.

Tabela 3 - Quantidade de unidades das listas obtidas após o processo de eliminação.

	Corpúsculo EP	Corpúsculo GP	Corpúsculo EI	Corpúsculo GI
<i>Esperança Mútua</i>	153	1.115	140	348
<i>NSP</i>	260	1.556	262	645

4.4. Avaliação das Listas Obtidas

Utilizando-se os extratores automáticos de unidades multipalavras foram obtidas no total dezesseis listas de unidades multipalavras: oito listas originais, obtidas a partir da aplicação das técnicas de extração aplicadas aos corpúsculos construídos, e oito listas processadas, obtidas a partir das listas originais após um processo de eliminação de unidades não relevantes. Com base nos resultados obtidos foi então feita uma comparação entre as listas geradas seguindo alguns critérios pré-determinados para que fosse possível avaliar as técnicas de extração utilizadas.

Os critérios estabelecidos para a comparação das listas foram: 1) a quantidade de unidades geradas; 2) a quantidade de unidades eliminadas (somente aplicado às listas originais); 3) a quantidade de unidades coincidentes e 4) quantidade de unidades contidas no conjunto diferença das listas geradas.

A avaliação consistiu de duas etapas, a primeira delas foi feita utilizando-se as listas originais e a outra utilizando-se as listas processadas. Todo o processo de avaliação é descrito de forma detalhada nas próximas subseções.

4.4.1. Etapa 1 - Antes da Eliminação (Listas Originais)

A primeira etapa foi realizada utilizando-se as listas originais exatamente da maneira como foram geradas pelas técnicas de extração, sem nenhum processamento ou eliminação de unidades não relevantes. Para possibilitar uma melhor visualização dos resultados obtidos, foram montados, durante o processo de avaliação, tabelas e gráficos comparativos mostrando o desempenho de cada um dos extratores de multipalavras utilizados neste trabalho.

A Tabela 2, apresentada na subseção 4.3.1, mostra a comparação baseada no primeiro critério de avaliação estabelecido, que contrasta a quantidade de unidades geradas pela técnica baseada em *Esperança Mútua* e pelo pacote *NSP*. O gráfico obtido com base nos valores contidos na Tabela 2 pode ser visualizado na Figura 8 mostrada abaixo.

Figura 8 - Comparação entre a quantidade de unidades geradas pelas técnicas de extração.

Observando-se a tabela (Tabela 2) e a figura (Figura 8) obtidas nesta primeira comparação, pode-se perceber que o pacote *NSP* sempre gera mais unidades (aproximadamente cinco vezes mais) do que a técnica baseada em *Esperança Mútua*. Isto

pode ser explicado pelo fato de que o pacote *NSP* gera todas as possíveis combinações de *n-gramas* existentes no *cópus* em questão, enquanto que a *Esperança Mútua* só gera os *n-gramas* que têm um elevado grau de associação entre as palavras que o compõem.

O segundo parâmetro consiste em estabelecer uma comparação entre a quantidade de unidades eliminadas em cada uma das listas geradas pelos extratores automáticos. A Tabela 4 foi montada de acordo com os resultados obtidos depois do processo de exclusão das unidades que não satisfaziam os critérios estabelecidos na subseção 4.3.2. A Figura 9 apresenta o gráfico gerado a partir dos dados da Tabela 4.

Tabela 4 - Quantidade de unidades eliminadas das listas originais.

	Cópus EP	Cópus GP	Cópus EI	Cópus GI
<i>Esperança Mútua</i>	12.044	22.068	8.554	16.878
<i>NSP</i>	90.243	184.094	30.254	73.293

Figura 9 - Comparação entre a quantidade de unidades eliminadas das listas geradas pelas técnicas de extração.

Analisando-se a Tabela 4 e a Figura 9 é possível observar que o número de unidades que foram eliminadas foi bastante elevado tanto para a técnica de extração baseada em *Esperança Mútua* quanto para o pacote *NSP*, fazendo com que as Tabela 2 e Tabela 4 e Figura 8 e Figura 9 ficassem bastante semelhantes. Isto comprova o fato de que a maioria dos extratores de multipalavras existentes hoje necessita de melhoramentos à medida que

não geram apenas unidades multipalavras relevantes (para as quais necessita-se considerar um grupo de palavras para obter sua definição).

O próximo critério focaliza as unidades coincidentes nas listas equivalentes geradas pelos dois extratores. A idéia é obter a quantidade de unidades contidas na intersecção entre as listas geradas pela técnica de *Esperança Mútua* e suas equivalentes geradas pelo pacote *NSP*. Os resultados desta operação são apresentados na Tabela 5.

Tabela 5 - Quantidade de unidades coincidentes nas listas originais.

	Cópus EP	Cópus GP	Cópus EI	Cópus GI
<i>Esperança Mútua</i> $\cap$ <i>NSP</i>	5.095	9.288	2.933	5.831

Um aspecto interessante que foi observado, e que vale a pena ser mencionado, é que as listas provenientes da operação de intersecção continham algumas unidades não relevantes comuns às duas técnicas de extração, apesar de se esperar que a intersecção das listas nos fornecesse uma certa confiança para as unidades multipalavras geradas. Isto significa que, apesar de utilizarem estratégias diferentes para encontrar as unidades multipalavras, estas técnicas podem coincidir nas falhas, gerando falsas multipalavras comuns.

Por fim, o quarto e último critério procura encontrar as unidades que estão contidas em uma lista, mas não estão presentes em sua equivalente (lista gerada a partir do mesmo cópus, mas com outra técnica de extração automática). A Tabela 6 mostra as diferenças encontradas nesta comparação e a Figura 10 ilustra esta diferença.

Tabela 6 - Quantidade de unidades da diferença entre as listas geradas pelos métodos extratores.

	Cópus EP	Cópus GP	Cópus EI	Cópus GI
<i>Esperança Mútua</i> – <i>NSP</i>	7.102	13.908	5.761	11.395
<i>NSP</i> – <i>Esperança Mútua</i>	85.408	176.362	27.583	68.111

Figura 10 – Diferença entre as listas geradas pelos métodos extratores.

A primeira linha da Tabela 6 mostra quantas unidades estão contidas nas listas geradas pela técnica de extração baseada na *Esperança Mútua* e não estão incluídas nas listas geradas pelo pacote *NSP*. De forma análoga, a segunda linha fornece o número de unidade que são encontradas nas listas obtidas pelo pacote *NSP*, mas não são encontradas nas listas obtidas pela técnica de *Esperança Mútua*.

Depois de realizada esta etapa de avaliação nas listas originais é mais marcante ainda a necessidade de um processo de eliminação dos *n-gramas* que estão inseridos na lista, mas que não devem ser considerados como unidades multipalavras. Na subseção a seguir serão apresentados os resultados das comparações realizadas com as listas processadas.

4.4.2. Etapa 2 - Depois da Eliminação (Listas Processadas)

A etapa seguinte, etapa 2, foi realizada com as listas processadas, que são as listas obtidas depois da eliminação das unidades que não deveriam ser consideradas como multipalavras (unidades não relevantes). Os critérios utilizados para esta avaliação foram os mesmos utilizados na etapa realizada com as listas originais, com exceção do critério de número 2, pois nestas listas as unidades não relevantes já foram excluídas. Como na etapa de

avaliação anterior, foram construídos tabelas e gráficos comparativos para exibir o desempenho obtido por cada um dos extratores de unidades multipalavras utilizados.

Com base no primeiro critério estabelecido, quantidade de unidades geradas, foi montada a Tabela 3 já mostrada na subseção 4.3.3. A tabela mostra que as listas foram bastante “enxugadas” ficando com um número bem reduzido de unidades em relação às listas originais, sendo eliminados todos os excessos (unidades não relevantes). Como na comparação das listas originais, o pacote *NSP* obteve um desempenho superior (mais unidades geradas) ao da técnica baseada em *Esperança Mútua*. O gráfico da Figura 11 permite visualizar a diferença entre a quantidade de unidades geradas pelos dois extratores para cada um dos *cópus* utilizados como entrada.

Figura 11 - Comparação entre a quantidade de unidades obtidas após o processo de eliminação.

O próximo critério de avaliação, terceiro critério da etapa 1, estabelece uma comparação entre as unidades coincidentes contidas nas listas de cada uma das técnicas de extração. A Tabela 7 mostra os valores obtidos.

Tabela 7 – Quantidade de unidades coincidentes nas listas processadas.

	Corpus EP	Corpus GP	Corpus EI	Corpus GI
<i>Esperança Mútua</i> ∩ <i>NSP</i>	75	718	82	218

Com as listas obtidas nesta comparação foi possível certificar algumas ocorrências de unidades multipalavras, permitindo obter uma lista básica de unidades relevantes.

O último critério explora a propriedade da diferença entre as listas equivalentes gerada pelas diferentes técnicas de extração. A Tabela 8 foi montada com os valores obtidos na comparação deste critério e o gráfico da Figura 12 mostra a relação entre tais valores.

Tabela 8 – Quantidade de unidades no conjunto diferença entre as listas processadas.

	Cópus EP	Cópus GP	Cópus EI	Cópus GI
<i>Esperança Mútua – NSP</i>	76	397	58	130
<i>NSP – Esperança Mútua</i>	185	840	181	429

Figura 12 - Diferença entre as listas obtidas após o processo de eliminação.

A diferença na etapa 2 possui a mesma definição apresentada na avaliação das listas originais (etapa 1).

Avaliando-se todos os resultados apresentados neste Capítulo, em relação ao desempenho dos protótipos utilizados na tarefa de extração, foi possível perceber que o pacote *NSP*, embora utilize uma heurística bem mais simples, obteve melhor desempenho do que o extrator baseado na técnica de *Esperança Mútua*, que produziu uma quantidade aproximadamente duas vezes menor de unidades multipalavras válidas.

No próximo capítulo serão apresentados os alinhadores lexicais utilizados neste trabalho, detalhando-se o seu funcionamento e a avaliação dos resultados gerados por cada um deles.

Capítulo 5

Implementação e Avaliação dos Alinhadores Lexicais

5.1. Implementação dos Alinhadores Lexicais

Como já foi mencionado no Capítulo 3, dois métodos de alinhamento lexical foram selecionados para serem implementados e avaliados. Um deles é representante da abordagem empírica, o *SIMR*, e o outro, representante da abordagem híbrida, o *LWA*.

A implementação de tais alinhadores foi feita utilizando a linguagem de programação PERL. O protótipo do alinhador empírico (*SIMR*) foi desenvolvido para ser executado em plataforma Unix, já o protótipo do alinhador híbrido (*LWA*) foi projetado para ser executado em plataforma Windows.

A seguir, serão descritos alguns detalhes para exemplificar o funcionamento dos alinhadores *SIMR* e *LWA*.

5.1.1. Alinhador *SIMR*

O *SIMR* (descrito no Capítulo 2) baseia-se em reconhecimento de padrões para estabelecer as correspondências entre as unidades lexicais dos textos paralelos (texto fonte e texto alvo) e para isto utiliza a medida *LCSR* (definida na subseção 2.1.2) para identificar cognatos. Um ponto muito importante que deve ser mencionado é que este alinhador só considera no seu processo de alinhamento as palavras simples (não considera unidades multipalavras).

O alinhador vê os bitextos como eixos perpendiculares, sendo que cada palavra no texto fonte corresponde a uma posição no eixo x e cada palavra do texto alvo corresponde a uma posição no eixo y. Na Tabela 9 é mostrado um fragmento de um bitexto considerado pelo *SIMR*.

Tabela 9 – Fragmento de bitexto visto pelo *SIMR*.

BITEXTO	
Texto Fonte	Texto Alvo
0 <EOS>	0 <EOS>
2.5 Este	2.5 This
9.5 trabalho	11.5 dissertation
17.5 propõe	22.5 proposes
23 uma	28 a
30 modelagem	34.5 linguistic
41 linguística	44.5 modeling
49 dos	50.5 of
54 itens	56 lexical
61.5 lexicais	63 items
67.5 do	67.5 of
74 português	74 Brazilian
80.5 do	84.5 Portuguese
85.5 Brasil	91 ,
90 ,	94 a
94 uma	100.5 relational
101 modelagem	110.5 modeling
111.5 relacional	117 and
118 e	121 its
121 sua	130.5 implementation
130 implementação	139.5 in
138.5 na	143 the
143 forma	147.5 form
147.5 de	151.5 of
151 uma	154 a
155.5 Base	159 Lexical
159.5 de	167.5 Database
164 Dados	173 .
171.5 Lexicais	176 <EOS>
177 .	
180 <EOS>	

O objetivo do *SIMR* é encontrar um mapeamento entre as coordenadas do texto fonte e do texto alvo, que mais se aproxime dos Pontos de Correspondência Verdadeiros (PCVs). Para determinar se um par de palavras corresponde a um par de traduções o *SIMR* utiliza a medida *ICSR*, produzindo mapeamentos como os mostrados na Tabela 10.

Tabela 10 – Mapeamentos produzidos pelo *SIMR*.

MAPEAMENTO	
Texto Fonte	Texto Alvo
0	0
30	44.5
41	34.5
54	63
61.5	56
74	84.5
90	91
101	110.5
111.5	100.5
130	130.5
143	147.5
171.5	159
177	173

Um programa em PERL foi construído para transformar as coordenadas nas unidades lexicais correspondentes, produzindo as listas de correspondências (alinhamentos) para as unidades lexicais dos bitextos (outros exemplos de alinhamentos propostos pelo *SIMR* podem ser vistos no Apêndice 2). A Tabela 11 apresenta a lista correspondente ao mapeamento mostrado na Tabela 10.

Tabela 11 – Exemplo de alinhamentos propostos pelo *SIMR*.

Unidades Correspondentes	
Texto Fonte	Texto Alvo
modelagem	Modeling
lingüística	linguistic
itens	items
lexicais	lexical
português	Portuguese
modelagem	modeling
relacional	relational
implementação	implementation
forma	Form
Lexicais	Lexical
resultante	resulting
centralização	centralization
facilitando	facilitating
processo	process
Lingüístico	Linguistic

É interessante mencionar que, antes de produzir os alinhamentos com o cópús *PE*, foi realizada uma fase de otimização com o alinhador *SIMR*, a fim de se determinar o limiar (valor mínimo aceitável) da medida de cognatos *LCSR* bem como o tamanho máximo da

cadeia (grupo de Pontos de Correspondência Verdadeiros) a ser considerado na busca dentro do retângulo de busca. Para esta otimização foram utilizados como entrada dois corpúscos distintos: um manual de PHP (contendo 7.092 palavras em português e 6.887 palavras em inglês) e a Constituição brasileira (contendo 5.936 palavras em português e 6.641 palavras em inglês). Os valores determinados nesta fase são: o limiar do *LCSR* igual a 0,66, com tamanho mínimo das palavras para o cálculo do *SIMR* de 4 caracteres; o tamanho máximo da cadeia igual a 6; nível de ambigüidade máximo da cadeia igual a 0,11; dispersão máxima do ponto igual a 15 e o desvio máximo do ângulo igual 0,65.

5.1.2. Alinhador *LWA*

O *LWA* (também descrito no Capítulo 2) é um alinhador híbrido que utiliza medidas estatísticas de co-ocorrência (frequência da ocorrência de um par de palavras no bitexto) e três módulos de conhecimento lingüístico superficial para determinar as correspondências entre as unidades lexicais (palavras e unidades multipalavras) do bitexto.

Para entender o funcionamento do alinhador, considere-se como entrada o fragmento de bitexto mostrado na Tabela 12.

Tabela 12 – Exemplo de bitexto usado como entrada para o *LWA*.

BITEXTO	
TEXTO FONTE	TEXTO ALVO
<pre> <text lang=pt id=quali3R> <p><s><w>Este_PRON</w> <w>trabalho_N</w> <w>propõe_VERB</w> <w>uma_ART</w> <w>modelagem_N</w> <w>lingüística_ADJ</w> <w>dos_PREP+ART</w> <w>itens_N</w> <w>lexicais_ADJ</w> <w>do_PREP+ART</w> <w>português_N</w> <w>do_PREP+ART</w> <w>Brasil_NP</w> , <w>uma_ART</w> <w>modelagem_N</w> <w>relacional_ADJ</w> <w>e_CONJ</w> <w>sua_PRON</w> <w>implementação_N</w> <w>na_PREP+ART</w> <w>forma_N</w> <w>de_PREP</w> <w>uma_ART</w> <w>Base_N de_PREP Dados_NP</w> <w>Lexicais_NP</w> .</s> <s>...</s> </p> </text> </pre>	<pre> <text lang=en id=quali3A> <p><s><w>This_DT</w> <w>dissertation_NN</w> <w>proposes_VBZ</w> <w>a_DT</w> <w>linguistic_JJ</w> <w>modeling_NN</w> <w>of_IN</w> <w>lexical_JJ</w> <w>items_NNS</w> <w>of_IN</w> <w>Brazilian_JJ</w> <w>Portuguese_NN</w> , <w>a_DT</w> <w>relational_JJ</w> <w>modeling_NN</w> <w>and_CC</w> <w>its_PRP\$</w> <w>implementation_NN</w> <w>in_IN the_DT</w> <w>form_NN</w> <w>of_IN</w> <w>a_DT</w> <w>Lexical_JJ</w> <w>Database_NNP</w> .</s> <s>...</s> </p> </text> </pre>

Pode-se observar que, tanto no texto fonte, quanto no texto alvo, estão marcadas não só as unidades simples como também as unidades multipalavras. Tais unidades são obtidas numa fase de pré-processamento através de extração manual ou automática (a atividade de extração automática é descrita no Capítulo 4) e são fornecidas ao módulo de unidades multipalavras para que sejam identificadas e marcadas nos textos.

Depois disso, inicia-se o processo de estabelecimento de correspondência entre as unidades lexicais. Para tanto, é calculado um escore de associação utilizando a medida *T-score* (descrita no Capítulo 2), sendo que os escores de associação para as unidades multipalavras são calculados semelhantemente aos das palavras simples. Para o cálculo desta medida, respeita-se a restrição estabelecida pelo módulo de classe fechada, que determina que unidades de classe fechada (pronomes, artigos, preposições e conjunções) só podem ser alinhadas com unidades da mesma classe, assim como unidades de classe aberta (demais categorias morfológicas) só podem ser alinhadas com unidades da mesma classe. Além disso, respeita-se também a restrição do módulo de pesos de posição que determina uma janela de correspondência (no caso deste trabalho foi determinada uma janela com n igual a três unidades) entre os bitextos, limitando a busca pelas unidades correspondentes.

Obedecendo a todas estas restrições, o escore de associação é finalmente calculado para cada uma das unidades do texto fonte produzindo os resultados mostrados na Tabela 13.

Tabela 13 – Escores de associação produzidos pelo *LWA*.

Unidades do Texto Fonte	Unidades do Texto Alvo	<i>T-score</i>
ESTE PRON	THIS DT	5.62470404668098
TRABALHO_N	PROPOSES VBZ	1.94127199461015
	DISSERTATION NN	4.90603519137623
PROPÔE_VERB	DISSERTATION NN	1.39130213017628
	PROPOSES VBZ	1.99594979273173
	LINGUISTIC JJ	0.996962344548801

A unidade candidata do texto alvo é obtida através do maior escore de associação, satisfazendo duas condições: que o escore seja maior que um determinado limiar e que a frequência global do par seja suficientemente alta. Para determinar o limiar de escore de associação e o limiar de frequência para um par de unidades lexicais, foi realizada uma etapa de otimização utilizando um corpus composto por 10 bitextos (dissertações de

mestrado e teses de doutorado da área de Ciência da Computação da Universidade Federal de Minas Gerais¹²), contendo 2.545 palavras em português e 2.412 palavras em inglês. Foram obtidos os seguintes resultados: 1) limiar do escore de associação = 0,99 e 2) limiar de frequência = 2. Para auxiliar no desempenho do método, é utilizado também um teste de cognatos, utilizando a medida *LCSR*, para ajudar na escolha da unidade candidata.

Com base nos escores da Tabela 13, alguns exemplos dos alinhamentos finais produzidos pelo *LWA* (outros exemplos de alinhamentos propostos pelo alinhador *LWA* podem ser vistos no Apêndice 2) são apresentados na Tabela 14.

Tabela 14 – Exemplos de alinhamentos propostos pelo *LWA*.

ALINHAMENTOS PROPOSTOS		
TEXTO FONTE	TEXTO ALVO	<i>T-score</i>
ESTE PRON	THIS DT	5.62470404668098
TRABALHO N	DISSERTATION NN	4.90603519137623
PROPÕE VERB	PROPOSES VBZ	1.99594979273173

As avaliações de desempenho dos métodos descritos nesta seção são descritas na próxima seção.

5.2. Avaliação dos Alinhadores Lexicais

Depois de realizado o processo de alinhamento, é necessário realizar uma etapa de avaliação dos resultados obtidos, a fim de determinar o desempenho dos alinhadores (*SIMR* e *LWA*). Para tal avaliação foram utilizadas as métricas revocação, precisão e medida-F, conforme descritas no Capítulo 2.

Nas próximas subseções as avaliações dos alinhadores lexicais são descritas em detalhes.

5.2.1. Alinhador *SIMR*

Para a avaliação do alinhador *SIMR* foi utilizado para teste o corpus de teste *PL*, composto por 65 pares de textos paralelos, contendo 11.341 palavras nos textos fonte e 10.217 palavras nos textos alvo. Neste processo foram propostos 2.046 alinhamentos.

¹² Disponível em: <http://www.npd.dcc.ufmg.br/orientacoes/orientacoes%20concluidas.htm>

Como resultados da avaliação realizada no corpus *PE*, foram encontrados os valores contidos na Tabela 15 mostrada a seguir.

Tabela 15 – Valores de revocação, precisão e medida-F obtidos na realização da tarefa de alinhamento.

	Revocação (%)	Precisão (%)	Medida-F (%)
<i>SIMR</i>	19,21	86,70	31,45

Além da avaliação considerando apenas os alinhamentos corretos, foi também realizada uma avaliação levando em consideração os alinhamentos considerados parcialmente corretos (aqueles em que falta uma parte do alinhamento, mas a correspondência proposta faz parte do alinhamento correto), segundo as métricas descritas no final do Capítulo 2. Os novos valores de revocação, precisão e medida-F obtidos são apresentados na Tabela 16.

Tabela 16 – Valores de revocação, precisão e medida-F considerando alinhamentos parcialmente corretos.

	Revocação (%)	Precisão (%)	Medida-F (%)
<i>SIMR</i>	20,27	92,93	33,28

Como se pode observar, este alinhador possui valor de revocação muito baixo. Isto se deve ao fato de que o meio utilizado para estabelecimento entre as unidades do texto é a técnica de identificação de cognatos; nenhuma outra métrica é utilizada. Assim, entre todas as unidades que compõem os textos fonte só serão alinhadas aquelas que possuem cognatos nos textos alvo. A utilização desta técnica acarreta alguns erros de alinhamento, pois existem alguns falsos cognatos (palavras com grafia semelhante, mas que não são traduções uma da outra). Como exemplo de falso cognato, podemos citar o verbo “pretender” do português e o verbo “pretend” do inglês, que apesar de possuírem grafias semelhantes não são traduções mútuas, as traduções corretas para cada uma destas unidades é “pretender:intend” e “fingir:pretend”.

Um outro ponto importante do funcionamento do *SIMR* é que só são consideradas no processo de alinhamento as unidades simples (palavras); as unidades multipalavras não

são levadas em consideração, acarretando erros de estabelecimento de correspondência entre as unidades dos textos paralelos.

Adicionalmente, foi realizada uma comparação do método *SIMR* quando aplicado ao alinhamento lexical e quando aplicado ao alinhamento sentencial (Caseli, 2003). Pode-se observar que, no alinhamento lexical, o método *SIMR* possui uma precisão boa (86,70%), porém, uma revocação muito baixa (19,21%). Por outro lado, no alinhamento sentencial, os alinhamentos lexicais produzidos pelo *SIMR* são utilizados como entrada para um algoritmo que alinha sentenças. Os alinhamentos sentenciais são determinados a partir dos mapeamentos retornados pelo *SIMR* e de informações sobre as fronteiras das sentenças. Os valores obtidos, nesse caso, foram bons tanto para a revocação (99,28%) quanto para a precisão (99,04%), mostrando assim que o *SIMR* desempenha um papel importante neste tipo de alinhamento.

5.2.2. Alinhador *LWA*

Durante a realização das primeiras tarefas de alinhamento usando o alinhador baseado no método *LWA*, foi possível verificar que alguns dos textos presentes no cópulo de teste (cópulo *PE*) causavam problemas na determinação de correspondência entre as unidades lexicais (palavras e unidades multipalavras) do texto, uma vez que o par de textos paralelos não possuía a mesma quantidade de sentenças (ou seja, texto fonte possui mais sentenças que o texto alvo, ou, texto alvo possui mais sentenças que o texto fonte).

Dessa forma, dos 65 pares de textos contidos no cópulo *PE*, foi necessário excluir 19 pares, eliminando com isso todos os textos que não possuíam a mesma quantidade de sentenças, permitindo que as possíveis correspondências entre as unidades lexicais fossem estabelecidas. O cópulo de teste final foi formado por 46 pares de textos paralelos português-inglês, contendo um total de 7.327 palavras nos textos fonte (sendo 1.718 palavras não repetidas).

O alinhador *LWA* foi utilizado em quatro tarefas distintas de alinhamento, uma para cada lista de unidades multipalavras extraída automaticamente via método da *Esperança Mútua* e via Pacote *NSP* (detalhadas no capítulo 2). O número de alinhamentos propostos é apresentado na Tabela 17 e os valores de revocação, precisão e medida-F são apresentados na Tabela 18 e na Figura 13.

Tabela 17 – Total de alinhamentos propostos em cada uma das tarefas de alinhamento.

	Listas Utilizadas	Alinhamentos Propostos
Esperança Mútua	<i>EP/EI</i>	7.032
	<i>GP/GI</i>	7.167
Pacote NSP	<i>EP/EI</i>	7.029
	<i>GP/GI</i>	7.171

Tabela 18 – Valores de revocação, precisão e medida-F obtidos nas tarefas de alinhamento utilizando listas extraídas automaticamente.

	Listas	Revocação (%)	Precisão (%)	Medida-F (%)
Esperança Mútua	<i>EP/EI</i>	38,65	33,94	36,14
	<i>GP/GI</i>	41,50	35,76	38,41
Pacote NSP	<i>EP/EI</i>	44,24	38,86	41,37
	<i>GP/GI</i>	44,76	38,54	41,42

Figura 13 - Contraste entre os valores de revocação, precisão e medida-F das tarefas de alinhamento utilizando as listas extraídas automaticamente.

É possível observar que os todos os valores de revocação e precisão são muito baixos, e que os melhores foram obtidos com a utilização das listas de unidades multipalavras geradas a partir do pacote *NSP*, tanto para o corpus específico quanto para o corpus geral.

Diferentemente da avaliação do método descrita por Ahrenberg, Andersson e Merkel (2000), a qual apresentou uma boa precisão (entre 83,9 e 96,7%), uma revocação (entre 50,9 e 67,1%) e uma medida-F (entre 69,79 e 70,10%) mais baixas, neste experimento os resultados obtidos não apresentaram valores bons. Com isso, decidiu-se iniciar uma etapa de identificação dos fatores que estariam levando aos baixos valores de revocação e precisão.

Identificando Fatores que Levam a Baixos Valores de Revocação e Precisão

Alguns pontos foram focados na investigação realizada, entre eles: 1) a influência das listas extraídas automaticamente no processo de alinhamento, 2) a influência das traduções muito livres, 3) o tamanho do corpus utilizado e 4) a influência da etiquetagem POS no processo de alinhamento.

Etapa 1 – Influência das listas extraídas automaticamente no processo de alinhamento

Para avaliar o primeiro dos fatores foi realizada uma tarefa de extração manual das unidades multipalavras. Os produtos desta tarefa de extração manual foram duas listas de unidades multipalavras, uma para o português brasileiro, contendo 297 unidades e uma para o inglês, contendo 304 unidades. As unidades multipalavras em questão foram então marcadas nos 46 pares de textos que foram utilizados no experimento anterior e um novo processo de alinhamento das unidades lexicais foi realizado, gerando 6.656 alinhamentos. É importante ressaltar que essa providência garante, por um lado, que toda unidade multipalavra seja identificada, porém não garante necessariamente que as multipalavras sejam corretamente alinhadas, pois isso dependerá fortemente da frequência relativa das unidades multipalavras no corpus. Como resultado da avaliação deste processo foram obtidos os valores mostrados na Tabela 19.

Tabela 19 – Valores de revocação, precisão e medida-F obtidos utilizando listas de unidades multipalavras extraídas manualmente.

	Revocação (%)	Precisão (%)	Medida-F (%)
<i>LWA</i>	51,93	48,18	49,98

A análise de tais valores permite constatar, como já era esperado, que as listas de multipalavras extraídas automaticamente afetaram negativa e significativamente o desempenho do alinhador. Isso pode ser explicado pelo fato de que as listas geradas são dependentes dos corpúis utilizados no processo de extração, sendo tão limitadas quanto tais corpúis. É importante mencionar que a revocação obtida nesta etapa já alcançou o intervalo da revocação na avaliação do método descrita em (Ahrenberg, Andersson & Merkel, 2000), entre 50,9 e 67,2%.

Etapa 2 – Influência das traduções muito livres

Considerando o segundo item da investigação, foi feita uma análise dos textos presentes no corpúis *PE* e foi possível verificar que alguns dos textos eram traduzidos muito livremente. Assim, selecionou-se um subconjunto do corpúis (aproximadamente 20% do corpúis *PE*) formado por 10 pares de textos, contendo 1.409 palavras nos textos fonte (com 477 palavras não repetidas). Esse subconjunto de pares de textos com traduções mais literais foi utilizado como entrada para o alinhador *LWA*, sendo marcadas as unidades multipalavras contidas nos textos com base nas listas extraídas manualmente. Foram propostos para este subconjunto 1.274 alinhamentos. Os valores da avaliação são exibidos na Tabela 20.

Tabela 20 – Valores de revocação, precisão e medida-F obtidos utilizando um subconjunto do corpúis *PE* e as listas de unidades multipalavras extraídas manualmente.

	Revocação (%)	Precisão (%)	Medida-F (%)
<i>LWA</i>	57,82	47,56	52,19

É possível observar, portanto, que a subtração de textos com traduções mais livres melhorou um pouco (~10%) a revocação do método, porém não contribuiu na precisão, que até abaixou um pouco. Com isso, excluímos este item da lista de fatores que prejudicam o bom desempenho do alinhador.

Etapa 3 – Tamanho do corpúis utilizado

Em seguida, foi avaliada a influência do tamanho do corpúis no desempenho do alinhador, considerando que, no experimento realizado em (Ahrenberg, Andersson e Merkel, 2000), os corpúis utilizados possuíam entre 66.693 (romance) e 169.779 (manual de

um programa de computador) palavras nos textos fonte, que são números consideravelmente maiores que a quantidade de palavras nos textos do *cópus PE*, de 7.327 palavras. Para esta etapa, foi construído um novo *cópus* adicionando-se 41 pares de textos aos 46 pares já disponíveis no *cópus PE*. Estes novos textos são resumos e *abstracts* de dissertações de mestrado e teses de doutorado da área de Ciência de Computação da Universidade Federal de Minas Gerais, pertencentes a subdomínios variados da Computação. Os novos textos foram marcados, assim como o *cópus PE*, com etiquetas indicando início e fim de parágrafo, sentença e palavra, além disso, foram etiquetados com o etiquetador MXPOST.

O total de palavras contidas nos textos fonte do novo *cópus* é de 11.713 (o *cópus* original foi aumentado em 4.386 palavras nos textos fonte – contendo 2.867 palavras não repetidas), correspondendo a um aumento de aproximadamente 45% no tamanho do *cópus* original. Este novo *cópus* foi utilizado como entrada para o alinhador *LWA*, sendo utilizadas as listas extraídas manualmente para a marcação das unidades multipalavras. Para este novo *cópus* foram propostos 12.560 alinhamentos. O resultado deste processo é mostrado na Tabela 21.

Tabela 21 – Valores de revocação, precisão e medida-F obtidos utilizando o novo *cópus* (contendo 87 pares de textos paralelos).

	Revocação (%)	Precisão (%)	Medida-F (%)
<i>LWA</i>	55,20	51,48	53,27

Os valores obtidos mostram que o aumento de 45% do tamanho do *cópus* provocou um aumento de cerca de 3% na precisão do alinhador. Isto pode ser explicado pelo fato de que o *escore* de associação utilizado para determinar a correspondência entre as palavras dos textos fonte e alvo considera a frequência da ocorrência das unidades lexicais (palavras e/ou unidades multipalavras) nos textos e da co-ocorrência dos pares de palavras e/ou unidades multipalavras nos bitextos. Por isso, é necessário que a frequência dos pares de unidades lexicais seja relativamente alta para que os alinhamentos sejam corretamente estabelecidos. Apesar de ter ocorrido um aumento na precisão do alinhador, o aumento ainda foi muito pequeno, indicando que o *cópus* deve ser aumentado muito mais para que se alcance uma precisão satisfatória.

Durante esta etapa da investigação, foi possível determinar dois aspectos que influenciam diretamente no desempenho do *cópus*: a quantidade de palavras não repetidas contidas no *cópus* e a falta de padronização das traduções.

A não repetição de palavras afeta o desempenho, pois tais palavras não obedecem à exigência de alta frequência exigida pelo método para o cálculo adequado do *escore* de associação para determinação da correspondência entre as unidades lexicais, como mencionado anteriormente.

A falta de padronização nas traduções acontece quando uma unidade lexical tem vários sinônimos para sua tradução, como, por exemplo, a unidade “trabalho” que, nos textos alvo, é geralmente traduzida como “*work*”, “*thesis*”, “*dissertation*”. Assim, “trabalho” possui uma frequência maior do que suas traduções, uma vez que, para uma mesma unidade existem várias traduções sendo utilizadas. Este aspecto também dificulta o estabelecimento de correspondências, pois, no cálculo do *escore* de associação, a unidade “trabalho” pode ser alinhada com uma outra unidade em inglês que não seja a sua tradução, desde que a unidade em inglês possua uma frequência relativamente mais próxima de sua frequência.

Etapa 4 – Influência da etiquetação POS no alinhamento

Finalizando o processo de investigação, foi avaliada a influência da etiquetação morfológica (POS) no processo de alinhamento, a fim de determinar se o processo de etiquetação realizado com o etiquetador MXPOST provocava algum tipo de alteração na precisão do alinhador *LWA*. Como, para isso, era necessário ter um *cópus* de teste corretamente etiquetado, formou-se um sub*cópus* menor, com 10 pares de textos paralelos (o mesmo da etapa 2), o qual foi verificado manualmente quanto à etiquetação. Se fosse verificada a influência eventual da correção das etiquetas, proceder-se-ia da mesma forma com o *cópus* completo.

Como nos demais casos apresentados, as unidades multipalavras foram marcadas nos textos de acordo com as listas extraídas manualmente. Os resultados gerados nesta fase são mostrados na Tabela 22.

Tabela 22 – Valores de revocação, precisão e medida-F obtidos utilizando o subconjunto do corpus *PE* com as etiquetas morfológicas corrigidas.

	Revocação (%)	Precisão (%)	Medida-F (%)
<i>LWA</i>	57,92	47,59	52,25

Comparando estes resultados com aqueles obtidos na Tabela 19 (em que a avaliação foi realizada com o mesmo subconjunto do corpus *PE*) observa-se que o processo de etiquetagem automática realizado com o auxílio da ferramenta MXPOST não exerce influência negativa no desempenho do alinhador lexical implementado. Com isso, decidiu-se não realizar a correção das etiquetas no corpus completo.

Finalizadas as quatro etapas da investigação, foi necessário recalcular a revocação e a precisão, desta vez, porém, considerando os alinhamentos parcialmente corretos além dos alinhamentos corretos. Para esta avaliação, foi utilizado como entrada para o alinhador o corpus construído na etapa 3 da tarefa de investigação, que havia produzido o maior valor de precisão. Os novos valores de revocação, precisão e medida-F obtidos são apresentados na Tabela 23.

Tabela 23 – Valores de revocação, precisão e medida-F considerando alinhamentos parcialmente corretos.

	Revocação (%)	Precisão (%)	Medida-F (%)
<i>LWA</i>	51,63	54,05	52,81

Pode-se verificar, comparando a Tabela 23 com a Tabela 21, que o valor de revocação diminuiu. Isto pode ser explicado pelo fato de que na fórmula da revocação convencional (na qual são considerados apenas os alinhamentos corretos) é utilizada a razão entre o número de alinhamentos corretos e o número de alinhamentos de referência (todos os corretos possíveis) apenas. Enquanto que, na fórmula da revocação, que considera alinhamentos parcialmente corretos, se utiliza a razão entre o número de alinhamentos corretos e parcialmente corretos e o número de alinhamentos propostos e não propostos pelo método (que, neste caso, é maior que o número de alinhamentos de referência). Por outro lado, houve um aumento da precisão, pois, com a nova fórmula para o cálculo da precisão, são considerados também os alinhamentos parcialmente corretos.

É interessante mencionar que, como mostrado no Capítulo 2, o *LWA* também foi avaliado no projeto ARCADE, juntamente com outros 4 sistemas (CEA, LILA, RALI, XEROX), sendo obtida uma precisão de 75% para o melhor sistema. Nesta tarefa foi considerado um corpus inglês-francês (contendo aproximadamente 1,1 milhão de palavras). Os resultados obtidos pelo *LWA* nesta tarefa foram: 60% de precisão, 57% de revocação e 58% de medida-F. Estes números mostram que existe uma influência marcante das línguas envolvidas no processo de alinhamento, pois quando foi utilizado o par de línguas inglês-francês, sendo o francês língua de origem latina assim como o português, o desempenho do *LWA* foi inferior em relação ao seu desempenho aplicado ao par inglês-sueco, sendo ambas as línguas de origem germânica. Os valores da tarefa de avaliação do projeto ARCADE se aproximam muito dos melhores valores obtidos neste trabalho (Tabela 21).

Tomando como base os resultados obtidos nas quatro etapas da investigação, foi possível identificar dois aspectos que têm influência direta no desempenho do alinhador *LWA*: extração automática de unidades multipalavras e tamanho do corpus de teste. A utilização do processo de extração automática de unidades multipalavras resulta em baixa precisão do método. Já o aumento do tamanho do corpus provoca o aumento da precisão, uma vez que, aumentando o corpus, aumenta-se a frequência das unidades lexicais gerando escores de associação mais altos, portanto, maior é o número de correspondências corretas estabelecidas pelo alinhador.

Capítulo 6

Conclusões

Este trabalho teve como objetivo a investigação, implementação e avaliação de técnicas de alinhamento lexical de textos paralelos português-inglês. Entre os métodos selecionados estavam o *SIMR*, representante da abordagem empírica, baseado em reconhecimento de padrões, e o *LWA*, representante da abordagem híbrida, baseado na distribuição e co-ocorrência das unidades lexicais nos segmentos do bitexto e na utilização de conhecimento lingüístico, como etiquetagem morfológica e listas de unidades multipalavras.

Para o tratamento de unidades multipalavras, foram utilizadas duas técnicas de extração automática, a *Esperança Mútua* e o pacote *NSP*. A extração automática de unidades multipalavras é uma tarefa difícil, pois na maioria das vezes envolve a utilização de aspectos específicos da língua (como, por exemplo, os filtros que são utilizados para eliminação de unidades não relevantes) para melhorar o desempenho dos extratores. Um outro ponto que deve ser considerado é a dependência dos corpú que são utilizados. Quantas e quais serão as unidades multipalavras que estarão presentes nas listas geradas depende dos corpú usados como entrada para os extratores automáticos. Desta forma, se os corpú não forem suficientemente abrangentes, as listas poderão não ser tão completas quanto deveriam. Foi possível observar que a maioria dos extratores de multipalavras existentes hoje necessita ainda de melhoramentos à medida que não geram apenas unidades multipalavras relevantes. Neste trabalho, o pacote *NSP* foi o que apresentou melhor desempenho para os corpú em questão: 44,24 – 44,76 % de revocação e 38,86 – 38,54% de precisão. Esses valores mostraram-se baixos demais para a aplicação alvo, o alinhamento lexical.

As tarefas de avaliação dos métodos de alinhamento lexical permitiram observar que o não tratamento das unidades multipalavras leva a baixos valores de revocação, conforme resultados obtidos pelo alinhador empírico *SIMR*, apesar de este alinhador ter obtido alta precisão 86,70% (considerando-se apenas alinhamentos corretos) e 92,93% (considerando também alinhamentos parcialmente corretos). Por outro lado, o tratamento de tais unidades é custoso e sua extração automática influencia negativamente a precisão do alinhador, como mostraram os resultados do alinhador *LWA*.

O tamanho do *cópus* é outro ponto que influencia diretamente na precisão do alinhador *LWA*, uma vez que este alinhador utiliza uma medida, para determinar a correspondência entre as unidades lexicais do bitexto, que se baseia na frequência da ocorrência dos pares de unidades. Unidades lexicais de *cópus* maiores possuem frequências mais altas.

Apesar de não terem sido obtidos índices de revocação e precisão significativos, mas que se aproximam bastante dos valores obtidos na tarefa de avaliação desenvolvida pelo projeto ARCADE, foi possível determinar quais fatores e em que medida eles contribuem para o alcance desses índices. Além disso, ressalta-se o fato de este trabalho ser inédito na área de alinhamento lexical envolvendo o português do Brasil.

Como trabalho futuro imediato, sugere-se a compilação e preparação de um grande *cópus* paralelo. Embora não seja uma tarefa complexa, ela é trabalhosa e demanda disponibilidade de textos e tempo para sua preparação – pré-processamento, alinhamento, etiquetagem, etc.

Adicionalmente, os resultados obtidos aqui sugerem que pode ser promissor um método de alinhamento de dois passos, que combine o tratamento de multipalavras do alinhador *LWA*, pelo seu bom índice de revocação, com o alinhador *SIMR*, pela sua promissora precisão, da seguinte maneira: utiliza-se o *LWA* para gerar os candidatos a alinhamento e o *SIMR* para certificar as correspondências propostas pelo *LWA*.

Finalmente, seria interessante utilizar os alinhadores obtidos para a criação de outros recursos, como dicionários, glossários e terminologias bilíngües, regras de gramáticas de tradução, entre outros. A avaliação do recurso produzido pelo alinhador ou da aplicação que dele se utiliza levantaria outros resultados importantes para essa área e, em particular, para aquela voltada à Língua Portuguesa do Brasil.

Referências Bibliográficas

- AHRENBORG, L.; ANDERSSON, M.; MERKEL, M. (1998). A simple hybrid aligner for generating lexical correspondences in parallel texts. *In: Proceedings of Association for Computational Linguistics*. p. 29-35.
- AHRENBORG, L.; ANDERSSON, M.; MERKEL, M. (2000). A knowledge-lite approach to word alignment. *In: Véronis, J., ed. Parallel text processing: Alignment and use of translation corpora*. s.l.: Kluwer Academic Publishers, p. 97-116.
- AHRENBORG, L.; ANDERSSON, M.; MERKEL, M. (2002). A System for Incremental and Interactive Word Linking. *In: Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, 29-31 May. p. 485-490.
- AHRENBORG, L.; MERKEL, M.; HEIN, A. S.; TIEDEMANN, J. (2000). Evaluation of word alignment systems. *In: 2nd International Conference on Language Resources & Evaluation (LREC 2000)*. p. 1255-1261.
- AIRES, R. V. X.; ALUÍSIO, S. M. (2001). Criação de um corpus com 1.000.000 de palavras etiquetado morfossintaticamente. Série de Relatórios do NILC. NILC-TR-01-8, Outubro. 14p.
- CASELI, H. M. (2002). Alinhamento Sentencial de textos paralelos Português-Inglês. Monografia de Qualificação. ICMC-USP. 56p.
- CASELI, H. M. (2003). Alinhamento Sentencial de textos paralelos Português-Inglês. Dissertação de Mestrado. ICMC-USP. 101p.
- CASELI, H.M.; NUNES, M.G.V. (2002). *TagAlign*: Uma ferramenta de pré-processamento de textos. Série de Relatórios do NILC. NILC-TR-02-09, Junho. 37p.

- CHURCH, K. W. (1993). Char_align: A program for aligning parallel texts at the character level. *In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH. p. 1-8.
- DAGAN, I.; CHURCH, K. W.; GALE, W. (1993). Robust bilingual word alignment for machine-aided translation. *In: Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*. Columbus, Ohio. p. 1-8.
- DIAS, G.; KAALEP, H. (2002). Automatic Extraction of Multiword Units for Estonian: Phrasal Verbs. *In: H. Metslang, M. Rannut (eds.) Languages in Development*, Linguistic Edition 41, Lincom-Europa, München.
- KAY, M.; RÖSCHEISEN, M. (1993). Text-translation alignment. *In: Proceedings of Association for Computational Linguistics*. 19(1), p.121-42.
- MARTINS, M. S.; CASELI, H. M.; NUNES, M. G. V. (2001). A construção de um corpus de textos paralelos inglês-português. Série de relatórios do NILC. NILC-TR-01-5. 62p.
- MELAMED, I. D. (1997b). A portable algorithm for mapping bitext correspondence. *In: Proceeding of the 35th Annual Meeting of the Association for Computational Linguistics*. p. 305-312.
- MELAMED, I. D. (1999) Bitext maps and alignment via pattern recognition. *In: Computational Linguistics*. 25(1). p. 107 –130.
- MELAMED, I. D. (2000). Pattern recognition for mapping bitext correspondence. *In: Véronis, J., ed. Parallel text processing: Alignment and use of translation corpora*. s.l.: Kluwer Academic Publishers. p. 25-47.
- MELAMED, I. D.; AL-ADHAILEH, M. H.; KONG, T. E. (2001). Malay-English bitext mapping and alignment using *SIMR/GSA* algorithms. *In: Malaysian National*

Conference on Research and Development in Computer Science (REDECS'01).
Selangor Darul Ehsan, Malaysia.

MERKEL, M.; ANDERSSON, M. (2000). Knowledge-lite Extraction of Multi-word Units with Language Filters and Entropy Thresholds. *In: Proceedings of RIAO'2000, Collège de France, Paris, France, April 12-14. Volume 1. p. 737-746.*

RATNAPARKHI, A. (1996). A Maximum Entropy Part-Of-Speech Tagger. *In: Proceedings of the Empirical Methods in Natural Language Processing Conference, May 17-18. University of Pennsylvania. p. 133-141.*

SCHONE, P.; JURAFSKY, D. (2001). "Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?". *In: Proceedings of the 2001 Empirical Methods in Natural Language Processing (EMNLP2001), Pittsburgh, PA. p. 100-108.*

SIMARD, M. (2000) Multilingual text alignment: Aligning three or more versions of a text. *In: Véronis, J., ed. Parallel text processing: Alignment and use of translation corpora. s.l.: Kluwer Academic Publishers. p. 49-68.*

SMADJA, F. A.; MCKEOWN, K. R.; HATZIVASSILOPOULOS, V. (1996). Translating collocations for bilingual lexicon: A statistical approach. *In: Computational Linguistics. Volume 22, Nº 1. Março. p. 1-38.*

TELINHA, M. F.; MANFRIM, A. M. P.; ALUÍSIO, S. M. (2003). Extração Automática de Termos de Textos em Português: Aplicação e Avaliação de Medidas Estatísticas de Associação de Palavras. Série de Relatórios do NILC. NILC-TR-03-10, Outubro. 42p.

VERÓNIS, J. (2000). From the Rosetta stone to the information society: A survey of parallel text processing. *In: Véronis, J., ed. Parallel text processing: Alignment and use of translation corpora. s.l.: Kluwer Academic Publishers. p. 1-24.*

VERÓNIS, J; L. ANGLAIS, P. (2000). Evaluation of parallel texts alignment systems: *The ARCADE project*. In: Véronis, J., ed. *Parallel text processing: Alignment and use of translation corpora*. s.l.: Kluwer Academic Publishers. p. 369-388.

Apêndice 1: Unidades Multipalavras Extraídas pelas Técnicas Implementadas

Neste apêndice são apresentadas algumas das unidades relevantes geradas pelas técnicas de extração utilizadas neste trabalho. As listas estão separadas de acordo com a técnica de extração utilizada para gerar as unidades multipalavras para cada um dos quatro corpuses (corpus EP, corpus GP, corpus EI e corpus GI) utilizados como entrada.

1. Esperança Mútua

A Tabela A mostrada a seguir contém algumas das unidades multipalavras (unidades relevantes) obtidas a partir da aplicação do extrator automático construído com base na técnica de *Esperança Mútua*.

Tabela A – Algumas unidades relevantes extraídas pela técnica de *Esperança Mútua*.

Cópus EP	Cópus GP	Cópus EI	Cópus GI
ad hoc	às vezes	case study	Bar Mitzvah
Além disso	abrir mão de	finite state automata	be used to
autômato finito determinístico	além disso	in fact	breaking apart
Ciclo de vida	apesar disso	in this	carry out
código fonte	Bob Marley	Marry Monads	Ed Thompson
de modo que	cara ou coroa	of the	fell apart
Donald Hung	em nível de	of this	find out
Engenharia de Software	Fernando Henrique Cardoso	pay attention	for example
et al	flauta doce	points of view	Garden of Eden
Gill Harel	ou seja	points out	make up
guerra fria	por isso	taken into account	of that
levam em conta	por que	there are	of these
Ponto de vista	Rio de Janeiro	Software Engineering	San Antonio
Por exemplo	São Paulo	vice versa	so far
Von Neumann	Semana Santa	World Wide Web	sooner or later

2. Pacote NSP

Na Tabela B apresentada abaixo estão listados exemplos das unidades multipalavras (unidades relevantes) extraídas com a utilização do pacote *NSP*.

Tabela B – Algumas unidades relevantes extraídas pelo pacote *NSP*.

Córpus EP	Córpus GP	Córpus EI	Córpus GI
Assim como	ao acaso	as well as	Ariel Sharon
Bases de dados	assim como	at least	based on
Engenharia Reversa	através de	based on	be used to
Estados Unidos	César Maia	break out	C I A
et al	Daniela Mercury	et al	how many
Klein Joshi	Euclides da Cunha	finite state machine	in the
ou seja	frente fria	in the	New York
Ponto flutuante	no entanto	instead of	of the
Por alto	Pedro Paulo Diniz	look at	point of view
Por engano	por favor	of this	Saddam Hussein
Por exemplo	por isto	on the	set up
Por outro lado	por meio de	Operating Systems	so much
Por que	pouco a pouco	points of view	sooner or later
São Paulo	São Carlos	taken into account	take into account
Tom Jobim	Via Crucis	World Wide Web	United States

Apêndice 2: Alinhamentos Corretos Propostos pelos alinhadores *LWA* e *SIMR*.

Neste apêndice são apresentados alguns exemplos dos alinhamentos corretos propostos pelos alinhadores utilizados neste trabalho (*LWA* e *SIMR*).

1. Alinhador *SIMR*

Na Tabela C são mostrados alguns dos alinhamentos corretos propostos pelo método *SIMR*, sendo utilizado como entrada o corpus *PE*.

Tabela C – Exemplos de alinhamentos corretos produzidos pelo alinhador *SIMR*.

ALINHAMENTOS	
TEXTO FONTE	TEXTO ALVO
validação	validate
suporta	supports
documento	document
agentes	agents
cenários	scenarios
heurísticas	heuristics
evolução	evolution
modelo	model
modelos	models
software	software
informais	informal
problemática	problematic
documentação	documentation
reflete	reflects
implementado	implemented
contexto	context
projeto	project
documentar	documenting
experimento	experiment
descrito	described
parte	part

2. Alinhador LWA

Na Tabela D são apresentados alguns exemplos dos alinhamentos corretos propostos pelo método *LWA*, quando utilizado como entrada o corpus de teste construído no Capítulo 5, na etapa 3 da tarefa de identificação dos problemas que acarretam baixos índices de revocação e precisão.

Tabela D Exemplos de alinhamentos corretos produzidos pelo alinhador *LWA*.

ALINHAMENTOS	
TEXTO FONTE	TEXTO ALVO
neste	in this
artigo	paper
uma	a
ferramenta	tool
para	to
e	and
essa	this
suporta	supports
a	the
abordagem	approach
baseia-se	is based
no	on the
documento	document
requisitos	requirements
e	and
propõe	proposes
seus	its
agentes	agents
cenários	scenarios
em seguida	then
os	the
são	are
também	also
heurísticas	heuristics
evolução	evolution
modelo	model
para	to
análise	analysis
exemplificadas	illustrated