PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# EXTRAÇÃO DE VOCABULÁRIO MULTILÍNGUE A PARTIR DE DOCUMENTAÇÃO DE SOFTWARE

LUCAS WELTER HILGERT

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Profª. Renata Vieira

Porto Alegre
2014

**Dados Internacionais de Catalogação na Publicação (CIP)**

H644e Hilgert, Lucas Welter
Extração de vocabulário multilíngue a partir de documentação de software / Lucas Welter Hilgert. – Porto Alegre, 2014.
97 p.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. Renata Vieira.

1. Informática. 2. Linguística Computacional. 3. Tradução Automática. 4. Engenharia de Software. I. Vieira, Renata. II. Título.

CDD 006.35

**Ficha Catalográfica elaborada pelo**
**Setor de Tratamento da Informação da BC-PUCRS**

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

# TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Extração de Vocabulário Multilíngue a partir de Documentação de Software" apresentada por Lucas Welter Hilgert como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 26/03/2013 pela Comissão Examinadora:

Profa. Dra. Renata Vieira – Orientadora — PPGCC/PUCRS

Profa. Dra. Vera Lúcia Strube de Lima – PPGCC/PUCRS

Profa. Dra. Helena de Medeiros Caseli – UFSCar

Homologada em 06/03/2014, conforme Ata No. 002 pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS
**Campus Central**
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

"Do or do not. There is no try."
(Star Wars: Episode V)

## AGRADECIMENTOS

Em primeiro lugar aos meus pais, à minha irmã, aos meus queridos avós (que tanto fizeram por mim) e a todos os meus familiares, por todo o apoio concedido e por toda a paciência e compreensão demonstrados ao longo destes dois anos.

À minha orientadora Renata por todos os ensinamentos, pelo apoio e, pela paciência demonstrada em relação a este aprendiz de pesquisador.

À minha querida Fernanda, que me mostrou o caminho mas infelizmente não pôde me ver trilhá-lo até o fim.

Aos meus colegas do grupo de pesquisa Roger, Larissa, Clarissa, Lucelene, Evandro, Daniela, Sandra, Tiago e Marlo, por todo o auxílio e apoio prestado nestes dois anos.

Um agradecimento especial a Dr.ª Aline Vanin pela prestatividade e pelo auxílio prestados durante a avaliação dos resultados e por todo o auxílio prestado.

Aos meus amigos e colegas: Lucas, Claiton, Luciana, Eli, Viviane, Joaquim, Leonardo, Bernardo, Thiago, Maria, Gabriel e tantos outros pelo festival de aleatoriedades que foi o tempo em que convivemos, por todo o apoio e incentivo. Vocês são os melhores.

Aos meus amigos de infância, mas irmãos de convivência Eduardo Henrique Spies, William Schneider e Ricardo Rohden, por tantas coisas que fica até complicado listar.

À minha querida namorada Viviane Linck Lara pela paciência, compreensão e apoio prestados principalmente durante as fases mais críticas da elaboração deste trabalho.

À todos aqueles que, embora não mencionados, sabem que tiveram uma participação importante nesta conquista.

Thank you! You are the best!

# EXTRAÇÃO DE VOCABULÁRIO MULTILÍNGUE A PARTIR DE DOCUMENTAÇÃO DE SOFTWARE

## RESUMO

Ferramentas e serviços de tradução de máquina (automática) em tempo real têm sido investigadas como uma alternativa à utilização de idiomas comum (*Lingua Franca*) durante reuniões de equipes com diferentes idiomas nativos. No entanto, como demonstrado por diferentes pesquisadores, este tipo de tecnologia ainda apresenta alguns tipos problemas que dificultam a sua utilização neste contexto, dentre os quais destaca-se neste trabalho as traduções inconsistentes (diferentes traduções atribuídas a uma mesma palavra em um mesmo contexto).

Dentre as soluções apontadas na literatura para melhorar a qualidade das traduções, destaca-se a construção de vocabulários multilíngues específicos de domínios. Sendo assim, neste trabalho é proposto um processo para a extração de vocabulário multilíngue a partir de documentos de *software*.

O processo proposto seguiu um conjunto de etapas consolidadas na literatura, tendo apresentado, como principal diferencial a forma pela qual o vocabulário de domínio é identificado: mediante a utilização de *softwares* extratores de terminologia.

Uma avaliação manual dos dicionários gerados pelo processo demonstrou uma precisão de 81% na tradução de palavras simples e 39% na tradução de expressões multipalavras. Estes valores demonstraram-se condizentes com os trabalhos relacionados.

**Palavras-Chave:** Vocabulário Multilíngue, Desenvolvimento Global de Software, Tradução de Máquina.

# MULTILINGUAL VOCABULARY EXTRACTION FROM SOFTWARE DOCUMENTATION

## ABSTRACT

Real-time machine translation tools and services have been investigated as an alternative approach to the utilization of a common language (*lingua franca*) during distributed meetings involving teams with differet native languages. However, as presented by different research works, this kind of technologies presents a set of problems that difficults the communication.

Among the solution proposed in the literature, the construction of domain specific vocabularies are highlited. This work propose a multilingual vocabulary extraction process for multilingual dicionary entries extraction from software user guides.

The process here proposed follows a well stablished set of steps presenting as the main difference the way in wich the domain vocabulary is identified: through the utilization of terminology extraction softwares.

A manual evaluation of the dictionaries generated by the process has shown a precision of 81% for simple world translation and 39% for multiword expressions. This values are consistent with the related work.

**Keywords:** Multilingual Vocabulary, Global Software Development, Machine Transaltion.

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS

PLN. – Processamento da Linguagem Natural

IEEE. – *Institute of Electrical and Electronics Engineers*

IRC. – *Internet Relay Chat*

PLN. – Processamento de Linguagem Natural

GMA. – *Geometric Mapping and Alignment*

PoS. – *Part-of-Speech*

NLTK. – *Natural Language Toolkit*

HMM. – *Hidden Markov Model*

MI. – *Mutual Information*

LCSR. – *Longest Common Subsequence Ratio*

XML. – *eXtensible Markup Language*

PDF. – *Portable Document Format*

HTML. – *Hypertext Markup Language*

# SUMÁRIO

# 1. INTRODUÇÃO

Serviços de tradução de máquina em tempo real têm sido considerados como alternativas promissoras ao auxílio na comunicação durante a execução de tarefas colaborativas envolvendo equipes multilíngues [6] [8] [7] [52] [51].

Das áreas nas quais este tipo de tecnologia tem sido explorada, destacam-se neste trabalho as pesquisas relacionadas ao Desenvolvimento Global de *Software* [6] [7], área na qual a comunicação entre times com diferentes idiomas nativos é frequentemente realizada por intermédio de idiomas comuns (inglês, por exemplo), nos quais nem sempre os participantes apresentam a fluência necessária [7].

No entanto o desempenho dos serviços e ferramentas de tradução ainda está longe da perfeição, apresentando um alguns problemas a serem solucionados [7] [51]. Dentre os problemas identificados, tanto na bibliografia quanto em uma análise conduzida sobre registros do experimento de Calefato *et al.* [7], destacam-se neste trabalho aqueles relacionados ao vocabulário empregado como, por exemplo, a inconsistência de tradução de palavras e termos. Este tipo de inconsistências, segundo Yamashita *et al.* [51], tende dificultar o estabelecimento de um conhecimento comum (*common ground*) entre os participantes da reunião, condição necessária ao sucesso da execução de tarefas colaborativas [52].

Uma das possíveis soluções apontadas na literatura [37] para o problema das inconsistências, e consequente melhoria da qualidade das ferramentas e serviços de tradução, consiste na especialização dos vocabulário bilíngue utilizado por essas mediante a ampliação de seus dicionários com vocabulário bilíngue específico dos domínios sobre as quais serão utilizadas.

Sendo assim, este trabalho propõe um processo para a extração de vocabulário multilíngue (português-inglês) a partir da documentação de *software*, tendo esse como objetivo a ampliação dos dicionários da ferramenta Apertium [15], (ferramenta de tradução) a ser utilizada em futuras etapas do projeto no qual este trabalho encontra-se inserido (apresentado na seção 2.1).

Este documento encontra-se organizado de modo que o Capítulo 2 apresenta o contexto no qual este trabalho encontra-se inserido, o Capitulo 3 apresenta o corpus construído e demais recursos levantados, o Capítulo 4 apresenta os principais conceitos relacionados ao processo proposto, o Capítulo 5 apresenta os trabalhos relacionados a esse, o Capítulo 6 apresenta e descreve o processo proposto por este trabalho e, por fim, os capítulos 7 e 8 apresentam respectivamente uma análise dos resultados obtidos e as principais discussões em relação a esse.

# 2. CONTEXTUALIZAÇÃO DA PESQUISA

Este capítulo descreve o contexto no qual o presente trabalho de pesquisa encontra-se inserido. Nesse, são apresentados o projeto do qual este trabalho faz parte (Seção 2.1), os problemas identificados e alternativas encontradas na literatura para a solução desses (Seção 2.2). Por fim, são apresentados os objetivos do trabalho e os métodos utilizados no decorrer da pesquisa(Seção 2.4).

## 2.1 Projeto

O Desenvolvimento Global de *Software* consiste em uma abordagem de desenvolvimento na qual as equipes envolvidas e os *stakeholders* (interessados no *software*) encontram-se geograficamente dispersos a nível global [1]. Este tipo de desenvolvimento explora questões como a diversidade cultural e as diferenças de fusos horários visando diminuir o tempo e os custos do processo de desenvolvimento de *softwares*, além de promover a aproximação das empresas com os clientes ao redor do globo [1]. Assim, esta estratégia tem se demonstrado como um diferencial competitivo para as empresas de desenvolvimento [7].

A dispersão geográfica, no entanto, tende a trazer uma série de dificuldades, das quais destacam-se as linguísticas principalmente no que tange a diferença de idiomas. Esta diferença é apontada tanto como um dos maiores obstáculos à execução de atividades colaborativas [31] quanto como um dos principais fatores de sucesso dos projetos de desenvolvimento em países como Índia, Filipinas, Irlanda e Singapura. Estes países se destacam por seus altos índices de proficiência na língua inglesa, frequentemente empregada como idioma comum (*common language*) durante reuniões de equipes multilíngues.

No entanto, existem países considerados competidores no mercado global de desenvolvimento de *software* que não possuem um número suficiente de profissionais fluentes na língua inglesa. No caso do Brasil, por exemplo, tem-se um crescimento médio anual de 6,5% de empregos no setor de TI (Tecnologia da Informação) [7] sendo que apenas aproximadamente 5,4% da população tem fluência no idioma inglês (aproximadamente 10 milhões de pessoas), contra 90 milhões da Índia.

Este trabalho ocorre no contexto do projeto "*O Efeito do Processamento da Linguagem Natural no Desenvolvimento da Capacidade do Brasil no Mercado Global de Desenvolvimento de Software*", que tem como objetivo auxiliar na inserção de equipes brasileiras no mercado global de desenvolvimento de *software* mediante a experimentação e utilização de métodos, técnicas e ferramentas da área de Processamento da Linguagem Natural (PLN) .

O projeto tem como foco principal a experimentação, utilização e adaptação de de serviços de tradução de máquina em tempo real, tendo como principais objetivos:

- Mapear as etapas e práticas usuais na engenharia de requisitos, identificando tarefas de comunicação;
- Fazer um levantamento de vocabulário específico destas práticas;
- Propor métodos de aquisição de vocabulário nas comunicações entre equipes distribuídas;
- Investigar, propor e avaliar da geração de vocabulário controlado multilíngue orientado a tarefas;
- Estudar o uso de ferramentas *off the shelf* na comunicação entre equipes distribuídas, em especial envolvendo idiomas distintos;
- Estudar o incremento de ferramentas através do mapeamento de vocabulários específicos.

## 2.2 Tradução de Máquina em Atividades Colaborativas Distribuídas

A utilização de serviços de tradução de máquina em tempo real tem sido considerada por pesquisadores como uma alternativa promissora à utilização de idiomas comuns (*lingua franca*) durante a comunicação envolvendo participantes com diferentes línguas nativas [7] [52] [51].

No contexto da execução de tarefas colaborativas entre participantes geograficamente distribuídos, destacam-se os trabalhos de Calefato *et al.* [6] [8] [7], Yamashita e Ishida [52] e Yamashita *et al.* [51], cujas constatações motivaram a realização do presente trabalho.

Calefato *et al.* [6] [8] [7] investigaram a utilização de tradutores automáticos (*Google Translator*) durante reuniões de equipes distribuídas (e multilíngues) de desenvolvimento de *software*. A investigação foi realizada através de experimentos controlados e teve como alvo reuniões nas quais eram discutidos requisitos de *softwares*.

Yamashita e Ishida [52] pesquisaram os efeitos da tecnologia de tradução de máquina em tarefas colaborativas relacionadas à ordenação de imagens. No experimento realizado, um participante organizava um conjunto de imagem de acordo com as instruções (automaticamente traduzidas) fornecidas por um segundo participante. Posteriormente em Yamashita *et al.* [51] um novo experimento foi conduzido (nos moldes do experimento anterior), porém, utilizando trios de participantes no lugar de duplas, levantando uma novo conjunto de dificuldades.

Os trabalhos anteriormente apresentados demonstram que não houve uma diferença significativa entre a utilização do inglês como um idioma comum e da utilização de serviços de tradução simultânea durante a intermediação da comunicação entre participantes (e equipes) multilíngues. Este fato é atribuído, de forma unânime, aos problemas de tradução encontrados durante a utilização dos serviços de tradução automática.

Estes problemas (traduções incorretas ou inconsistentes) foram apontados como a principal causa das dificuldades no estabelecimento de um conhecimento comum entre os participantes (*common ground* [51]), levando a atrasos na comunicação uma vez que era necessário um maior

número mensagens de esclarecimento para que os participantes pudessem compreender de forma clara a informação passada pelos demais participantes.

Uma análise (qualitativa) dos registros de comunicação do experimento de Calefato *et al.* [7] (apresentada em Hilgert *et al.* [19]) demonstrou problemas de tradução ordem estrutural (referentes à reestruturação incorreta de sentenças) e relacionados ao vocabulário (traduções incorretas ou inconsistentes).

Em relação à traduções incorretas foram identificados casos nos quais essas foram provocadas por características da sentença de entrada como, por exemplo, erros de digitação e abreviações. Como exemplo deste tipo de problemas pode ser mencionada a abreviação das palavras "*ring tone*" (abreviado como "*ring*") e "*(*bluetooth)" (abreviado como "*blue*") para as quais foram geradas, respectivamente, as traduções *"anel"* e "*azul*", ambas corretas em relação a versão abreviada, porém incorretas de acordo com o contexto.

Outro problema identificado decorre da tradução inconsistente de palavras. Considera-se neste trabalho como inconsistente a atribuição de diferentes traduções para uma mesma palavra dados contextos iguais. Como exemplo de tradução inconsistente pode ser destacado o termo "*release*"(versão), para o qual foram atribuídas três diferentes traduções: "*versão*", "*lançamento*", "*entrega*", "*liberação*". Foram, ainda, identificados casos para os quais o termo não foi traduzido. Por fim, observou-se, ainda, a ocorrência frequente de erros de digitação (*typos*) sendo que, para o termo "*Bluetooth*"(tecnologia para transmissão de dados), por exemplo, foram encontradas 5 diferentes grafias: "*bluetooth*", "*blutooth*" ,"*bluetoth*", "*blutoofh*""*bluetooh*". Estes erros, apesar de simples, podem vir a ter impacto na qualidade da tradução e mesmo no entendimento dos participantes.

## 2.3 Motivação

Os problemas anteriormente mencionados dificultam a adoção de ferramentas e serviços de tradução de máquina para a intermediação da comunicação entre equipes multilíngues. Sendo este tipo de tecnologia apontado como uma solução promissora para a falta de profissionais proficientes na língua inglesa, buscou-se neste trabalho colaborar para o aumento de qualidade destes servições e ferramentas.

Tendo sido observado que os problemas relacionados ao vocabulário (traduções incorretas ou inconsistente) ocorreram mais frequentemente e causaram maiores impactos ao entendimento entre os participantes, optou-se, inicialmente, por buscar soluções para esses.

Dentre as possíveis soluções apontadas na bibliografia, destaca-se a construção de vocabulários bilíngues específicos do domínio [37]. Apesar deste objetivo ser encontrado em outros trabalhos [12][11][49], esses não tem como foco a especialização do vocabulário para um domínio específico.

Sendo assim, este trabalho propõem um processo para auxiliar na extração de vocabulário multilíngue inglês-português, tendo como foco (domínio) manuais de *software* do nível de usuário.

## 2.4 Objetivo Geral

Dado o contexto do projeto e os estudos experimentais iniciais, este trabalho tem como objetivo auxiliar no tratamento dos problemas de tradução relacionados ao vocabulário.Neste sentido, propomos o estudo e desenvolvimento de um processo para a extração de vocabulário multilíngue a partir de textos paralelos de um domínio específico.

Assim, pretende-se colaborar com o projeto mediante a utilização do processo proposto na extração de vocabulário multilíngue a partir de documentação de *software*, sendo que neste trabalho foram investigados manuais de usuário multilíngues, principal fonte de material encontrada dentro do domínio abordado.

### 2.4.1 Objetivos Específicos

A extração de vocabulário multilíngue foi mapeada nos seguintes objetivos específicos:

- Levantamento de materiais relacionados à documentação de *software*;
- Construção de um corpus paralelo para os idiomas inglês-português (Brasil);
- Investigação de métodos e técnicas para extração de vocabulário multilíngue a partir de corpora paralelo;
- Proposta de um método para a extração de vocabulário multilíngue a partir da documentação de *software*;
- Avaliação dos resultados obtidos a partir da execução do processo;

O objetivo inicial deste trabalho era a construção do vocabulário bilíngue relacionado às práticas usuais de Engenharia de *Software*, no entanto, devido à escassez de material encontrado, optou-se por focar na extração de vocabulário a partir de manuais de usuário (maior número de documentos paralelos), que constituem uma das principais fontes de vocabulário do domínio de aplicação de *softwares*.

### 2.4.2 Escopo do Trabalho

Como pode ser identificado mediante a descrição do projeto, o mesmo pode ser dividindo em duas principais linhas de pesquisa, sendo uma relacionado à experimentação [8] [7] com tecno-

logias relacionadas à tradução de máquina e a outra à pesquisa linguística para melhoramentos ou especializações das tecnologias empregadas.

Este trabalho se encontra inserido na fase inicial da pesquisa linguística, estando voltado à coleta e construção de recursos a serem empregados nas etapas futuras do projeto. A Figura 2.1 apresenta o escopo do trabalho proposto.

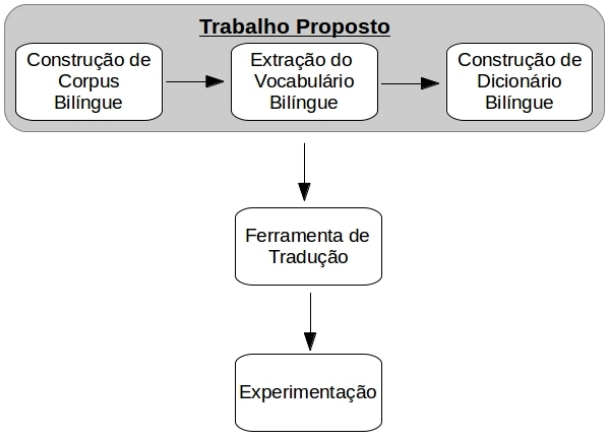


Figura 2.1 – Escopo do trabalho proposto.

Como pode ser visto na Figura 2.1, o trabalho proposto é composto pelas seguintes etapas:

1. Construção do Corpus: Nesta etapa foi construído um corpus paralelo português-inglês composto por manuais de *software*;

2. Extração do Vocabulário Bilíngue: A partir da aplicação do processo proposto sobre o corpus criado, foi extraído um vocabulário bilíngue;

3. Construção de Dicionário Bilíngue: O vocabulário bilíngue extraído foi filtrado eliminando símbolos não úteis ao dicionário como, por exemplo, símbolos de pontuação e numerais. Posteriormente as entradas mais relevantes do vocabulário bilíngue restante foram formatadas de acordo com a estrutura dos dicionários do bilíngues do *Apertium*.

Assim, a saída do processo proposto consistem em um dicionário bilíngue, construído a partir do vocabulário extraído dos manuais de *software*. O dicionário criado pode tanto ser utilizado de forma individual em conjunto com a ferramenta *Apertium* [15] ou utilizado para expandir os dicionários padrão da ferramenta com vocabulário específico de um domínio (tendo em vista que os dicionários padrão são de domínio genérico).

Como o corpus encontrado é escasso, o processo aqui apresentado tem como um dos objetivos permitir sua aplicação sobre os diferentes tipos corpus que possam vir a ser construídos no decorrer do projeto.

No entanto, vale ressaltar que a avaliação do vocabulário extraído mediante sua utilização na ferramenta selecionada (avaliação extrínseca), bem como a realização de experimentos utilizando

o vocabulário em questão, são atividades que encontram-se fora do escopo deste trabalho, sendo executada em outras etapas do projeto.

O vocabulário, bem como os demais materiais levantados durante a execução deste trabalho, podem ser utilizados em outras etapas do processamento linguístico como, por exemplo, a construção automática de corpora multilíngues, construção de estruturas hierárquicas e ontologias multilíngues, entre outras. No entanto, como anteriormente ressaltado, o escopo deste trabalho termina na geração dos dicionários, sendo a utilização dos recursos realizada durante outras etapas do processo.

# 3. LEVANTAMENTO DE RECURSOS

Neste capítulo serão apresentados os recursos linguísticos levantados durante a primeira etapa deste trabalho, tendo como foco principal a apresentação do corpus bilíngue compilado a partir de manuais de *softwares*.

O corpus aqui apresentado foi compilado com o objetivo de ser utilizado como recurso linguístico inicial para estudos relacionados à extração de vocabulário multilíngue e para a avaliação do processo de extração proposto neste trabalho.

Apesar do foco na construção de um corpus, no decorrer da busca por documentos bilíngues, outros recursos considerados como úteis ao escopo do trabalho foram encontrados como, por exemplo, vocabulários da área de Engenharia de *Software* e registros de comunicação entre desenvolvedores.

O capítulo encontra-se organizado de modo que, inicialmente, é apresentado o corpus paralelo construído e utilizado durante este trabalho (Seção 3.1). A seguir, a Seção 3.2 apresenta um conjunto de materiais complementares relevantes ao projeto que foram levantados durante a busca por documentos para o corpus. Por fim (Seção 3.3), são justificadas as principais escolhas tomadas durante a construção do corpus.

## 3.1 Corpus Bilíngue

Um corpus pode ser definido como uma coleção de textos, compilada de acordo com critérios de seleção preestabelecidos, de modo a torná-la empregável na pesquisa linguística [45] [41] [3]. Este tipo de recurso é imprescindível a pesquisas na área de Processamento da Linguagem Natural (PLN) sendo utilizado em diferentes atividades das quais destaca-se, no contexto deste trabalho, a extração de vocabulário multilíngue.

No que tange à questão multilíngue, um corpus pode ser classificado como paralelo ou comparável [14][36]. Corpus paralelo (também conhecido como *bitext*) consiste em um conjunto de textos acompanhados por suas respectivas traduções para outros idiomas. Corpus comparável, por sua vez, consiste em um conjunto de textos em diferentes idiomas que compartilham determinadas características como, por exemplo, assunto (tema), autor, época, entre outras.

Dentre os corpus multilíngues encontrados, o único relacionado ao domínio de documentos de *software* foi o *OPUS* [49]. No entanto, apesar de ser um corpus grande em número de palavras (aproximadamente 3 milhões) esse é composto apenas por dois manuais. Assim sendo, optou-se pela construção de um novo corpus paralelo.

### 3.1.1 Corpus Construído

O objetivo inicial desta etapa foi a construção de um corpus a partir da documentação relacionada a projetos de desenvolvimento de *software* como, por exemplo, documento de requisitos, documento de análise, documento de projeto, registros (*logs*) de reuniões multilíngues, entre outros.

No entanto, devido a pequena quantidade de material técnico bilíngue encontrado, optou-se por recorrer aos manuais de usuário, mais facilmente encontrados em múltiplos idiomas, sendo priorizados manuais de projetos de código livre (*open source*).

Em projetos *open source*, manuais de usuário são construídos e traduzidos de forma colaborativa por usuários, muitas vezes, geograficamente dispersos. Neste processo, inicialmente, uma versão em inglês do manual é disponibilizada para posteriormente ser traduzida para diferentes idiomas.

A utilização de manuais de usuário de projetos *open source* para a construção de corpus paralelo não é novidade, tendo sido anteriormente utilizada por Tiedemann no *OPUS Corpus* [49], tendo como principal vantagem a possibilidade de disponibilização posterior desse, assim como dos resultados do seu processamento, devido ao tipo de licenças utilizadas nestes projetos.

Esta decisão foi tomada de modo a consolidar, inicialmente, o processo de extração do vocabulário multilíngue enquanto a busca por material de nível técnico é realizada de forma paralela para utilização em futuras etapas do projeto.

As dificuldades associadas à construção de corpus para domínios específicos, principalmente quando relacionados a trabalhos multilíngues (especialmente para documentos paralelos), é destacada na literatura por diferentes pesquisadores [45] [14] [36].

Para a construção do corpus, optou-se pela utilização de uma abordagem manual devido ao baixo desempenho obtido durante o teste de abordagens automáticas (utilização da ferramenta *Bootcat* [2]).

Os documentos componentes do corpus construídos, assim como a quantidade de palavras (segunda e terceira colunas) componentes desses e seus respectivos domínios são apresentados na Tabela 3.1. Vale ressaltar que o cálculo do número de palavras foi realizado sobre os arquivos originais, sem que nenhum tipo de filtragem ou pré-processamento tenha sido utilizado.

Em relação ao tamanho do corpus construído, pode-se observar, de acordo com a Tabela 3.2, que esse é similar ao dos corpora utilizados nos trabalhos relacionados, com exceção ao utilizado por Tiedemann [49].

Em relação à disparidade no tamanho do corpus utilizado por Tiedemann (2,8 milhões de palavras) deve-se ressaltar que aproximadamente 2,6 milhões das palavras utilizadas pelo referido autor foram retiradas a partir dos arquivos de tradução do ambiente gráfico *KDE* (mesmo documento), logo, existindo um viés no que tange a extração de vocabulário.

A Tabela 3.3 apresenta o número de sentenças componentes do corpus para cada um dos idiomas envolvidos.

Tabela 3.1 – Tamanho do corpus construído em número de palavras.

| Manual | Inglês | Português | Domínio |
|---|---|---|---|
| Android 2.3.4 | 72.371 | 81.412 | Sistemas Operacionais |
| Blender 2.6 | 31.661 | 25.671 | Edição Gráfica |
| Debian | 149.709 | 140.367 | Sistemas Operacionais |
| LibreOffice 3.3 | 100.978 | 108.450 | Edição de Documentos |
| Slackbook | 65.549 | 64.744 | Sistemas Operacionais |
| TortoiseSVN 1.7 | 79.008 | 79.696 | Versionamento |
| Ubuntu | 54.057 | 55.687 | Sistemas Operacionais |
| **Total** | **553.333** | **556.027** | - |

Tabela 3.2 – Corpora em trabalhos relacionados.

| Trabalho | Idiomas | Número de Palavras |
|---|---|---|
| Ha et al. [14] | Inglês/Espanhol | 31.498/30.344 |
| Caseli e Nunes [11] | Português/Inglês | 532.121/494.391 |
| Kwong et al. | Chinês/Inglês | 113.000/120.000 |
| Tiedemann [49] | Português/Inglês | 2.9 milhões/2.8 milhões |

Tabela 3.3 – Tamanho do corpus construído em número de sentenças.

| Manual | Inglês | Português |
|---|---|---|
| Android 2.3.4 | 4.023 | 4.184 |
| Blender 2.6 | 2.125 | 1.687 |
| Debian | 6.577 | 6.577 |
| LibreOffice 3.3 | 5.591 | 5.576 |
| Slackbook | 3.332 | 3.292 |
| TortoiseSVN 1.7 | 4.560 | 4.500 |
| Ubuntu | 2.806 | 2.803 |
| **Total** | **29.770** | **28.619** |

Como pode ser observado na Tabela 3.3, o número de sentenças dos textos em português (segunda coluna) e inglês (terceira coluna) é similar, sendo que a diferença foi resolvida posteriormente durante o alinhamento sentencial mediante a atribuição de alinhamentos *n:m* ou o descarte de sentenças.

## 3.2 Materiais Complementares

Nesta seção serão apresentados outros tipos de materiais linguísticos encontrados durante a construção do corpus. Apesar de não terem sido empregados diretamente no processo proposto (da mesma forma que o corpus), esses serviram como material de apoio durante etapas como, por exemplo, a avaliação do vocabulário extraído em relação à terminologia de domínio.

Esta encontra-se organizada de modo que a Seção 3.2.1 apresenta um conjunto de termos extraídos a partir de glossários de livros técnicos de Engenharia de *Software* e de documentos técnicos publicados por agências como, por exemplo, a *IEEE*. A Seção 3.2.2 apresenta um conjunto de

documentos técnicos referentes a diferentes etapas do processo de desenvolvimento de *software*. Por fim, a Seção 3.2.3 apresenta um conjunto de registros de comunicação entre equipes desenvolvidas.

### 3.2.1 Glossários

Os glossários encontrados [22] [26] consistem em listas de palavras (simples e compostas) empregadas nas diferentes práticas da Engenharia de *Software*, acompanhadas por suas respectivas definições.

O principal atrativo deste material é o fato de ter sido construído por especialistas da área e compilado por instituições como a IEEE (*Institute of Electrical and Electronics Engineers*), podendo ser considerado como relevante e confiável.

A Tabela 3.4 apresenta os vocabulários encontrados bem como a quantidade de termos componentes desses (segunda coluna).

Tabela 3.4 – Glossários da Engenharia de *Software*.

| **Glossário** | **Termos** |
|---|---|
| System and Software Engineering - Vocabulary [26] | 3.349 |
| Standard Glossary of Software Engineering Terminology [22] | 1.300 |

Os vocabulários apresentados na Tabela 3.4 foram encontrados apenas para o idioma inglês. Mesmo monolíngues, estes vocabulário podem ser utilizados como sementes (*seeds*) para a construção automática de corpus multilíngue (etapas futuras do projeto).

Outra fonte identificada para a extração de vocabulário da área foram o glossários e listas de assuntos de livros relacionados à Engenharia de *Software* como, por exemplo, o livro "*Software Engineering*"[46] do qual foram extraídos 167 termos a partir do glossário e 1.600 a partir da lista de assuntos.

No entanto, a utilização deste tipo de recurso implica questões relacionadas a diretos autorais (*copyright*), bem como a dificuldade de obtenção de cópias digitais desses materiais para a língua portuguesa.

Outra possível utilização destes glossários é na ampliação dos dicionários morfológicos utilizados por ferramentas de extração morfológica do processo de extração de vocabulário (Seção 6) para a identificação de expressões multipalavras.

### 3.2.2 Padrões Técnicos

Além dos vocabulários apresentados na Seção 3.2.1, foram encontrados padrões técnicos relacionados às diferentes etapas do processo de desenvolvimento de *software*, também compilados por instituições como a IEEE, a ISO, e outras. A Tabela 3.5 apresenta exemplos desses.

Tabela 3.5 – Exemplos de padrões técnicos da Engenharia de *Software*.

| **Padrão** | **Etapa** |
|---|---|
| IEEE Std 730-2002 [24] | Descreve o formato e o conteúdo de plano para a garantia da qualidade de software. |
| IEEE Std 1058-1998[23] | Descreve o formato e conteúdo para planos de gerência de software. |
| IEEE Std 828-2005 [25] | Especifica o conteúdo de um plano de gerenciamento de configuração do processo de desenvolvimento software, assim como da atividade de especificação de requisitos. |

Os padrões apresentados na Tabela 3.5 foram selecionados a partir da lista apresentada no Apêndice C do SWEBOK (*Software Engineering Book of Knwoledge*) [21] disponível no formato eletrônico.

Assim como no caso dos vocabulários (Seção 3.2.1), apenas versões em inglês dos padrões técnicos puderam ser encontradas. Sendo assim, esses não puderam ser utilizados na compilação do corpus construído.

### 3.2.3 Registros de Comunicação

Levando em consideração o fator comunicação entre equipes (escopo do projeto) outro recurso buscado nesta primeira etapa foram registros de comunicação entre equipes distribuídas, preferencialmente envolvendo diferentes idiomas. O objetivo da coleta deste tipo de recursos foi a identificação do vocabulário utilizado em situações reais de comunicação durante processos de desenvolvimento.

Assim como nos recursos anteriormente apresentados, optou-se pela pesquisa em projetos de desenvolvimento *open source*, devido à maior disponibilidade de material por parte destes projetos bem como ao tipo de licença de utilização adotada por esses.

Como resultados desta busca, dois tipos de registros foram encontrados: síncronos (mensagens instantâneas) e assíncronos (listas de discussão). Registros de comunicação assíncrona encontrados constituem-se, basicamente, por listas de *e-mail* (*mail list*) coletadas a partir de diferentes projetos.

Em relação aos registros de comunicação síncronos (mensagens instantâneas), esses foram obtidos a partir de gravações de conversas entre desenvolvedores utilizando o IRC (*Internet Relay Chat*), protocolo de mensagens instantâneas na Internet.

Entre os registros síncronos, destaca-se o registro de comunicação entre desenvolvedores participantes de projetos da fundação *Mozilla*, mais especificamente do *Thunderbird* (gerência de *e-mails*) composto por aproximadamente 161.316 mensagens e 1.669.000 palavras.

Listas de discussão podem ser obtidas em diferentes níveis, variando do nível mais leigo, no qual usuários discutem problemas encontrados no *software* e requisitam novas funcionalidades, até

o nível mais especialista, no qual desenvolvedores discutem aspectos do desenvolvimento do *software*. Das listas encontradas, destaca-se a do Debian (distribuição Linux) que possui um repositório contendo aproximadamente 18 anos de registros de discussão para múltiplos idiomas.

Quanto ao critério multilíngue, registros de comunicação síncrona foram encontrados somente para o idioma inglês, enquanto registros de comunicação assíncrona foram encontrados para diferentes idiomas.

Mediante uma análise do vocabulário encontrado nos registros levantados, pode-se constatar uma maior incidência de termos do domínio de aplicação do *software* do que do vocabulário da Engenharia de Software (obtido a partir das listas apresentadas na Seção 3.2.1).

Por exemplo, a partir do vocabulário extraído dos registros do experimento de Calefato *et al.* [7], cujo domínio foi telefonia móvel, termos como "*ring tone*" e "*Bluetooth*" apresentaram maior incidência do que termos da Engenharia de *Software* como, por exemplo, "*release*" (versão).

## 3.3 Corpus Utilizado na Pesquisa

Neste capítulo foram apresentados quatro tipos de recursos levantados: (1) corpus paralelo construído a partir de manuais de *software*, (2) conjunto de padrões técnicos, (3) vocabulários monolíngues da área de Engenharia de *Software* e (4) registros de comunicação entre desenvolvedores.

Dos recursos anteriormente listados, os conjuntos de padrões técnicos e os vocabulários de domínio são os que apresentam uma maior relação com a área de Engenharia de *Software*, no entanto, apenas versões monolíngues para o idioma inglês puderam ser encontradas, sendo assim, esses não puderam ser utilizados no processo de extração do vocabulário multilíngue.

A comparação do vocabulário coletado com as listas de termos extraídos a partir dos registros de comunicação demonstrou um baixo índice de intersecção, o que indica que o vocabulário técnico é pouco empregado nas conversas entre desenvolvedores, enquanto o vocabulário do domínio de aplicação é mais frequentemente empregado (o termo"agenda" para o domínio de telefones móveis, por exemplo).

A principal utilidade dos registros de comunicação coletados, foi a identificação do tipo de vocabulário mais utilizado durante a comunicação entre os desenvolvedores, sendo que apenas registros assíncronos comparáveis puderam ser encontrados para múltiplos idiomas, e sua utilização na construção de corpora não foi possível devido a dificuldades de alinhamento.

Sendo assim, optou-se pela compilação e utilização de um corpus multilíngue (português-inglês) a partir de documentos de *software* (a nível de usuário) para estudos da extração de vocabulário multilíngue e avaliação do processo proposto. A construção de outros corpora, de domínio mais técnico, está prevista para etapas futuras do projeto, tendo sido priorizada neste primeiro momento, a consolidação do processo proposto.

# 4. EXTRAÇÃO DE VOCABULÁRIO MULTILÍNGUE

Vocabulários multilíngues (também conhecidos como *translation lexicons*) são recursos de grande importância para pesquisas na área de Processamento de Linguagem Natural(PLN) e fundamentais para estudos em tradução de máquina[11] [47] [36] [43].

Este tipo de recurso linguístico especifica relações de correspondência entre palavras de diferentes idiomas, sendo aplicado em outras áreas de pesquisa multilíngues como, por exemplo, na construção de corpora [32] [2], recuperação de informações, tradução assistida por computadores [9], entre outras.

Neste capítulo serão apresentados os principais conceitos relacionados ao processo de extração de vocabulário multilíngue. Inicialmente (Seção 4.1), um processo geral (baseado na literatura) é apresentado, sendo suas etapas descritas de modo que a Seção 4.2 apresenta a etapa de alinhamento sentencial, a Seção 4.3 a análise morfológica e a Seção 4.4 a etapa de alinhamento léxico. A Seção 4.5 apresenta a estratégia empregada para o tratamento de expressões multipalavras e, por fim , são apresentadas estratégias para a avaliação de léxicos bilíngues (Seção 4.6).

## 4.1 Processo de Extração

A construção de recursos multilíngues (vocabulário, terminologia, modelos de tradução, regras de tradução, etc.) encontra-se bem consolidada na literatura, não podendo ser considerada como algo inédito [12][47][14][53].

Sendo assim, a partir dos trabalhos relacionados, com destaque para os trabalhos de Caseli e Nunes [12], Tiedemann [47] e Ha *et al.* [14], um processo genérico para a extração de vocabulários bilíngues foi identificado. Esse é composto por 3 etapas:

1. Alinhamento Sentencial;
2. Análise Morfológica;
3. Alinhamento Lexical;

De acordo com o processo, os documentos componentes de um corpus paralelo são decompostos em sentenças para as quais são buscadas correspondências entre os documentos considerados paralelos (diferentes idiomas). Esta atividade é conhecida como alinhamento sentencial.

Após alinhadas, as sentenças são decompostas em suas unidades léxicas (palavras, expressões multipalavras, etc.) para as quais correspondências são buscadas na sentença correspondente (alinhada) do documento paralelo (alinhamento lexical).

No entanto, como muitas vezes uma palavra (ou grupo de palavras) pode estar relacionada à mais de uma possível tradução, é importante que se estabeleça uma forma de diferenciar o contexto

no qual cada uma dessas é empregada (desambiguação). O objetivo da anotação morfológica (segunda etapa) é fornecer informações (rótulos morfossintáticos) que permitam ao alinhador lexical realizar esta distinção, ainda durante o alinhamento lexical.

Após a etapa de alinhamento lexical, os trabalhos relacionados divergem, de acordo com seu objetivo, em relação à utilização dos recursos produzidos. Das diferentes utilidades desses, destacam-se o auxílio à extração terminológica [14][53], a indução de léxicos bilíngues [12] [47] e a construção de dicionários multilíngues, como apresentado dos principais trabalhos relacionados.

Além das etapas anteriormente citadas, costuma-se, ainda, realizar uma etapa de pré-processamento do corpus, na qual são executadas atividades como, por exemplo, conversão de formato de arquivos e codificação dos textos, remoção de ruídos(símbolos indesejados), separação de sentenças e unidades léxicas, entre outras [48].

O processo apresentado é considerado genérico pois, além de comumente encontrado em trabalhos da área, diferentes abordagens podem ser adotadas na implementação das etapas que o compõem, como melhor exemplificado nas seções posteriores.

## 4.2 Alinhamento Sentencial

O alinhamento sentencial consiste no estabelecimento de correspondências (*links*) entre sentenças de textos paralelos [47][12],ou seja, em determinar possíveis traduções para essas.

A Tabela 4.1 apresenta exemplos de alinhamentos sentenciais extraídos a partir de versões paralelas (português-inglês) do manual de usuário do *Android* (sistema operacional de dispositivos móveis).

Tabela 4.1 – Exemplos de alinhamento sentencial.

| **Inglês** | **Português** |
|---|---|
| "Checking the time and setting alarms." | "Como verificar o horário e definir alarmes." |
| "Calculating the solutions to math problems." | "Como calcular soluções de problemas matemáticos." |
| "Widgets are applications that you can use directly on the Home screen." | "Widgets são aplicativos que podem ser utilizados diretamente na tela "Página inicial"." |

As sentenças demonstradas na Tabela 4.1 foram retiradas de textos paralelos sentencialmente alinhados. Sendo assim, presumindo-se o alinhamento como correto, pode-se estabelecer a sentença *"Como calcular soluções de problemas matemáticos"* (português) como uma provável tradução da sentença *"Calculating the solutions to math problems"* (inglês).

Em relação a esta etapa, duas principais questões devem ser consideradas: (a) tipos dos alinhamentos; (b) métodos de alinhamento. Estas questões serão apresentadas nas subseções que seguem.

### 4.2.1 Tipos de Alinhamento

O alinhamento sentencial nem sempre é realizado entre exatamente uma sentença do texto fonte e uma sentença do texto alvo (alinhamento do tipo *1:1*), podendo variar em diferentes tipos como, por exemplo, *1:2* (uma sentença do texto fonte com duas do texto alvo), *2:1* (duas sentenças do texto fonte com uma sentença do texto alvo), entre outras possíveis combinações [47][12].

No entanto, como demonstrado em [11], a maioria dos alinhamentos são do tipo *1:1*. Em [11], Caseli demonstrou que, em alinhamentos entre sentenças do português e do inglês a taxa de alinhamentos *1:1* foi de 93,97%enquanto para alinhamentos português-espanhol foi de 98,32% (manualmente revisados).

Existem, ainda, alinhamentos do tipo *1:0* e *0:1*, conhecidos como alinhamentos vazios (*empty*)[47] [35] ou omissões [12], que ocorrem quando não é possível determinar as equivalências de uma determinada sentença entre os idiomas envolvidos.

Sentenças com alinhamento vazio costumam ser descartadas durante o alinhamento sentencial, sendo assim, nem sempre o número de sentenças alinhadas será o mesmo que o número total de sentenças dos textos originais.

### 4.2.2 Métodos de Alinhamento

Sendo inviável o alinhamento manual de corpora e tendo-se como objetivo a automatização do processo, métodos automáticos (assim como ferramentas) foram propostos[35] [10] [16] [20] [38]. Os métodos de alinhamento sentencial pesquisados [35] [47] [38] baseiam-se em três abordagens básicas:

- Tamanho das sentenças (*length-based*);
- Conteúdo das sentenças (*lexical-based*);
- Híbridas, combinação das duas abordagens anteriores;

Abordagens baseadas no tamanho da sentença (*length-based*) [16] partem do princípio de que existe uma relação estatística entre o tamanho de sentenças paralelas de determinados idiomas [47].

Esta relação pode ser calculada tanto a partir do número de caracteres, quanto do número de unidades léxicas (palavras, números, sinais de pontuação, etc.) componentes das sentença. Destas alternativas, Gale e Church [16] relatam ter obtido melhores resultados com cálculos baseados no número de caracteres.

Abordagens baseadas no conteúdo das sentenças (*lexical-based*) exploram os símbolos (palavras, numerais, etc.) componentes dessas [35]. Em relação ao tipo de informações utilizadas,

essas podem ser baseadas em: (a) dicionários bilíngues (*dictionary based*), (b) símbolos âncora (*anchor based*).

Métodos baseados em dicionários utilizam dicionários bilíngues iniciais (domínio geral) para realizar uma tradução de palavras conhecidas (existentes no dicionário) da sentença fonte, gerando uma tradução inicial dessa (ainda que pouco precisa). Esta tradução inicial é então comparada, mediante cálculo de similaridade, com as prováveis sentenças alvo.

A principal desvantagem desta técnica é a necessidade da utilização de dicionários bilíngues externos, os quais podem nem sempre encontrar-se disponíveis.

Métodos baseados em símbolos âncora (*anchor*) [10] calculam a similaridade de duas sentenças de acordo com a ocorrência desses. Símbolos âncora são símbolos que mantêm sua forma original (não são traduzidos) em diferentes idiomas como, por exemplo, números e palavras cognatas.

Os símbolos âncora podem ser obtidos tanto a partir de listas externas [10], quanto a partir das próprias sentenças alinhadas, mediante algoritmos como sequência comum mais longa (*Longest Common Sequence*), por exemplo.

Por fim, abordagens híbridas [35][10] combinam (de diferentes formas) as duas abordagens anteriormente apresentadas. No *Bilingual Sentece Aligner* [35], por exemplo, o alinhamento baseado no tamanho das sentenças é inicialmente aplicado para posteriormente ser refinado utilizando-se um alinhamento baseado no conteúdo das sentenças.

Torna-se importante conhecer os métodos de alinhamentos pois estes delimitarão o tipo de pré-processamento que será empregado sobre o corpus.

Métodos baseados em símbolos âncora, por exemplo, são diretamente influenciados pela etapa de remoção de ruído. A remoção de determinados símbolos (como número, por exemplo) impacta de forma direta no desempenho destes métodos.

Já métodos baseados no tamanho da sentença são diretamente influenciados pelo procedimento de separação de sentenças. Sentenças incorretamente separadas (divididas ao meio, por exemplo) tem seu tamanho alterado, influenciando diretamente no estabelecimento de correspondências.

### 4.2.3 Ferramentas

No decorrer dos estudos relacionados ao alinhamento sentencial, um conjunto de ferramentas foi identificado, destacando-se entre essas as apresentadas na Tabela 4.2.

Tabela 4.2 – Alinhadores sentenciais levantados.

| **Ferramenta** | **Método** |
|---|---|
| Bilingual Sentence Aligner [35] | Híbrido |
| TCAalign [10][17] | Híbrido |
| Hunalign [50] | Híbrido ou Baseado em Dicionário |

Além das ferramentas apresentadas pela Tabela 4.2 foram identificados, ainda, o *WinAlign* (utilizado por Ha *et al.*[14]) e o *GMA* (*Geometric Mapping and Alignment*)[34]. No entanto, estas ferramentas não foram consideradas neste trabalho.

*Bilingual Sentence Aligner*

O *Bilingual Sentence Aligner* [35] (ferramenta empregada por este trabalho), utiliza uma abordagem híbrida, valendo-se tanto do tamanho da sentença quanto do conteúdo da mesma (palavras), sendo independente de conhecimento prévio das línguas envolvidas.

O método proposto por Moore é composto por três passos:

1. Um alinhamento sentencial inicial é realizado, utilizando uma abordagem baseada no tamanho das sentenças (em relação à quantidade de caracteres), adaptado a partir do trabalho de Brown *et al.* [5];
2. Os alinhamentos de maior pontuação (atribuída de acordo com o cálculo de tamanho) são utilizados para o treinamento de um modelo estatístico, no qual um alinhamento lexical inicial é realizado entre as unidades léxicas componentes da sentença, criando assim um dicionário bilíngue inicial;
3. O dicionário criado durante o passo anterior é utilizado em um realinhamento das sentenças, agora baseado no conteúdo sentencial.

O *Bilingual Sentence Aligner* foi testado sobre um corpus formado por manuais de *software*[35], domínio relacionado ao presente trabalho. Os testes, nos quais foram utilizadas as métricas de *Precision Error* e *Recall Error*, demonstrou que a abordagem híbrida utilizada obteve um desempenho melhor do que a abordagem baseada somente no tamanho sentencial sem um aumento substancial no tempo.

Em um dos casos de avaliação realizados, 300 sentenças foram aleatoriamente removidas de um dos textos. Nesta avaliação os valores de *Precision Error* e *Recall Error* da abordagem híbrida foram, respectivamente, *0,042%* e *0,052%*, enquanto para a abordagem baseada no tamanho sentencial esses foram de *0,552%* e *1.967%*.

A ferramenta permite, ainda, a definição de um valor *threshold* (valor entre 0 e 1) que delimita um valor mínimo para a probabilidade de alinhamento entre duas sentenças para que estas sejam consideradas alinhadas. Em relação a este parâmetro, quanto maior o valor utilizado, maior a acurácia dos alinhamentos produzidos, no entanto, o número de alinhamentos tende a diminuir.

*TCAalign*

O *TCAalign* [10], baseado no *Translation Corpus Aligner* [20], é uma ferramenta construída durante o projeto PESA(*Portuguese-English Sentence Alignment*) e utilizada em Caseli e Nunes [12] para alinhamentos português-inglês/inglês-português e português-espanhol/espanhol-português.

Uma das características do *TCAalign* é a possibilidade da utilização de listas de palavras âncoras durante o alinhamento sentencial. Ainda que opcional, a utilização deste recurso é apontada como um dos fatores que melhoram o desempenho do sistema [12].

Apesar de constituírem um recurso externo, além de sua utilização ser opcional, listas de palavras cognatas para a dupla português-inglês, produzidas durante o projeto PESA, encontram-se disponíveis no sítio do projeto[1].

Os alinhamentos realizados de forma automática durante os experimentos apresentados em [12] foram manualmente revisados e corrigidos. Os alinhamentos corrigidos foram posteriormente utilizados como padrão de referência no teste da ferramenta. Após o teste foi observada uma precisão de 97.10% e uma abrangência (cobertura) de 98.23%.

O *TCAalign* possui uma versão para *download* (*standalone*), e outra versão para utilização *online*, conhecida como *VisualTCA*[2] [17].

*Hunalign*

O *Hunalign* [50] é uma ferramenta que emprega abordagens baseadas em dicionários e no tamanho das sentenças (sendo que pode-se optar pela utilização de apenas um destes modos), tendo sido originalmente proposto para o alinhamento do idioma húngaro devido à falta de similaridade desse com os demais idiomas.

Quando utilizado o modo orientado a dicionários, o dicionário inicial é utilizado para a realização de uma tradução inicial do texto, posteriormente comparada com o texto alvo. Esta comparação pode ser conduzida de acordo com o número de palavras ou com o tamanho da sentença.

Quando utilizado o modo orientado ao tamanho, a contagem de caracteres do texto original é incrementada em uma unidade e a pontuação é baseada na variação da mais longa para a mais curta. Os marcadores de limite de parágrafo são tratados como sentenças com pontuação especial.

A pontuação de similaridade é calculada para cada par de sentenças ao redor da diagonal da matriz de alinhamento (pelo menos 500 sentenças vizinhas são calculadas, aproximadamente 10% do total neste caso).

Quando um dicionário inicial não é disponibilizado, um alinhamento 1:1 entre os textos fonte e alvo é estabelecido, de modo a gerar um dicionário inicial. Neste passo, qualquer sentença com probabilidade de alinhamento maior do que 0.5 é considerada. Como pode ser observado, a abordagem empregada neste caso é muito similar à empregada por Moore [35].

Uma das principais vantagens desta ferramenta em relação às demais é a a velocidade de alinhamento, diretamente relacionada à sua implementação na linguagem de programação *C++*, enquanto as demais encontram-se implementadas em *Perl* (linguagem interpretada).

[1] http://www.nilc.icmc.usp.br/nilc/projects/pesa.htm
[2] www.nilc.icmc.usp.brnilctoolspagina-visualtcavisualtcatca.htm

## 4.3 Análise Morfológica

Dependendo do contexto, uma palavra (ou grupo de palavras) pode assumir mais de uma tradução válida, sendo assim, torna-se necessária a utilização de algum tipo de informação para a distinção entre estes contextos (desambiguação). Informações morfológicas são frequentemente empregadas com esta finalidade [12] [14].

Durante etapa de análise morfológica podem ser atribuídos diferentes tipos de informações morfológicas às palavras formadoras de um documento, destacando-se aquelas empregadas no presente trabalho:

- Forma superficial: Forma pela qual as palavras aparecem no texto;
- Lema: Forma canônica, ou seja, palavra desprovida de sufixos modificadores de gênero, número, etc.;
- Classe Gramatical: Classe gramatical à qual a palavra pertence (*Part-of-speech*) por exemplo, substantivo, adjetivo, etc;
- Flexões: Variações de número (singular e plural), gênero(masculino e feminino), entre outras.

A Tabela 4.3 apresenta um exemplo de análise morfológica da palavra "região" ("*region*") para o inglês e para o português, proveniente da saída da ferramenta *Lttoolbox* (analisador morfológico do *Apertium* [15]).

Tabela 4.3 – Exemplos de análise morfológica.

| **Idioma** | **Exemplo** |
|---|---|
| Português | regiões/região <n><f><pl> |
| Inglês | regions/region<n><pl> |

Analisando a primeira entrada "*região*", tem-se:

- regiões: forma superficial;
- região: forma canônica (lema);
- <n> : classe gramatical(morfológica), neste caso *noun* (substantivo);
- <f> : variação de gênero, "feminino" para este caso;
- <pl> : variação de número, "plural" para este caso.

Como observado, pode haver uma diferença entre o número de características disponibilizadas pelos analisadores, sendo que para o inglês o número de informações foi menor. Esta diferença pode ser atribuída a limitações dos dicionários utilizados pelo analisador morfológico.

Além da desambiguação de sentido, estas informações podem, ainda, ser empregadas para outras finalidades, como, por exemplo, no trabalho de Ha *et al.* [14], no qual os rótulos morfológicos (*PoS*) também são utilizados no processo de identificação de terminologia.

Tiedemann [49], por sua vez, utiliza informações morfossintáticas durante a geração de árvores de dependência, empregadas na visualização do corpus, assim como de seus termos.

### 4.3.1 Ferramentas

A Tabela 4.4 apresenta uma lista de ferramentas que podem ser utilizadas para diferentes tipos de anotação morfológica. Na tabela, um "X" representa que a determinada ferramenta apresenta a característica definida pela coluna. Dentre as ferramentas, destaca-se o *Apertium* [15] (descrito na próxima seção), tendo esta sido selecionada para a utilização neste trabalho.

Tabela 4.4 – Ferramentas para anotação morfossintáticas de textos.

| **Ferramenta** | **In** | **Pt** | **Lematização** | **Análise Morfológica** | **Análise Sintática** |
|---|---|---|---|---|---|
| OpenNLP [3] | X | X | - | X | X |
| NLTK [4] | X | X | X | X | X |
| Apertium [15] | X | X | X | X | - |
| Malt [53] | X | X | | X | X |
| HunPos [50] | X | X | - | X | - |
| LX Center | X | - | X | X | - |
| FreeLing [40] | X | X | - | X | X |

Vale ressaltar que várias das ferramentas apresentadas possuem particularidades a serem consideradas. O *Apertium*, por exemplo, depende da utilização de dicionários externos para análise morfológica, enquanto outras ferramentas como o *NLTK* [4] , por exemplo, necessitam que modelos próprios para cada idioma sejam treinados.

Outra particularidade que deve ser considerada é o conjunto de rótulos (*tagset*) utilizados para representar cada tipo de informação. Frequentemente, ferramentas utilizam diferentes padrões como, por exemplo, no caso do *Lttoolbox* que utiliza a letra "n"(minúscula) para representar substantivos, enquanto o *Freeling* utiliza "N"(maiúscula). Sendo assim, os conjuntos de rótulos devem ser padronizados.

#### *Apertium*

O *Apertium* é uma plataforma de código livre (*open source*) para tradução de máquina, sendo composta por um motor de tradução, um conjunto de ferramentas e um conjunto de dados para a implementação de sistemas de tradução de máquina baseada em regras [15].

Dentre as funcionalidades disponibilizadas por esta plataforma, destaca-se, nesta seção, a análise morfológica, conduzida em três etapas [15] [11]:

1. Desformatação: Texto (sentenças) é separado da formatação que o acompanha;

2. Análise Morfológica: O texto de entrada é decomposto em suas unidades léxicas (palavras) para as quais são atribuídas informações (através de rótulos) como, por exemplo, forma superficial, lema, categoria gramatical e flexões (gênero,número, etc.). Nesta etapa são realizadas, ainda, a resolução de casos de contração ("*do*" igual a "*de+o*") e a identificação de expressões multipalavras;

3. Desambiguação Categorial: Unidades para as quais mais de um rótulo gramatical (*PoStag*) foi atribuído são processadas por um etiquetador morfológico (*PoS-tagger*) para a definição do rótulo e ou forma canônica mais adequado ao contexto.

A Tabela 4.5 apresenta um exemplo de sentença anotada morfologicamente pelo *Apertium* [15]. A primeira coluna da tabela apresenta a sentença original enquanto a segunda coluna apresenta o resultado da anotação.

Tabela 4.5 – Exemplo de anotação morfológica

| **Sentença Original** | **Sentença Anotada** |
|---|---|
| "Como monitorar o status do seu telefone." | Como<rel><adv> *monitorar o<det><def><m><sg> status<n><m><sp>de<pr>+o seu<det><pos><m> <sg> telefone<n><m><sg> .<pun> |

Uma relação dos rótulos utilizados pela ferramenta para a anotação e seu significado pode ser obtida em http://wiki.apertium.org/wiki/List_of_symbols.

As etapas de anotação morfológica e desambiguação categorial são realizadas, dentro da arquitetura do *Apertium*, por ferramentas distintas. A análise morfológica é conduzida pelo pacote *Lttoolbox*[4], enquanto a desambiguação categorial é realizada pela ferramenta *apertium-tagger*.

O *Lttoolbox* é um conjunto de ferramentas do *Apertium* para o processamento lexical, e é composto por três ferramentas:

- *lt-comp*: Ferramenta utilizada para a compilação de dicionários morfológicos;
- *lt-proc*: Ferramenta de processamento léxico, utilizada para funcionalidades como, por exemplo, lematização e análise morfológica;
- *lt-expand*: Ferramenta utilizada para a expansão dos dicionários morfológicos. Utilizada principalmente para a visualização completa de seu conteúdo.

O *apertium-tagger*, por sua vez, é um rotulador morfológico estatístico (*statistical Part-of-Speech tagger*) baseado em um modelo de Markov oculto (HMM) de primeira ordem. Para realizar a escolha, esta ferramenta necessita de um modelo treinado especificamente para o idioma

[4] http://wiki.apertium.org/wiki/Lttoolbox

utilizado. Modelos treinados para vários idiomas podem ser obtidos no repositório do *Apertium* e encontram-se identificados pela terminação "*.prob*".

As ferramentas descritas são utilizadas em sequência (*pipeline*) de modo que a saída do analisador morfológico (*Lttoobox*) é fornecida como entrada para o desambiguador categorial *apertium-tagger*.

## 4.4 Alinhamento Lexical

O alinhamento lexical (*word alignment*) consiste no estabelecimento de correspondências entre palavras (ou grupos de palavras) pertencentes a textos considerados como paralelos [47] [48] [12].

Partindo desta definição, torna-se importante ressaltar a diferença entre alinhamento lexical, extração de léxicos bilíngues e extração de terminologia bilíngue[47].

O alinhamento léxico, como previamente apresentado, consiste no alinhamento de todas as unidades léxicas (palavras, sintagmas, sinais de pontuação, etc.) componentes dos textos paralelos [47].

A extração de léxicos bilíngues, por sua vez, tem por objetivo a identificação de traduções de palavras (ou grupos de palavras) específicas, que posteriormente poderão ser utilizadas fora de seu contexto. Sendo assim, são filtrados elementos como, por exemplo, sinais de pontuação, funções gramaticais, traduções incertas, etc.

A extração de um subconjunto de palavras pertencentes a um domínio específico, a partir de um léxico bilíngue, é conhecida como extração de terminologia bilíngue [47].

Segundo Tiedemann [47], existem duas principais abordagens para o alinhamento léxico: (a) associativas e (b) estimativa. Essas são apresentadas nas próximas seções.

### 4.4.1 Abordagens Associativas

As abordagens associativas, também conhecidas como abordagens heurísticas ou abordagens de teste de hipótese [47], constituem uma das principais técnicas utilizadas por lexicógrafos no início dos estudos linguísticos em corpus paralelos.

Este tipo de abordagem, baseado na coocorrência de palavras equivalentes em documentos paralelos, em geral, é composta por três etapas:

- Segmentação léxica: Os limites das unidades léxicas (sentenças, palavras, sintagmas, etc.) são identificados para ambos os idiomas envolvidos;
- Correspondências: Identificação de possíveis associações entre unidades léxicas de acordo com algum critério de correspondência (frequência, contexto, etc.). Nesta etapa, muitas vezes, são

criados dicionários bilíngues nos quais palavras são vinculadas a suas respectivas traduções, utilizando algum tipo de peso para cada vínculo;

- Alinhamento e Extração: Mediante o dicionário criado no passo anterior, são selecionadas as traduções mais confiáveis. Esta extração pode ser realizada através de estratégias "gulosas"(*greedy*) como a " *Best First*", por exemplo.

A etapa de segmentação léxica também pode ser referenciada como *tokenization* e pode ser executada de diferentes formas, sendo a forma mais comum a utilização de espaços em branco como delimitadores de palavras.

No entanto, existem ferramentas específicas para esta finalidade como, por exemplo, as disponibilizadas pelas bibliotecas NLTK [4] e *OpenNLP* [5]. Essas, baseiam-se em modelos treinados que, além de identificar os limites das palavras, são capazes de resolver casos de contração ("*do* = *de+o*") e abreviações.

Após identificados os limites das unidades léxicas, o próximo passo consiste em medir o nível de associação entre as palavras dos documentos. Dentre as medidas mais utilizadas na literatura para o cálculo de associação, destacam-se as medidas de co-ocorrência e as de similaridade de strings.

Medida de Coocorrência

Medidas de coocorrência assumem que palavras equivalentes de textos paralelos co-ocorrem significativamente mais frequentemente devido a caracterizarem um alinhamento do que ao acaso [47]. Sendo assim, para este tipo de medida, cria-se uma hipótese inicial de que as unidades léxicas co-ocorrem ao acaso e, posteriormente tenta-se refutá-la através da utilização de cálculos de associação como o *t-test* e o coeficiente de *Dice*.

Ambas as mediadas de associação estatística são utilizadas de forma que, quanto mais forte a evidência para rejeitar a hipótese de independência, maior os valores resultantes do cálculo.

Abordagens baseadas em medida de coocorrência podem ser vistas em [14] [53]. Ha *et al.*[14] faz uso da medida de *Loglikelihood* enquanto Zhang[53] utiliza um conjunto de medidas: *MI (Mutual Information)* , *Dice*, $X^2$ e *LogLikelihood*.Vale ressaltar que a medida de *LogLikelihood* é utilizada em ambos os trabalhos devido a sua capacidade de capturar associações de baixa frequência [14] [53].

Esta abordagem demanda a utilização de textos alinhados sentencialmente [14] [53] [47], para as quais são construídas tabelas de contingência que servem como base para a aplicação das métricas apresentadas.

A Tabela 4.6 representa um exemplo de tabela de contingência para um termo "*I*" (inglês) e seu suposto termo equivalente "*P* " (português).

Na Tabela 4.6, o valor "*a*" corresponde à quantidade de sentenças nas quais tanto o termo "I" quanto "P" co-ocorrem, "*b*" ao número de sentenças nas quais o termo "P" ocorre mas o termo

[5] http://opennlp.sourceforge.net

Tabela 4.6 – Exemplo de tabela de contingência.

| | **Palavra *I* ocorre** | **Palavra *I* não ocorre** | - |
|---|---|---|---|
| **Palavra *P* ocorre** | a | b | a+b |
| **Palavra *P* não ocorre** | c | d | c+d |
| - | a+c | b+d | *N*=a+b+c+d |

"I" não, o valor "*c*" ao número de sentenças nas quais o termo "I" ocorre e o termo "P" não e, por fim, o valor "*d*" representa a quantidade de sentenças nas quais nenhum dos termos ocorre.

A utilização de abordagens associativas costuma ser empregada pra extração de palavras ou expressões multipalavras (*Multiword Expressions*) previamente conhecidas[47]. Sendo assim, costuma ser precedida por um processo de extração terminológica, como pode ser observado nos trabalhos de Ha *et al.* [14] [53].

A aplicação das métricas anteriormente apresentadas faz uso destes valores para o cálculo de co-ocorrência. O cálculo de *Dice*, por exemplo, utiliza a seguinte fórmula:

$$Dice(E, P) = 2 * a/((a + b) * (a + c))$$

Medidas de Similaridade de *String*

Técnicas relacionadas a medidas de similaridade de *strings*, utilizam-se de medidas como, por exemplo, LCSR (*Longest Common Subsequence Ratio*) , subsequência comum mais longa, para a extração de unidades léxicas similares, também conhecidas como cognatas.

A identificação de cognatas é utilizada principalmente durante o processo de alinhamento sentencial. Dentre os sistemas encontrados, apenas o *LIHLA* [11] faz utilização desta abordagem, ainda que de forma auxiliar.

### 4.4.2 Abordagens Estimativas

Neste tipo de abordagem, modelos probabilísticos (estatísticos) treinados a partir de corpus paralelos são utilizados para a identificação de correspondências entre palavras [47]. Um exemplo de método que utiliza abordagens estimativas é a tradução de máquina estatística (*Statistical Machine Translation*), que consiste na aplicação do modelo do canal com ruído (*noisy channel model*) da teoria da informação à tradução de máquina [28].

Uma das ferramentas mais utilizadas por soluções que adotam esta abordagem é o *Giza++* [39], que implementa os 5 modelos de tradução da IBM [39]. Nesta ferramenta, os modelos de tradução são representados como um conjunto de conexões ocultas (*hidden*) entre palavras do texto fonte e do texto alvo. Nesta abordagem, o treinamento consiste na definição dos pesos das transições entre os estados (palavras).

Em relação a abordagens estimativas, duas questões devem ser levadas em consideração: (1) alinhamentos não simétricos, (2) identificação de expressões multipalavras (*Multi-Word Expressions*).

O fato de os alinhamentos não serem simétricos implica resultados diferentes de acordo com a direção de alinhamento escolhida (português-inglês ou inglês-português). Sendo assim, dados dois textos paralelos nos idiomas português e inglês, o alinhamento de um termo X do texto em inglês com um termo Y do texto em português (na direção inglês-português), não implicará que o termo Y seja alinhado ao termo X se a direção do alinhamento for invertida (português para inglês, por exemplo).

Esta diferença de alinhamento ocorre principalmente em palavras compostas, ou expressões multi-palavras uma vez que neste tipo de alinhamento uma palavra fonte só pode ser alinhada a uma palavra alvo por vez.

Uma das possíveis soluções para este problema são os algoritmos de simetrização propostos por Och e Ney [39] e utilizados por Tiedemann [47] e Caseli e Nunes [12]. Esses realizam a intersecção ou a união entre as matrizes de alinhamento geradas em ambos os sentidos fonte/alvo e alvo/fonte.

Outra técnica utilizada durante o alinhamento de expressões multipalavras é a conexão das palavras formadoras dessas mediante a utilização de símbolos como "_"(*sublinhado*) [12] como em "*telefone_móvel*", por exemplo. Esta conexão deve ser realizada antes do processo de alinhamento léxico.

Existem duas principais desvantagens associadas a esta técnica (conexão de palavras). A primeira consiste na necessidade que as expressões sejam previamente conhecidas, o que exige a utilização de métodos ou ferramentas para a identificação deste tipo de estrutura. A segunda desvantagem é que, por unir as palavras, o número de ocorrências individuais dessas é decrementado, podendo interferir em seu processo de alinhamento.

### 4.4.3 Ferramentas

A partir dos trabalhos relacionados, foram levantadas as ferramentas *Uplug* [49] e Giza++[39] para a realização do alinhamento lexical. No entanto, como o Giza++ é utilizada na maioria dos trabalhos pesquisados, optou-se por sua utilização. A ferramenta selecionada é apresentada na próxima seção.

#### *Giza++*

O *Giza++* [39] é uma ferramenta utilizada para o alinhamento lexical de documentos paralelos, implementando, com esta finalidade, os 5 modelos estatísticos de tradução da IBM [39].

Definido um corpus paralelo, o Giza++ realiza de forma simultânea o treinamento do modelo estatístico de tradução e o alinhamento lexical das palavras (ou grupos de palavras) componentes desses.

A ferramenta permite, ainda, que arquivos de configuração previamente treinados e dicionários bilíngues sejam utilizados como auxílio durante o processo, sendo que estas informações influenciam diretamente na qualidade do alinhamento.

Dicionários externos influenciam tanto no alinhamento das palavras multilíngues quanto na identificação de expressões multipalavras durante o alinhamento. Este, constitui-se de um arquivo indicando uma determinada palavra e seu respectivo equivalente para um segundo idioma.

Em sua configuração padrão, a ferramenta executa 5 iterações dos modelos 1,2 e 4 da IBM seguido pelo processo de alinhamento utilizando Modelos Ocultos de Markov (HMM).

## 4.5 Identificação do Vocabulário

Como apresentado na Seção 4.4, o alinhamento lexical, quando conduzido através de abordagens estimativas, estabelece relações de equivalência entre todas as unidades léxicas de textos paralelos, sem distinguir entre os tipos de símbolos alinhados [47].

Sendo assim, ao final do processo, tem-se uma lista de equivalências multilíngues na qual podem ser encontrados tanto termos de domínio quanto símbolos de menor importância (para o contexto do trabalho) como sinais de pontuação, por exemplo. Logo, é importante que sejam utilizados métodos que distinguam entre as entradas de acordo com sua importância, sendo o vocabulário do domínio priorizado.

Ainda no que tange a identificação do vocabulário, deve-se considerar expressões multipalavras (*Multi-Word Expressions*), combinações de palavras para as quais as propriedades sintáticas ou semânticas da expressão como um todo não podem ser obtidas a partir de suas partes constituintes [42] [13]. Este tipo de estrutura compõe, segundo Ramish *et al.* [42], pelo menos, 50% do vocabulário de um domínio especializado.

Vale ressaltar que a necessidade de identificação do vocabulário do domínio, bem como das expressões multipalavras (componentes desse) não é uma exclusividade das abordagens estimativas (estatísticas) sendo empregada também em abordagens associativas.

No restante desta seção serão apresentadas as principais técnicas utilizadas para a extração de palavras relevantes ao domínio e expressões multipalavras. Ao final, serão apresentadas as duas ferramentas utilizadas neste trabalho para extração dessas.

### 4.5.1 Identificação do Vocabulário de Domínio

A identificação do vocabulário do domínio, que também inclui o conjunto de expressões multipalavras pertencentes a esse, pode ser realizada de acordo com três principais abordagens[27] [30]:

- Abordagens estatísticas;
- Abordagens heurísticas;
- Abordagens híbridas.

Abordagens estatísticas baseiam-se na frequência de ocorrência de *ngramas* (unigramas, bigramas, trigramas, etc.), que consistem em grupos formados por *n* palavras contínuas.

Este tipo de abordagem, em geral, é composto por 3 etapas[30]:

1. Construção de *ngramas*;
2. Contabilização da frequência dos *ngramas*;
3. Filtragem dos *ngramas*.

Inicialmente, a partir do corpus definido, são construídos grupos formados por *n* palavras, também conhecidos como *ngramas* sendo que cada um desses é considerado como um candidato à expressão multipalavra.

Na segunda etapa do processo, os conjuntos de palavras (*ngramas*) são contabilizados de acordo com sua frequência de ocorrência e organizados em uma lista ordenada (em geral com o elemento mais frequente no topo).

Por fim, a lista de *ngramas* é filtrada de acordo com algum critério, de modo que apenas um sub-conjunto dos elementos considerados como mais representativos será extraído. O critério mais comumente empregado é a frequência de ocorrência dos elementos que pode ser considerada de forma estática (os 1000 elementos mais frequentes, por exemplo) ou dinâmica (10% dos elementos).

A lista de *ngramas* pode, ainda, ser refinada de acordo com heurísticas que realizam tarefas como a remoção de artigos e símbolos indesejados, por exemplo [30].

Abordagens heurísticas (ou linguísticas), por sua vez, baseiam-se em características linguísticas das palavras como, por exemplo, rótulos gramaticais (*Part-of-Speech tag*). Este tipo de abordagem costuma ser realizado de acordo com os seguintes passos:

1. Anotação morfossintática;
2. Extração de palavras candidatas;

3. Filtragem dos candidatos extraídos.

Inicialmente, os documentos são anotados (por ferramentas específicas) com informações morfossintáticas como, por exemplo, rótulos gramaticais (substantivo, verbo, etc.).

Em seguida, padrões sintáticos ("substantivo + substantivo", por exemplo) são utilizados sobre os rótulos atribuídos para a identificação e extração dos *ngramas* candidatos à entradas do vocabulário.

Por fim, os candidatos extraídos são filtrados de acordo com métricas e heurísticas para a seleção dos mais adequados a serem incluídos no vocabulário.

Assim como na etapa anterior, os candidatos podem ser pré-processados de acordo com heurísticas. No entanto, neste tipo de abordagem, as heurísticas costumam utilizar informações linguísticas no lugar de estatísticas como a frequência, por exemplo.

Por fim, abordagens híbridas combinam diferentes aspectos das abordagens anteriormente apresentadas. Uma das possíveis configurações (utilizada pelo *TTC TermSuite* [44]) é a utilização de uma abordagem linguística baseada em padrões gramaticais para a extração dos termos candidatos, seguida do cálculo de frequência destes para a realização da filtragem.

Posteriormente, as unidades léxicas extraídas são filtradas mediante um cálculo de *termhood* (baseado métricas específicas como *TF/IDF* [30] [14] [53], por exemplo) para determinar o grau de relevância desses para um determinado domínio. As unidades mais relevantes são considerados como termos do domínio.

O cálculo de *Termhood* mede o quanto um determinado termo é significativo para um domínio específico, sendo utilizado para verificar quais das palavras (e expressões multipalavras) podem ser considerados como termos do domínio em questão [14].

A identificação do vocabulário da área e de expressões multipalavras, pode ser realizada através de ferramentas extratoras de terminologia como, por exemplo o *ExATOlp*[29] e o *TTC Term Suite*[44], utilizados neste trabalho. As ferramentas mencionadas serão a seguir apresentadas.

### 4.5.2 Ferramentas

Para a identificação do vocabulário de domínio e de expressões multipalavras foram utilizadas as ferramentas *ExATOlp*[29] (português) e *TTC TermSuite*[44] (inglês).

Cogitou-se, ainda, a utilização da ferramenta *MWEtoolkit* [42], no entanto não foram encontrados os padrões morfológicos necessários para a extração das expressões multipalavras, optando-se por inserir a utilização dessa na lista de trabalhos futuros.

*ExATOlp*

O *ExATOlp* [29] é uma ferramenta para a extração da terminologia de domínio a partir de corpora em português. Dentre os recursos linguísticos disponibilizados por esta ferramenta, destacam-se:

- Lista de termos, juntamente com sua anotação morfossintática e de frequência;
- Concordanciador que disponibiliza uma lista de todas as sentenças nas quais um determinado termo está contido;
- Nuvem de conceitos (*Concept Cloud*), recurso visual que permite a visualização dos termos relevantes em um ambiente no qual o tamanho da fonte é proporcional à relevância desses;
- Hierarquia de conceitos, apresenta os conceitos extraídos hierarquicamente dispostos em uma árvore hiperbólica (*hyperbolic tree*).

Dos recursos anteriormente apresentados, aquele com maior relevância para este trabalho é a extração das lista de termos, utilizado na ampliação dos dicionários monolíngues da língua portuguesa.

A ferramenta utiliza uma abordagem linguística, na qual, um corpus anotado por um analisador sintático (*parser*) é investigado de acordo com padrões linguísticos e, para os candidatos extraídos são aplicadas heurísticas de refinamento e, posteriormente, filtros estatísticos baseados em frequência.

Por fim, a ferramenta realiza uma cálculo de *termhood* que leva em consideração a frequência do termo tanto no documento a partir do qual este foi extraído quanto em documentos de diferentes domínios para determinar, assim sua especificidade. Uma descrição mais detalhada do processo utilizado pela ferramenta pode ser obtida em [29].

*TTC TermSuite*

O *TTC TermSuite* [44] [18] é uma ferramenta originalmente criada para a extração de terminologia bilíngue a partir de corpus comparável [44] [18]. Esta ferramenta encontra-se inclusa no escopo do projeto *TTC Project* que tem por objetivo a extração de terminologia de domínios específicos, a partir de documentos comparáveis extraídos da *Web*.

Como parte de processo, o *TTC TermSuite* possui funcionalidades para a extração da terminologia monolíngue, utilizando uma abordagem híbrida na qual padrões sintáticos são empregados para a extração de candidatos que posteriormente são filtrados de acordo com critérios como, por exemplo, frequência, especificidade, entre outros.

O processo de extração utilizado é composto por cinco passos [44]:

1. Reconhecimento de candidatos a termos simples e compostos;

2. Cálculo de suas frequências relativas e especificidade no domínio;

3. Detecção de *neoclassical words* (palavras técnicas) a partir do conjunto de palavras simples;

4. Agrupamento das variantes do termo (flexões);

5. Filtro de candidatos utilizando um limiar (*treshold*) que pode ser especificado tanto em relação à frequência quanto à especificidade em um domínio.

Como o desenvolvimento da ferramenta em questão ainda encontra-se em andamento, sendo que o processo de validação ainda não foi realizado, dados em relação ao seu desempenho não puderam ser encontrados. No entanto, este fato não impede sua utilização uma vez que a funcionalidade de extração monolíngue encontra-se implementada e é baseada em métodos conhecidos.

## 4.6 Avaliação de Léxicos Bilíngues

Existem duas formas principais de avaliação para recursos linguísticos como léxicos bilíngues: intrínseca e extrínseca. Essas podem, ainda, ser executadas de forma manual ou automática [11].

Em uma avaliação intrínseca, o léxico é avaliado de acordo com o conteúdo que apresenta (palavras e expressões multipalavras) mediante a utilização de alguma métrica que pode variar de acordo com o objetivo da avaliação.

Quando executada de forma manual, juízes humanos realizam a avaliação das entradas classificando-as de acordo com categorias predefinidas. Esta avaliação pode ser binária (correta ou incorreta) ou ainda levar em consideração outras categorias como, por exemplo, parcialmente correta.

Já quando realizada de forma automática, o léxico induzido é comparado com um léxico de referência, sendo que as entradas do léxico induzido também presentes no léxico de referência são consideradas como corretas.

Em uma avaliação extrínseca o léxico bilíngue é utilizado em alguma aplicação de PLN (extração de informações multilíngue, por exemplo). Nesta abordagem, a avaliação é realizada sobre os resultados da aplicação e, a partir desses, determina-se o quanto o léxico induzido contribuiu para a aplicação [11].

Assim como na avaliação intrínseca manual, a avaliação extrínseca é conduzida por juízes humanos que, no entanto, ao invés desses avaliarem o conteúdo do léxico, avaliam as saídas produzidas pela tarefa de PLN na qual esse foi utilizado.

Por fim, em uma avaliação automática extrínseca, assim como na intrínseca, o que é a avaliado são as saídas da atividade de PLN, no entanto, diferentemente da avaliação manual, esta é conduzida de forma automática empregando padrões de referência da saída avaliada.

A escolha entre uma abordagem intrínseca (utilizada neste trabalho) ou extrínseca depende diretamente do que se deseja avaliar. Uma avaliação intrínseca avalia o léxico como um produto final, enquanto a avaliação extrínseca avalia a colaboração desse na realização de uma determinada tarefa.

Neste trabalho, como o objetivo da avaliação foi avaliar o desempenho do processo de extração proposto (diretamente relacionada ao vocabulário extraído), optou-se pela utilização de uma abordagem intrínseca, na qual foi avaliada a identificação das equivalências multilíngues. A metodologia utilizada encontra-se descrita na Seção 7.1.

Em relação à escolha entre abordagens manuais ou automáticas, esta depende diretamente dos recursos disponíveis. Avaliações automáticas dependem de padrões de referência (léxico, dicionários, listas, etc.) [11], enquanto avaliações manuais dependem da disponibilidade de juízes humanos.

A principal vantagem de avaliações automáticas sobre manuais é a possibilidades de avaliar o impacto de modificações no processo de forma rápida e imediata. No entanto, léxicos de referências para textos de domínios específicos são recursos escassos e sua construção pode ser custosa.

Já a avaliação manual depende da disponibilidade de juízes humanos e na existência de um nível de concordância aceitável entre esses (podendo esta ser mensurada pelo cálculo de *Kohens Kappa*). No entanto, a abordagem manual proporciona uma avaliação mais detalhada, podendo auxiliar na identificação de questões até então não investigadas, como em relação a problemas na identificação das expressões multipalavras ocorridos neste trabalho, descritos na Seção 7.2.

Como para o presente trabalho a alocação de juízes humanos era menos custosa do que a construção de léxicos de referência, optou-se pela realização de uma avaliação manual como apresentado no Capítulo 7.

# 5. TRABALHOS RELACIONADOS

A extração automática de vocabulário multilíngue é um tema bastante explorado na literatura relacionada à Tradução de Máquina (MT). Neste capítulo serão apresentados os principiais trabalhos que embasaram a construção do processo proposto (Capítulo 6).

Este capítulo encontra-se organizado de modo que inicialmente (Seção 5.1) são apresentados os trabalhos de Caseli [11] e Caseli e Nunes [11], na Seção 5.2 são apresentados o processo de extração de vocabulário multilíngue [47] e o corpus [49] propostos por Tiedemann e, na Seção 5.3 são apresentados os trabalhos de Ha *et al.* [14] e Zhang [53]. Por fim, a Seção 5.4 discute a relação destes trabalhos com o presente (Seção 5.4).

## 5.1 *ReTraTos*

O *ReTraTos* [11] é uma ferramenta que tem por objetivo a indução automática de léxicos bilíngues e regras de tradução, ambos representados em um formato compatível com a plataforma de tradução de máquina *Apertium* [15].

Das publicações relacionadas ao *ReTraTos*, destaca-se a de Caseli e Nunes [12] que trata especificamente da indução de léxicos bilíngues a partir de textos paralelos, objetivo esse diretamente relacionado ao presente trabalho. Maiores informações sobre a indução das regras de tradução e detalhes da ferramenta podem ser obtidos em [11].

O processo proposto por Caseli e Nunes [12] tem início em uma etapa de pré-processamento dos documentos componentes do corpus. Nessa, os documentos foram decompostos em suas sentenças formadoras (identificadas por rótulos) e essas foram organizadas no formato de uma sentença por linha (arquivo de saída).

Na segunda etapa, os conjuntos de sentenças (um português e outro inglês) foram sentencialmente alinhados (de forma automática) mediante a utilização da ferramenta *TCAalign* [10], em conjunto com uma lista auxiliar de palavras cognatas. Os resultados desta etapa foram manualmente verificados e os alinhamentos corrigidos foram utilizados na construção de uma lista de referências.

Na terceira etapa foi conduzida a anotação morfológica das palavras componentes das sentenças. Para cada palavra foram atribuídos rótulos gramaticais (*POStag*), formas canônicas e flexões de número e gênero. O processo de anotação foi realizado através de ferramentas da plataforma *Apertium* (*Lttoolbox*) [15].

Vale ressaltar que a ferramenta de análise morfológica utilizada é baseada em dicionários externos. Os dicionário monolíngues utilizados para anotação dos documentos foram obtidos a partir do repositório da ferramenta *Apertium* [1] e, posteriormente ampliados com entradas de outros dicionários.

[1] http://sourceforge.net/projects/apertium/files/lttoolbox/

Os dicionários foram ampliados a partir de entradas extraídas dos dicionários produzidos pelo projeto Unitex[2], passando a abranger 337.861 formas superficiais (forma pela qual uma palavra se apresenta no texto) para a língua portuguesa e 61,601 para a língua inglesa.

Na quarta etapa do processo foi realizado o alinhamento lexical (a nível de palavra) dos documentos. Para esta tarefa, foi utilizada a ferramenta *LIHLA* no alinhamento dos documentos português-espanhol (*pt_es*), e o *Giza++* (versão 2.0) no alinhamento português-inglês.

A ferramenta *Giza++* (versão 2.0) foi treinada com todos os 17.397 exemplos de tradução (sentenças) gerados durante a segunda etapa, e utilizada de acordo com sua configuração padrão (iterações dos modelos IBM-1, IBM-3, IBM-4 e HMM). Os alinhamentos lexicais obtidos foram, ainda, submetidos ao algoritmo de simetrização (união) proposto por Och e Ney [39], em busca de maior precisão. Nesta etapa, Caseli e Nunes [12] ressaltam ter obtido uma precisão de 90,47% e uma cobertura (*recall*) de 92,34% (obtido após avaliação manual de 500 sentenças).

Por fim, as saídas dos alinhadores lexicais foram utilizadas para a indução dos léxicos bilíngues que, por sua vez, foram representados de acordo com os formalismos da ferramenta *Apertium* [15]. A Figura 5.1 demonstra um exemplo de saída do módulo de pós processamento. O processo de indução é descrito na Seção 5.1.1.

```
<e r="LR">
»       <p>
»       »       <l>phone<s n="adv"/></l>
»       »       <r>telefone<s n="n"/><s n="m"/><s n="sg"/></r>
»       </p>
</e>
```

Figura 5.1 – Exemplo de saída da ferramenta *ReTraTos*.

Vale ressaltar que a indução dos léxicos bilíngues difere da construção de dicionários bilíngues probabilísticos. Enquanto dicionários probabilísticos retornam a probabilidade duas palavras serem equivalentes, léxicos induzidos retornam padrões de tradução nos quais uma palavra, anotada com determinadas informações morfossintáticas, é considerada como equivalente de outra palavra (em outro idioma) quando essa encontrar-se anotada de uma forma específica (determinado conjunto de informações).

Para a avaliação do processo apresentado, Caseli e Nunes [12] conduziram um experimento. Nesse, dois corpus paralelos foram utilizados, um para os idiomas português (do Brasil) e Inglês, e o outro para os idiomas português (do Brasil) e espanhol.

O corpus português/espanhol utilizado era composto por 18.236 sentenças sendo formado por 594.391 palavras para o português e 645.866 palavras para o espanhol, enquanto o segundo corpus era composto por 17.397 sentenças paralelas contabilizando 494.391 palavras para o português e 532.121 palavras para o inglês. Os documentos utilizados foram levantados a partir da revista *Pesquisa FAPESP*.

[2] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html

A avaliação do léxico bilíngue *português-inglês* gerado ao final do experimento utilizou uma abordagem manual intrínseca, devido à padrões de referencia (*Golden Standards* não se encontrarem disponíveis [11].

As 19.191 entradas (15.949 palavras simples e 3.242 multipalavras) do léxico foram classificadas em 8 classes de acordo com seus atributos. Dessas foram avaliadas somente aquelas cujos atributos fossem iguais para as entradas fonte e alvo ou que apresentassem pequenas diferenças em relação a estes (atributos mais específicos ou gerais). Por fim, aproximadamente 10% das entradas resultantes da seleção foram avaliadas.

A escala de avaliação utilizada classificou as entradas avaliadas em três categorias:

- Válidas (V): A parte fonte é uma possível tradução da parte alvo, considerando-se o sentido de tradução especificado;
- Parcialmente Válidas (PV): A especificação seria válida se alguma alteração nas informações morfológica ou no sentido da tradução fossem realizadas;
- Não Válidas (NV): A correspondência entre as duas partes não é válida.

A avaliação foi conduzida por dois juizes humanos. Cada juiz avaliou manualmente 618 entradas (519 palavras simples e 99 multipalavras) sendo que parte dos conjuntos de avaliação era comum aos dois avaliadores. Uma verificação do nível de concordância entre os juízes (medida pelo cálculo de *kappa*) apresentou um valor de 0,48, indicando baixa concordância entre os avaliadores. No entanto, devido aos tipos de discordância (principalmente relacionado à expressões multipalavras) encontrados, os juízes foram considerados aptos para o procedimento.

### 5.1.1 Processo de Indução

O processo de indução dos léxicos bilíngues utilizado pela ferramenta *ReTraTos* é composto por 7 passos [11]:

- Criação de um léxico bilíngue para o sentido fonte-alvo;
- Criação de um léxico bilíngue para o sentido alvo-fonte;
- União dos léxicos criados nos passos anteriores;
- Generalização das entradas do léxico bilíngue;
- Tratamento de diferenças de gênero e número (opcional);
- Tratamento de multipalavras;

Antes da execução dos passos anteriormente apresentados, é executado um passo de pré-processamento, no qual os textos paralelos são lidos e armazenados em estruturas de dados específicas do processo.

No início do processo (passos 1 e 2), são buscadas todas as possíveis traduções para uma determinada entrada do texto fonte (palavra ou expressão multipalavra), considerando as diferentes combinações de informações morfológicas para ambos os idiomas. Após construída a lista de possíveis alinhamentos, a frequência de ocorrências desses é verificada, sendo que o par de maior frequência é considerado como a tradução mais adequada. Este procedimento é adotado durante o primeiro e segundo passos, alternando apenas os textos fonte e alvo.

No terceiro passo, os alinhamentos produzidos nos dois primeiros passos (fonte-alvo e alvo-fonte) são unidos (mediante algoritmo de união de Och e Ney [39]) e as ambiguidades são resolvidas. Neste passo é realizada a filtragem de expressões multipalavras mediante sua frequência de ocorrência. A frequência mínima para a inclusão de multipalavras pode ser manualmente definida pelo usuário da ferramenta.

No quarto passo tenta-se generalizar entradas similares do léxico bilíngue, as quais possuam apenas um atributo (informações morfológicas) com valores diferentes. Durante a generalização, estas entradas são unificadas e os possíveis valores para o atributo diferente são inseridos no mesmo campo, separados pelo símbolo "|". As frequências de ocorrência das entradas agrupadas são somadas.

O passo 5 (opcional) é responsável pelo tratamento de entradas nas quais o valor do atributo de gênero ou número não pode ser determinado de acordo com as informações contidas na entrada (palavras que possuem a mesma forma em diferentes contextos como *thesis*, por exemplo). Para este tratamento, é necessário que o usuário forneça uma lista contendo possíveis valores dos atributos de número e gênero de acordo com um formato de representação específico. Ainda durante este passo as entradas válidas para ambos os sentidos são divididas em duas novas entradas com uma indicação para o sentido de tradução.

Por fim, no passo 6 é realizada a impressão das entradas do léxico bilíngue no arquivo de saída. Durante esta impressão, as entradas são formatadas de acordo com o formalismo dos dicionários de entrada da ferramenta *Apertium* [15].

## 5.2 *Uplug* e *OPUS*

Os trabalhos de Tiedemann [47] [49] diferem entre si no processo utilizado, no entanto, ambos possuem elementos relacionados a este trabalho. O projeto *OPUS* [49] tem por objetivo suprir a comunidade científica com corpora paralelo para múltiplos idiomas e livremente acessível. Além de corpora, o projeto também agrega um conjunto de ferramentas para processamento, anotação e gerenciamento deste tipo de recurso.

Dentre os domínios dos corpora disponibilizados pelo *OPUS*, destacam-se o de legendas de filmes (construído a partir do repositório *open subtitles*), biomedicina (*EMEA*) e documentação de *softwares* (*open source*).

Antes de serem inseridos ao repositório os textos passam por um processo no qual são pré-processados (de acordo com seu tipo), convertidos para o formato XML(eXtensible Markup Language), seguindo uma estrutura própria do *OPUS*, e alinhados sentencialmente com seus textos equivalentes mediante a utilização da ferramenta *Hunalign* [50].

A etapa de pré-processamento varia de acordo com o tipo de texto a ser processado uma vez que as características de um documento contendo legendas de filmes é diferente das encontradas em documentos de *software*.

Entre as técnicas de pré-processamento aplicadas por Tiedemann [49], a utilizada em documentos do corpus de biomedicina (*EMEA*) foi considerada a mais similar à empregada sobre os documentos do corpus apresentado no Capítulo 3 uma vez que os documentos de ambos os corpora possuem o mesmo formato de entrada (*PDF*).

Em [49] Tiedemann conduz o alinhamento utilizando o *Hunalign*, já no *Uplug* (extrator automático de dicionários multilíngues) [47] permite que seja utilizado o *Hunalign* [50], o GMA[34] ou um alinhador padrão baseado no tamanho das sentenças.

Para o alinhamento lexical,Tiedemann [49] utiliza a ferramenta *Giza++* (versão 2.0) e treina a ferramenta sobre o corpus inteiro, realizando o alinhamento em ambas as direções (fonte/alvo e alvo/fonte) para, posteriormente, aplicar o algoritmo de união (simetrização) proposto por Och e Ney[39] sobre os alinhamentos resultantes.

Os textos (no formato *XML*) são então convertidos para o modelo de entrada de uma ferramenta de gerenciamento de corpus (*Corpus Workbench*) que disponibiliza um conjunto de funcionalidades para auxiliar na visualização de corpora como, por exemplo, concordanciadores multilíngues.

A maior parte das ferramentas de pré-processamento e conversão utilizadas por Tiedeman são provenientes do *Uplug* [47] que apresenta um conjunto de ferramentas para a extração de dicionários bilíngues probabilísticos.

A principal diferença encontrada no *Uplug* [47] é a forma pela qual este executa o alinhamento lexical, utilizando uma abordagem baseada em pistas ("*cues*") que se baseia em uma série de características da palavra (pistas) para a execução do alinhamento. Uma descrição detalhada do projeto pode ser encontrada em [47].

## 5.3 Ha *et al.* [14] e Zhang [53]

Os trabalhos de Ha *et al.* [14] e Zhang [53] são bastante similares, utilizando abordagens associativas para alinhamento lexical e tendo como objetivo principal auxiliar processos de extração terminológica.

Como corpus paralelo, Ha *et al.* [14] utilizaram um subconjunto do *MedlinePlus* construído a partir da *Medline*, considerada a maior biblioteca médica disponível. O subconjunto utilizado era composto por 9.250 sentenças paralelas sendo essas constituídas por 31.498 palavras para o inglês e 30.344 palavras para o espanhol.

Zhang[53], por sua vez, utilizou um corpus composto por textos acadêmicos sendo este composto por 460.000 documentos para o idioma chinês e 130.000 documentos para o inglês, distribuídos entre 23 categorias com uma média de 4.733 registros paralelos por categoria.

Em relação ao processo de extração utilizado, Ha *et al.* [14] inicialmente executaram uma etapa de pré-processamento na qual os documentos originais foram convertidos para texto sem formatação (*plain text*) tendo sido removidas as quebras de linha para evitar problemas posteriores com alinhadores sentenciais. Zhang [53] não apresenta o procedimento de pré-processamento empregado.

Na segunda etapa, ambos os trabalhos realizaram a identificação da terminologia de domínio. Esta identificação foi realizada de forma individual para cada um dos idiomas envolvidos. Nesta etapa, Ha *et al.* [14] utiliza uma abordagem baseada em padrões morfossintáticos para a extração dos termos, enquanto Zhang [53] utiliza uma abordagem baseada em palavras-chave.

Posteriormente, em ambos os trabalhos, foi calculado o *termhood*(especificidade de um termo em relação a um domínio) dos candidatos a termo previamente extraídos.Para este cálculo, foram utilizados critérios como frequência e métricas como *TF/IDF* (*Term Frequency/Inverse Document Frequency*), por exemplo.

A terceira etapa dos processos consistiu no alinhamento sentencial do corpus, procedimento para o qual Ha *et al.* [14] utilizaram a ferramenta *Trados WinAlign*, enquanto Zhang [53] não apresentaram a ferramenta utilizada.

Na quarta etapa, tabelas de contingência foram construídas para os termos candidatos (extraídos nos passos anteriores), como apresentado na Seção 4.4. A partir das tabelas de contingência, métricas de associatividade foram utilizadas para a determinação equivalências multilíngues (pares de palavras).

Zhang [53] fez uso de um conjunto de métricas composto por *LogLikeliHood*[33], *Dice,* $X^2$ *e MI (Mutual Information*.Ha *et al.* [14], utilizaram apenas *LogLikeliHood*[33]. Vale ressaltar que a métrica de *LogLikeliHood*[33] é utilizada pois, segundo os autores previamente referenciados, é capaz de identificar associações de baixa frequência.

Em relação à metodologia de avaliação utilizada por Han *et al.* [14] e Zhang [53], esta tem por objetivo avaliar a extração terminológica e não a extração de equivalências multilíngues. Sendo assim, a metodologia em questão não foi considerada por este trabalho.

## 5.4 Relação com o Presente Trabalho

O processo proposto neste trabalho (Capítulo 6), segue um conjunto de passos bastante similar aos dos trabalhos apresentados no decorrer deste capítulo, destacando-se a semelhança com o trabalho de Caseli e Nunes [12].

Das etapas componentes do processo, o pré-processamento e o alinhamento sentencial seguiram abordagens muito similares às empregadas nos demais trabalhos, tendo como principais alterações as ferramentas utilizadas.

O procedimento para a análise morfológica foi baseado nos trabalhos de Caseli e Nunes [12], Ha *et al.*[14] e Zhang[53]. Desses, a expansão dos dicionários e a escolha das ferramentas utilizadas basearam no trabalho de Caseli e Nunes, enquanto a utilização de termos mais específicos aos documentos do corpus foi baseada nos trabalhos de Ha *et al.* e Zhang.

Na etapa de alinhamento lexical, assim como em Caseli e Nunes [12] e Tiedemann [49], optou-se pela utilização de abordagens estimativas, conduzidas através da ferramenta *Giza++* (versão 2.0) [39], cujo treinamento utilizou o corpus inteiro, seguindo as configurações padrão da ferramenta.

Por fim, para a indução dos léxicos multilíngues e geração dos dicionários no formato de entrada da ferramenta *Apertium* foi utilizada a ferramenta *ReTraTos* [11], proposta e utilizada no trabalho de Caseli e Nunes.

Em relação à similaridade do presente trabalho com o de Caseli e Nunes [12], a principal diferença que pode ser apontada é forma pela qual a ampliação dos dicionários morfológicos é conduzida, sendo que o presente trabalho utiliza extratores automáticos de terminologia ao invés de listas previamente construídas.

Outra diferenças entre os dois trabalhos é relacionada ao domínio das palavras utilizadas para a ampliação dos dicionários. Enquanto Caseli e Nunes [12] utilizam dicionários de domínio geral, este trabalho utiliza palavras (e expressões multipalavras) mais específicas do domínio, extraídas por ferramentas de extração terminológicas.

# 6. PROCESSO PROPOSTO

Este capítulo apresenta o processo proposto para a extração de vocabulário multilíngue, cujo o objetivo é a extração de vocabulário multilíngue a partir do corpus paralelo formado por documentos de *software* (Seção 3.1). O processo segue as etapas apresentadas no Capítulo 4, sendo adicionadas uma etapa de pré-processamento (Seção 6.2) e outra de indução dos léxicos bilíngues (Seção 6.6).

Este capítulo encontra-se organizado de modo que inicialmente é apresentada uma visão geral do processo (Seção6.1), seguida pela descrição das etapas de pré-processamento (Seção 6.2), alinhamento sentencial (Seção 6.3), análise morfológica (Seção 6.4), alinhamento léxico (Seção 6.5) e, por fim, a indução do léxico (Seção 6.6).

## 6.1 Visão Geral do Processo

Nesta seção é apresentada uma visão geral do processo proposto, sendo demonstrada a interação entre as etapas apresentadas nas seções posteriores. A Figura 6.1 apresenta o processo como um todo.

A entrada do processo consiste em um corpus paralelo formado por dois conjuntos de textos equivalentes (paralelos) nos idiomas português ("*Corpus PT*") e inglês ("*Corpus EN*"). O corpus utilizado neste trabalho encontra-se descrito no Capitulo 3.

Os textos componentes do corpus são submetidos a uma etapa de pré-processamento na qual são convertidos para texto puro ("*txt*") e reorganizados no formato de uma sentença por linha (como descrito na Seção 6.2).

A segunda etapa do processo consiste no alinhamento sentencial do conjunto de documentos pré-processados. Nessa, são estabelecidas as equivalências entre as sentenças dos documentos em português com seus correspondentes em inglês. O alinhamento é realizado com os textos de cada conjunto ainda separados. O processo de alinhamento encontra-se descrito na Seção 6.3.

A próxima etapa consiste na anotação morfológica, na qual os documentos de cada um dos conjuntos produzidos são analisados e anotados com informações morfossintáticas como, por exemplo, rótulos gramaticais (*Part-of-Speech Tags*) e flexões de número e gênero. Ainda durante esta etapa é realizada a identificação das expressões multipalavras (*Multi-Word Expressions*). Esta etapa encontra-se descrita na Seção 6.4.

Após morfologicamente anotados, os arquivos fonte e alvo são utilizados como entrada para a etapa de alinhamento léxico, na qual são determinadas as equivalências entre os símbolos (palavras, expressões multipalavras etc.) componentes desses. O processo de alinhamento léxico é apresentado na Seção 6.5.

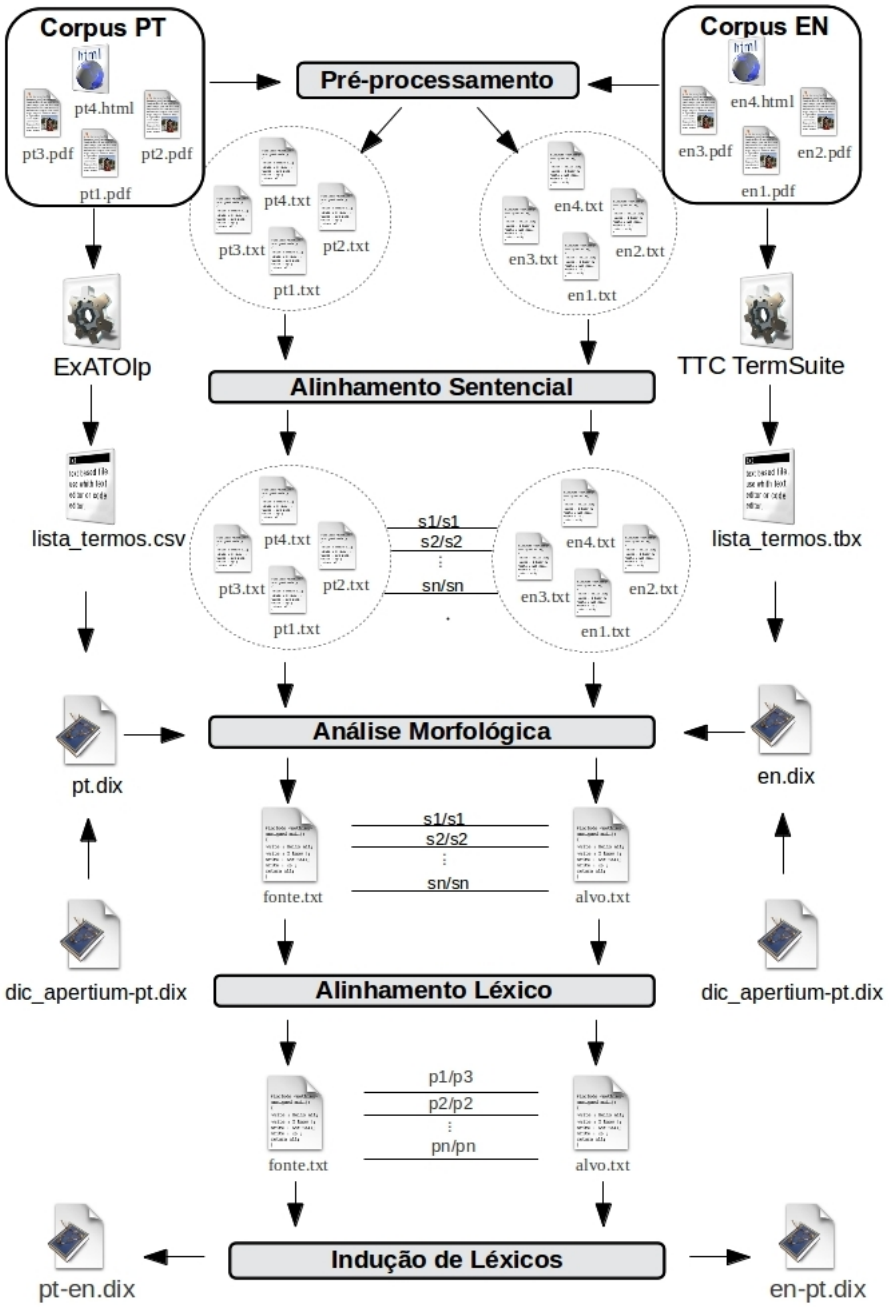


Figura 6.1 – Visão geral do processo proposto.

Como saída da etapa de alinhamento léxico são gerados dois arquivos, cada um contendo uma direção de alinhamento (português/inglês e inglês/português). Esses são então simetrizados através do algoritmo de união proposto por Och e Ney [39] e utilizados como entrada para a última etapa do processo, referente a indução de léxicos conduzida pela ferramenta "ReTraTos" [11].

A etapa de indução de léxicos consiste na extração de entradas bilíngues a partir dos arquivos de alinhamento, sendo estas compostas por palavras simples e expressões multipalavras. O processo de indução é descrito na Seção 6.6.

Por fim, a etapa de indução dos léxicos gera como saída dois dicionários bilíngues, uma na direção português/inglês e outro na direção inglês/português. Estes dicionários são gerados de acordo com modelo utilizado pelo *Apertium* [15].

Como pode ser observado na Figura 6.1 os conjuntos de documentos formadores do corpus são submetidos à ferramentas extratoras de terminologia (*ExATOlp* [29] e *TTC TermSuite* [44]) com o objetivo de identificar o vocabulário relevante do domínio e utilizá-lo para a ampliação dos dicionários morfológicos empregados (como apresentado na Seção 6.4.1). Uma descrição mais completa do processo proposto é apresentada nas próximas seções.

## 6.2 Pre-processamento

Nesta seção serão apresentados os procedimentos de pré-processamento empregados na preparação do corpus construído (Seção 3.1) para sua utilização em etapas posteriores do processo de extração do vocabulário bilíngue.

Esta etapa é constituída por três passos, a seguir apresentados:

1. Conversão dos documentos para texto sem formatação (*plain text*);
2. Remoção dos símbolos de formatação (eliminação do ruído);
3. Separação e organização das sentenças;

As próximas seções apresentam descrições detalhadas dos passos apresentados.

### 6.2.1 Conversão de Documentos

Os manuais de usuário, componentes do corpus, foram coletados em três formatos distintos: *PDF* (*Portable Document Format*) , *HTML* (*Hypertext Markup Language*) e "*txt*" (texto sem formatação).

Nesta etapa, os documentos do corpus foram convertidos para arquivos de texto no formato "*txt*" codificados em "*UTF-8*", de acordo com os procedimentos a seguir apresentados.

#### Conversão de Arquivos PDF

A conversão de documentos no formato PDF foi realizada através da ferramenta *pdftotext* [1] (assim como em [48]) configurada de acordo com os seguintes parâmetros:

- -nopgbrk: Remove marcadores de quebras de página;

[1] http://linux.die.net/man/1/pdftotext

- -enc UTF-8: Define a codificação do arquivo de saída para "UTF-8";
- -eol unix: Define o símbolo para quebra de linha de acordo com o padrão Unix (também utilizado por distribuições Linux);
- -raw: O arquivo original é lido no formato de uma sequência de palavras (*stream*).

Uma descrição mais detalhada da utilização da ferramenta, bem como dos parâmetros utilizados e seus possíveis valores pode ser obtida no manual da ferramenta[2].

O parâmetros "*-raw*" merece destaque pois delimita a forma pela qual o arquivo original é lido, sendo que, para o valor adotado, o texto é lido no formato sequencial desfazendo assim sua formatação original.

Tiedemann [48] defende a utilização do parâmetro "*-layout*" (que tenta manter a formatação original do arquivo), relatando ter obtido melhores resultados durante a conversão de documentos. No entanto, após alguns testes realizados sobre os documentos, optou-se por manter o valor "*-raw*", devido a este apresentar melhores resultados de acordo com o tipo de formatação apresentado pelos documentos do corpus.

Para ilustrar a diferença entre os dois parâmetros utiliza-se como exemplo um texto formatado em duas colunas. O parâmetro *"-raw"* converterá o texto para apenas uma coluna, enquanto a opção *"-layout"* manterá o formato de duas colunas.

O principal problema associado à opção *"-layout"*, no contexto deste trabalho é que, ao manter a formatação original de duas colunas, as sentenças do texto são fragmentadas em várias linhas, dificultando o processamento.

#### Conversão de Documentos HTML

Inicialmente, as páginas HTML componentes dos manuais foram capturadas com o auxílio da ferramenta *wget*[3] que disponibiliza um conjunto de funcionalidades para esta finalidade. Documentos HTML, diferentemente dos arquivos PDF, não precisam ter seu formato convertido. No entanto, precisam ter a marcação HTML(rótulos) removida, uma vez que esta não pertence ao texto original. A remoção dos marcadores HTML foi realizada através de funcionalidades (método "*clean_html()*") da biblioteca NLTK [4].

### 6.2.2 Tratamento do Corpus

A conversão de arquivos PDF para texto puro ("*.txt*") frequentemente acrescenta ruído ao texto resultante. Este ruído, em geral, se demonstra na forma de símbolos desconhecidos às ferramentas, utilizando codificações ou fontes não suportadas por essas.

[2] http://linux.die.net/man/1/pdftotext
[3] http://www.gnu.org/software/wget/

Além do ruído inserido durante a conversão dos arquivos PDF, existem outros símbolos, frequentemente pertencentes à formatação dos documentos como, por exemplo, sequências de pontos utilizados para reconstruir a estrutura de sumário nos documentos.

Remover este tipo de símbolos torna-se necessário pois esses tendem a causar problemas durante a leitura e processamento dos arquivo convertidos nas etapas posteriores do processo de extração (alinhamento léxico, por exemplo). No entanto, vale ressaltar que o tipo de ruído (símbolos) varia, de acordo com o tipo de documento convertido, não sendo possível assim, a automatização completa do processo de remoção.

Antes que o processo de remoção fosse realizado, duas questões foram levantadas: (a) Quais os símbolos a serem definidos como ruído; (b) O quanto é aceitável modificar corpus original mediante sua remoção.

A questão da definição de símbolos como ruído foi resolvida com a criação manual de uma lista contendo símbolos já identificados como tal (mediante observação de conversões anteriores). Exemplos de símbolos removidos podem ser visualizados na Tabela 6.1.

Em relação ao conteúdo a ser removido, foi definido que apenas símbolos relacionados à estrutura do documento (estilo de formatação visual) seriam removidos evitando assim, a remoção de informações relevantes ou a introdução de algum tipo de viés ao corpus [45].

Outra questão, posteriormente levantada, foi o momento no qual os símbolos deveriam ser removidos. Como mencionado no Capítulo 4, ferramentas de alinhamento sentencial frequentemente utilizam símbolos (pontuações e números, por exemplo) como âncoras (*anchors*), ou seja, símbolos comuns a ambos os idiomas que auxiliam nas identificação de sentenças correspondentes/equivalentes.

Testes utilizando diferentes abordagens de remoção conduziram a uma estratégia, na qual, parte dos símbolos são removidos antes do alinhamento sentencial e o restante após esse ser executado. A Tabela 6.1 apresenta os símbolos removidos durante a primeira etapa da remoção.

Tabela 6.1 – Símbolos removidos durante a primeira etapa.

| **Símbolos** | **Legenda** |
|---|---|
| ...[...], . . .[...] | Sequências de três ou mais pontos, frequentemente utilizadas para estrutura de sumário |
| ==[...] | Sequências de sinais de igualdade, frequentemente utilizados como separadores. |
| –[...],__[...] | Sequências de hifens e traços, utilizados principalmente na estruturação do documento. |

Um dos tratamentos considerado de suma importância é a remoção dos símbolos utilizados para a quebra de linha ("\n", em ambientes Unix/Linux) [48], especialmente em documentos convertidos a partir do formato PDF. Esta remoção corrige grande parte das quebras de linha incorretamente inseridas pelo *software* de conversão que, frequentemente, quebram sentenças antes de sua finalização, prejudicando a identificação e separação posterior dessas.

Outras informações também removidas durante a primeira etapa da limpeza dos documentos foram os marcadores (*bulletmarks*) e os números das páginas, sendo que esses números não colaboram para o processo de alinhamento.

Durante a segunda etapa do alinhamento foram removidos os seguintes símbolos: #,%, =, <, >, +, *, utilizados, principalmente, na formatação do *layout* dos arquivos. Esses foram eliminados após a observação de sua interferência negativa tanto no alinhamento léxico quanto na indução dos léxicos.

Vale ressaltar que a remoção dos símbolos deve ser realizada antes do alinhamento léxico, caso contrário corre-se o risco de interferir no alinhamento realizado. A remoção dos símbolos foi realizada através de expressões regulares implementadas na linguagem de programação *Python*.

### 6.2.3 Separação de Sentenças

Como última etapa do pré-processamento, os documentos foram decompostos em suas sentenças constituintes e os arquivos foram reorganizados no formato de uma sentença por linha, como exigido pelos alinhadores sentenciais considerados [35] [10].

A separação das sentenças (*sentence detection* ou *sentence splitting*), foi realizada através de funcionalidades da biblioteca *NLTK* [4] (*sent_tokenize()*), utilizando modelos treinados para os idiomas envolvidos, disponibilizados pela própria biblioteca.

A saída deste passo (que corresponde à saída da etapa) consistiu nos documentos do corpus convertidos em arquivos de texto sem formatação organizados no formato de uma sentença por linha. Vale ressaltar que, diferentemente do trabalho de Caseli e Nunes [11], nenhum tipo de notação foi utilizada para demarcar os limites (início e fim) das sentenças.

## 6.3 Alinhamento Sentencial

A partir do conjunto de ferramentas apresentado na Seção 4.2, o *Bilingual Sentence Aligner* [35] foi selecionado para a realização do alinhamento sentencial. Testes iniciais realizados com as ferramentas levantadas demonstram que, apesar da ferramenta selecionada ter obtido um desempenho muito similar às demais, essa se demonstrou mais adequada ao contexto (sem correções manuais). O fato da ferramenta realizar apenas alinhamentos *1:1* não demonstrou ter causado impactos negativos significativos no resultado do alinhamento.

O Hunalign [50] demonstrou uma vantagem significativa em relação ao tempo de alinhamento, no entanto, seu desempenho foi inferior às demais ferramentas. Estima-se que o baixo desempenho tenha sido causado pela falta de dicionário para os idiomas envolvidos.

Como entrada, o *Bilingual Sentence Aligner* exige dois arquivos paralelos organizados no formato de uma sentença por linha, sendo que os arquivos gerados durante a etapa de pré-processamento foram utilizados.

O valor de *"threshold"* (probabilidade mínima de alinhamento para a inclusão) padrão da ferramenta é de *0.75* no entanto, para este trabalho, foi adotado um valor de 0.85. Testes preliminares demonstraram uma diferença muito pequena no número de alinhamentos produzidos devido ao aumento deste valor.

Como saída, esta ferramenta produz uma lista de arquivos sendo os mais relevantes aqueles identificados pelo sufixo "*aligned*" que correspondem às sentenças resultantes do alinhamento. Sentenças para as quais não foram encontrados alinhamentos (vazias) foram automaticamente descartadas pela ferramenta.

Os arquivos alinhados encontram-se estruturados de modo que cada linha do arquivo "<fonte> .aligned" corresponde diretamente à linha de mesmo número do arquivo "<alvo>.aligned". Sendo assim, a linha 3 do arquivo <fonte>, por exemplo, encontra-se alinhada à linha 3 do arquivo <alvo>.

Como recomendado em trabalhos relacionados [35] os arquivos foram alinhados de forma individual, ou seja, cada manual foi alinhado com seu correspondente paralelo sem que seu conteúdo fosse concatenado ao dos demais manuais de mesmo idioma.

## 6.4 Análise Morfológica

Dentre as ferramentas listadas na Seção 4.3, optou-se pela utilização do *Apertium* [15] para a anotação morfológica das sentenças alinhadas. Além de prover os níveis de anotação desejados, este conjunto de ferramentas suporta diferentes idiomas e permite que seus dicionários morfológicos sejam ampliados ou mesmo especializados para um domínio específico.

Sendo baseadas em dicionários externos, as ferramentas componentes do *Apertium* são capazes de anotar textos em diferentes idiomas, tendo-se, assim, um padrão de anotação para os idiomas envolvidos (sem necessidade de adaptação). Neste aspecto, outra vantagem deste pacote é que dicionários para diferentes idiomas encontram-se livremente disponíveis no repositório do projeto.

Apesar de os dicionários disponibilizados serem de domínio geral, estes podem ser ampliados ou especializados mediante a inserção de termos de um domínio específico. Esta especialização do dicionário, muitas vezes, se torna necessária (como no caso deste trabalho) para que as palavras componentes do vocabulário específico de um domínio sejam corretamente anotadas. A Seção 6.4.1 descreve o processo de ampliação realizado neste trabalho.

O dicionário probabilístico utilizado pelo rotulador morfológico (*PoS-tagger*) do *Apertium* foi obtido a partir dos mesmos pacotes linguísticos que os dicionários morfológicos originais, previamente mencionados.

O *Apertium* é, ainda, utilizado em trabalhos relacionados [12], o que permite que informações sobre sua utilização, otimização, desempenho e recursos sejam obtidas, além de validar sua utilização no contexto do trabalho.

### 6.4.1 Ampliação do Dicionário

Como apresentado na Seção 4.3, o *Lttoolbox* (pacote do *Apertium* para análise morfológica)[4] utiliza dicionários externos para realizar a anotação morfológica dos documentos. Estes dicionários são compostos por palavras, juntamente com suas respectivas formas canônicas e flexões gramaticais (gênero e número).

Os dicionários padrão do *Apertium* podem ser obtidos a partir do repositório do projeto[5] . Os dicionários monolíngues encontram-se inclusos juntamente com os dicionários bilíngues nos pacotes de tradução. Sendo assim, no pacote *apertium-es-pt*, por exemplo, encontram-se os dicionários morfológicos monolíngues dos idiomas português e espanhol.

Inicialmente, cogitou-se a utilização dos mesmos dicionários utilizados em Caseli e Nunes [12], no entanto, testes preliminares demonstraram que estes não continham palavras e expressões multipalavras consideradas importantes ao domínio, optando-se assim pela expansão dos dicionários padrão com o vocabulário considerado de domínio extraído do corpus construído.

Sendo que a identificação das expressões multipalavras é realizada nesta etapa (seguindo o processo padrão do *Apertium*), constatou-se a necessidade da ampliação dos dicionários com a inserção do vocabulário dos manuais de usuário, obtido através da utilização das ferramentas *ExATOlp* [29] e *TTC TermSuite* [44]. O processo de extração deste vocabulário é apresentado na Seção 4.5.

Como apenas as versões compiladas dos dicionários utilizados por Caseli e Nunes [12] foram obtidas, estes não puderam ser ampliados, sendo assim, optou-se por realizar a ampliação dos dicionários disponibilizados no repositório do *Apertium*.

Para a língua portuguesa, foi utilizado o dicionário "*pt_BR.dix*" proveniente do pacote "*apertium-es-pt*" (versão 1.1.0), enquanto para o inglês foi utilizado o dicionário "*apertium-en-ca.en.dix*" contido no pacote "*apertium-en-ca (versão 0.8)*".

Durante a ampliação foram adicionados 78.062 formas superficiais ao dicionário da língua portuguesa, sendo esta adição composta por 41.757 expressões multipalavras e 36.307 palavras simples. Já para o dicionário da língua inglesa esta adição foi menor, tendo sido adicionadas 10.408 palavras das quais 6.466 expressões eram expressões multipalavras e 3.937 palavras simples.

Apesar de o número de entradas adicionadas ter sido inferior ao adicionado por Caseli e Nunes [11], as entradas adicionadas são mais específicas ao domínio, tendo sido extraídas a partir dos próprios documentos do corpus.

[4] http://sourceforge.net/projects/apertium/files/lttoolbox/
[5] http://sourceforge.net/projects/apertium/files/?source=navbar

A estrutura de um dicionário do *Apertium* é composta, basicamente por três tipos de definição:

- Alfabeto e conjunto de símbolos: Conjunto de símbolos a serem encontrados no dicionário, dentre os quais destacam-se os rótulos morfológicos utilizados na anotação das palavras;
- Paradigmas: Padrões de anotação que podem ser aplicados a diferentes palavras com o intuito de otimizar a organização do dicionário. Por exemplo, sendo que as palavras *carro* e *passo* possuem a mesma flexão de número (introdução do *s* para o plural), cria-se um paradigma para esta flexão que passa a ser empregado por ambas as palavras, evitando sua duplicação;
- Entradas: Nesta seção são declaradas as palavras (simples e compostas) em conjunto com suas formas canônicas, rótulos morfológicos (*PoStags*) e flexões como número (singular e plural) e gênero (masculino e feminino), entre outras.

A ampliação dos dicionários morfológicos padrão do *Apertium* [15] consistiu na adaptação das saídas dos programas de extração de terminologia (previamente mencionados) para o formato padrão dos dicionários utilizados por esse.

Durante esta adaptação, os rótulos morfológicos utilizados pelas ferramentas de extração tiveram de ser convertidos para o padrão utilizado pelo

*Lttoolbox*[6] . A notação utilizada pela ferramenta *ExATOlp* [29] pode ser visualizada em: http://beta.visl.sdu.dk/visl/pt/info/symbolset-floresta.html\#.

Vale ressaltar que, uma vez que tanto a extração dos termos quanto sua inserção foram realizadas de forma automática, os dicionários resultantes não foram otimizado com a utilização de paradigmas. A utilização deste tipo de recurso encontra-se listada como um dos possíveis trabalhos futuros.

A não utilização dos paradigmas pode implicar que varições das palavras (alguma flexão não prevista) não sejam reconhecidas, diminuindo assim a abrangência do dicionário. No entanto, pelo vocabulário ter sido extraído a partir dos textos paralelos empregados no processo, suas formas superficiais (forma pela qual a palavras se apresenta no texto) serão identificadas podendo, assim, seus equivalentes multilíngues serem extraídos.

Outra questão que deve ser ressaltada é que, nem sempre, a ferramenta de extração terminológica empregada fornecerá todos os dados necessários. O *TTC Term Suite*, por exemplo, fornece apenas o rótulo morfológico (*PoStag*), de forma que apenas esta informação será utilizada para o vocabulário do domínio.

Apesar de existirem rotuladores morfológicos para a língua inglesa, estas ferramentas, em geral, não realizam o reconhecimento de expressões multipalavras ou mesmo a priorização dos sintagmas extraídos. Sendo assim, preferiu-se utilizar o *TTC TermSuite*[44].

O *TTC TermSuite* baseia-se em uma abordagem híbrida, sendo assim, a extração da terminologia consistiu na aplicação automática dos padrões sintáticos já definidos pela ferramenta

[6]http://wiki.apertium.org/wiki/List_of_symbols

precedido pela aplicação manual de filtros de frequência para os quais, baseado no trabalho de Lopes *et al.* [30] utilizou-se uma frequência mínima de 5 ocorrências.

Mesmo o rótulo gramatical (*PoStag*) atribuído pelo extrator pode não estar correto ou, ainda, não ser determinado (principalmente em expressões multipalavras). Para esta etapa, estimasse que poderia ser utilizada a estratégia proposta por Tiedeman [47], na qual três etiquetadores são utilizados em um sistema de votação, sendo que o rótulo que tiver o maior número de votos é atribuído a palavra avaliada. Esta abordagem segue como dica de melhoria para trabalhos futuros.

Expressões multipalavras, no entanto, devem ser conectadas de forma que a ferramenta de alinhamento léxico (*Giza++*) considere-as como unidades durante o alinhamento de palavras. Assim como em Caseli e Nunes [11], o símbolo "_" foi utilizado para conectar as palavras como em "*mobile_phone*", por exemplo.

Após ampliados, os dicionários morfológicos foram compilados utilizando-se, para esta finalidade, a ferramenta "*lt-comp*" pertencente ao *Lttoolbox*. No processo de compilação os dicionários, até então definidos no formato *XML*, foram compilados em arquivos binários, identificados pela extensão *".bin"*.

### 6.4.2 Anotação Morfológica

A análise morfológica foi então conduzida utilizando o *"lt-proc"*[7] ferramenta componente do *Lttoolbox*, em conjunto com os dicionários morfológicos ampliados (descritos na Seção 6.4.1). Por fim, o rotulador gramatical (*PoS-tagger*) do *apertium-tagger* foi utilizado para determinar o rótulo gramatical mais adequado para cada palavra, principalmente para casos nos quais mais de um rótulo encontrava-se disponível. Estas duas ferramentas foram utilizadas em conjunto na forma de um *pipeline* (sequência) como demonstrado pelo exemplo a seguir:

```
"cat <arquivo_entrada>.txt | lt-proc <dicionario morfologico> |
 apertium-tagger -gp  <dicionario probabilístico>.prob > <arquivo_saida>.txt"
```

No exemplo anterior, "*cat*" representa um programa para leitura de arquivos de texto no terminal do Linux. Assim como na etapa anterior, os arquivos componentes do corpus foram analisados de forma individual, gerando como saída um conjunto composto pelos manuais morfologicamente anotados.

[7] http://wiki.apertium.org/wiki/Lttoolbox

## 6.5 Alinhamento Lexical

Seguindo os trabalhos de Caseli e Nunes [12] e Tiedemann [49] optou-se pela adoção de uma abordagem estimativa para o alinhamento lexical, conduzida com a utilização da ferramenta *Giza++* (versão 2.0) [39].

Como o conjunto inicial de recursos é escasso, sendo constituído por um corpus paralelo de tamanho médio, o treinamento do modelo foi realizado sobre o corpus inteiro, utilizando a configuração padrão do *Giza++* (iterações dos modelos IBM-1, IBM-3, IBM-4 e HMM), sem a utilização de dicionários externos, da mesma forma que os trabalhos de Caseli e Nunes [12] e Tiedemann [47].

Como anteriormente mencionado, uma das limitações do alinhamento padrão no *Giza++* é que uma palavra fonte só pode ser alinhada a uma palavra alvo (alinhamento 1:1), dificultando a identificação de expressões multipalavras como no caso de "celular" e "*mobile phone*" [39].

Para diminuir os efeitos desta limitação, foi adotada a técnica de simetrização proposta por Och e Ney [39] e implementada por Caseli e Nunes [12] e Tiedemann [47]. Sendo assim, o corpus foi lexicalmente alinhado em ambos os sentidos (português/inglês, inglês/português) gerando alinhamentos sobre os quais foi aplicado o algoritmo de união [39] visando maior abrangência.

A Tabela 6.2 apresenta um exemplo de um mesmo alinhamento léxico executado em ambas as direções, extraído a partir de um arquivo de saída da ferramenta *Giza++* (versão 2.0):

Tabela 6.2 – Exemplo de alinhamento léxico bi-direcional.

| **Direção** | **Alinhamento** |
|---|---|
| Português → Inglês | # Sentence pair (3) source length 6 target length 5 alignment score : 3.13643e-08<br>*Android mobile<adj> technology<n><sg> platform<n><sg> 2.3 <num><br>NULL ( ) Plataforma<n><f><sg> ( 1 ) de<pr> ( ) tecnologia<n><f><sg> ( 3 4 ) móvel<adj><mf><sg> ( 2 ) *Android ( ) 2.3<num> ( 5 ) |
| Inglês → Português | # Sentence pair (3) source length 5 target length 6 alignment score : 1.11139e-08<br>Plataforma<n><f><sg> de<pr> tecnologia <n><f><sg> móvel <adj><mf><sg> *Android 2.3<num><br>NULL ( 2 ) *Android ( 1 ) mobile<adj> ( ) technology<n><sg> ( 3 ) platform<n><sg> ( 4 5 ) 2.3<num> ( 6 ) |

De acordo com a saída da ferramenta *Giza++* (versão 2.0), a primeira linha de um alinhamento, identificada pelo símbolo "#" contém informações como, por exemplo, o número da linha das sentenças envolvidas "(3)", o tamanho de ambas as sentenças (em número de símbolos) e, por fim, a pontuação atribuída ao alinhamento "3.13643e-08" de acordo com métricas da ferramenta utilizada.

Na segunda linha do alinhamento, é inserida a sentença alvo (de destino), sem qualquer formatação específica. Por fim, a terceira linha apresenta a sentença fonte juntamente com as informações alinhamento, de forma que cada unidade léxica apresenta um índice para seu correspondente no texto paralelo (segunda linha). A entrada ("*technology<n><sg> ( 3 )*") por exemplo, indica que a tradução para a palavra "*technology*" encontra-se na terceira posição da sentença fonte.

A entrada "NULL ( )" (início da terceira linha do alinhamento) representa o alinhamento vazio (*empty*), que indicará todas as palavras (ou símbolos) da sentença alvo para as quais nenhum equivalente pode ser estabelecido na sentença fonte.

Como saída, esta etapa produz dois arquivos de alinhamento sendo um na direção fonte/alvo e o outro na direção alvo/fonte. Esses são identificados pelo sufixo "*.A3.final*" e posteriormente utilizados pelo algoritmo de união.

O algoritmo de união apenas complementa os alinhamentos existentes nos arquivos anteriores, sendo assim, a saída desta etapa consiste em dois arquivos contendo os alinhamentos léxicos, identificados pelo sufixo anteriormente mencionado.

## 6.6 Indução do Léxico

Após o alinhamento léxico, a próxima etapa do processo consiste na extração das equivalências estabelecidas e na utilização destas para a indução do léxico do domínio, representadas nos arquivos produzidos na seção anterior. Para esta finalidade, foi selecionada a ferramenta *ReTraTos* [11] (anteriormente apresentada).

O primeiro passo do processo de indução consiste na conversão dos arquivos de saída do *Giza++* (sufixo "*A3.final*") para o formato de entrada do *ReTraTos*. Esta conversão foi realizada mediante a utilização do *script* "*Giza_to_Lihla.pl*"[8].

Para cada um dos arquivos de saída do Giza++ o *script* mencionado gera dois arquivos de alinhamento, sendo que um representa os alinhamentos fonte/alvo enquanto o outro representa o alinhamento alvo/fonte, como mostrado na Tabela 6.3. Como pode ser observado nessa, os dois arquivos realizam um mapeamento entre os alinhamentos contidos em cada arquivo de saída do *Giza++* (de forma separada).

Os arquivos gerados são então utilizados como entrada para o *ReTraTos*, que realiza a indução do léxico a partir desses. Durante esta indução são levadas em consideração todas as informações morfossintáticas disponíveis para a palavra e seu equivalente. Como saída, o ReTraTos produz um dicionário bilíngue seguindo o modelo padrão da ferramenta *Apertium*.

A Figura 6.6 apresenta um exemplo de entrada bilíngue induzida pelo ReTraTos e armazenada no modelo padrão do *Apertium*.

[8] http://wiki.apertium.org/wiki/Talk:ReTraTos

Tabela 6.3 – Exemplo de entrada da ferramenta *ReTraTos*.

| **Fonte** | **Entrada** |
|---|---|
| Alinhamento *Giza++* | # Sentence pair (21) source length 4 target length 5 alignment score : 8.22616e-08 Como<adv><itg> usar<vblex><inf> a<pr> tela_de_toque<n><sg> 23<num> NULL ( ) Use<vblex><ger> ( 1 2 ) the<det><def><sp> ( 3 ) *touchscreen ( 4 ) 22<num> ( 5 ) |
| *ReTraTos* fonte → alvo | <s snum=20>Use<vblex><ger>:1 the<det><def><sp>:3 *touchscreen:4 22<num>:5</s> |
| *ReTraTos* alvo → fonte | <s snum=20>Como<adv><itg>:1 usar<vblex><inf>:0 a<pr>:2 tela_de_toque<n><sg>:3 23<num>:4</s> |

```
<e>
»       <p>
»       »       <l>google<b/>search<s n="n"/></l>
»       »       <r>a<s n="pr"/><j/>pesquisa<b/>do<b/>google<s n="np"/><s n="sg"/></r>
»       </p>
</e>
```

Figura 6.2 – Exemplo de entrada do dicionário.

Na Figura 6.6 a entrada entre os marcadores "<l>" e "<\l>" representa a palavra fonte enquanto a entre "<r>" e "<\r>" representa seu equivalente na linguagem alvo. Sendo assim, quando a palavra fonte for encontrada no contexto representado pela notação de substantivo esta será traduzida para a forma alvo com as características representadas por ela.

# 7. AVALIAÇÃO E RESULTADOS

Este capítulo apresenta a estratégia utilizada para avaliar os dicionários bilíngues extraídos pelo processo (descrito no Capítulo 6). Esse encontra-se organizado de modo que a Seção 7.1 descreve o procedimento (manual e intrínseco) utilizado na avaliação de uma amostra do dicionário inglês → português, e a Seção 7.2 apresenta e discute os resultados da avaliação. Ainda na Seção 7.2, é realizada uma comparação dos resultados encontrados com os apresentados em trabalhos de referência.

## 7.1 Avaliação do Vocabulário Bilíngue

O desempenho do processo proposto foi medido através da avaliação dos dicionários produzidos por esse, levando em consideração o número de entradas para as quais traduções corretas foram definidas. Para tal, utilizou-se um procedimento de avaliação manual e intrínseca, no qual avaliadores humanos classificaram as entradas selecionadas em categorias predefinidas, sem considerar o contexto sentencial (logo intrínseca).

Nesta seção são apresentados o objetivo da avaliação (Seção 7.1.1), a amostra considerada (Seção 7.1.2), os procedimentos e escala utilizados na avaliação manual (Seção 7.1.3) e, por fim, a métrica de desempenho adotada (Seção 7.1.4).

### 7.1.1 Objetivo

O foco da avaliação aqui apresentada foi determinar o desempenho (precisão) do processo proposto na extração de equivalências/correspondências bilíngues a partir de corpus paralelo. Ou seja, em avaliar as entradas do dicionário (traduções), sem levar em consideração as sentenças das quais foram extraídas. A utilização do vocabulário (dicionários) na tradução de sentenças (avaliação extrínseca) e a avaliação do processo de identificação dos termos de domínio não foram realizadas, sendo considerados como trabalhos futuros.

Vale ressaltar que o desempenho das ferramentas empregadas, bem como a qualidade das listas de termos utilizadas para a ampliação dos dicionários (obtidas a partir de ferramentas de extração terminológica) não foram avaliadas. Em relação às ferramentas utilizadas ao longo do processo, o desempenho foi considerado a partir do divulgado em suas publicações de referência.

### 7.1.2 Conjunto de Avaliação

Considerando o tamanho (em número de palavras) do vocabulário multilíngue extraído pelo processo, optou-se pela construção de subconjuntos sobre os quais as avaliações foram conduzidas. Esses foram construídos a partir do dicionário inglês $\rightarrow$ português.

O conjunto de avaliação utilizado foi composto por 500 entradas extraídas de forma automática e aleatória do dicionário gerado ao final do processo (arquivo "*.dix*"). Dessas, 400 entradas constituíam palavras simples e 100 expressões multipalavras.

Cada entrada do conjunto de avaliação foi composta por uma entrada do dicionário inglês acompanhada de suas respectivas informações morfológicas (rótulo gramatical e flexões de número e gênero), seguida pela tradução a essa atribuída (entrada do dicionário português) e suas respectivas informações morfológicas. Estas entradas foram organizadas no formato de planilhas eletrônicas, repassadas aos juízes humanos.

Vale ressaltar que, além das informações morfológicas, nenhum contexto foi fornecido aos avaliadores humanos durante a avaliação, tendo em vista que são estas as informações utilizadas pela ferramenta (*Apertium* [15]) para a determinação das possíveis traduções.

### 7.1.3 Processo de Avaliação

Devido à falta de listas de referência (*golden standards*), a avaliação foi conduzida de forma manual por dois juízes humanos. Como juízes, foram selecionados um linguista e um profissional da área de informática (de acordo com terminologia técnica do domínio selecionado). A estratégia de avaliação é apresentada a seguir.

Em relação ao conjunto de avaliação, das 500 entradas selecionadas 250 foram atribuídas para cada um dos juízes humanos (conjuntos distintos), das quais 200 constituíam palavras simples e as demais (50), expressões multipalavras. Um conjunto extra de 50 entradas (22 palavras simples e 28 palavras compostas), também aleatoriamente selecionadas de forma automática, foi construído para a avaliação do nível de concordância entre os juízes.

O nível de concordância foi medido através do cálculo de *Kohen's Kappa [28]*, para o qual obteve-se um valor 0,57. Apesar desse ser considerado baixo, uma avaliação dos casos de discordância entre os juízes demonstrou que essas não desvalidavam a avaliação (de acordo com sua natureza técnica). O mesmo problema relacionado ao baixo nível de concordância pode ser observado no trabalho de Caseli e Nunes [12], para o qual um valor de 0,47 foi encontrado.

A escala de avaliação utilizada pelos avaliadores para categorizar as equivalências bilíngues é similar à empregada por Caseli e Nunes [12]. Essa é composta por 3 categorias:

- Válida(V): Equivalências corretamente determinadas (tradução correta);

- Parcialmente Válida(PV): Equivalências parcialmente corretas nas quais a modificação de alguma das informações complementares (rótulos gramaticais, por exemplo) tornaria-a correta (válida);

- Não-Válida(NV): Equivalências não corretamente determinadas.

A principal diferença entre as escalas utilizadas nos dois trabalhos encontra-se na categoria "Parcialmente-Válida". Originalmente, essa compreende equivalências nas quais existe alguma informação morfológica diferente entre a palavra fonte e a palavra alvo (gênero, por exemplo). No presente trabalho esta categoria foi ampliada, podendo ser utilizada para classificar entradas nas quais o juiz não conseguiu determinar se a palavra era Válida (V) ou Não-Válida (NV).

### 7.1.4 Métricas de Desempenho Utilizadas

Na avaliação dos resultados do processo foi utilizada apenas a métrica de precisão. Demais métricas frequentemente utilizadas como acurácia, abrangência (*recall*) e *Medida-F* (*F-measure*) não puderam ser utilizada devido à falta de material de referência (*golden standard*) como listas e dicionários bilíngues, por exemplo. O cálculo de precisão utilizado é apresentado pela fórmula abaixo.

$$Precisão = \frac{Numero\ de\ Corretos}{Numero\ de\ Corretos\ +\ Numero\ de\ Incorretos}$$

Na formula apresentada, "*Número de Corretos*" representa o número de entradas na tabela classificadas como Válidas (V), enquanto "*Número de Incorretas*" correspondem à aquelas entradas classificas como Parcialmente-Válidas (PV) e Não-Válidas (NV). A categoria de Parcialmente-Válidas foi considerada como incorreta uma vez que, apesar de não serem completamente incorretas, também não podem ser consideradas como corretas. Sendo assim, optou-se por classificá-las desta forma.

## 7.2 Resultados

Esta seção apresenta uma contabilização do vocabulário extraído pelo processo proposto (Seção 7.2.1) e os resultados da avaliação de parte de suas entradas (conforme descrito na Seção 7.1). Para uma análise mais detalhada, as entradas avaliadas foram divididas em três seções: palavras simples (Seção 7.2.2), expressões multipalavras (Seção 7.2.3) e avaliação geral (Seção 7.2.4). Ainda na Seção 7.2.4, os resultados são comparados com os apresentados em trabalhos de referência.

### 7.2.1 Vocabulário Extraído

Conforme apresentado no Capítulo 6, a saída do processo proposto consiste em dois vocabulários bilíngues estruturados de acordo com o formalismo dos dicionários da ferramenta *Apertium* [15], um na direção de tradução português → inglês e o outro na direção inglês → português. O tamanho desses, em relação à quantidade de entradas (palavras e expressões multipalavras), é apresentado na Tabela 7.1.

Tabela 7.1 – Tamanho dos dicionários construídos (quantidade de entradas).

| **Dicionário** | **Palavras Simples** | **Exp. Multilavaras** | **Total** |
|---|---|---|---|
| Português → Inglês | 11.299 | 8.727 | 20.026 |
| Inglês → Português | 9.517 | 1.136 | 10.653 |

Como pode ser observado na Tabela 7.1, os dicionários apresentam tamanhos diferentes, sendo que o dicionário português → inglês é maior do que o dicionário inglês → português (em quantidade de entradas). Este fato decorre do alinhamento léxico conduzido pela ferramenta *Giza++* não ser simétrico [47], ou seja, nem sempre as mesmas equivalências são encontradas ao inverter-se a direção do alinhamento. Esta assimetria pode manifestar-se tanto no estabelecimento de diferentes equivalências, quanto na quantidade de equivalências identificadas.

De modo a estimar a abrangência dos dicionários bilíngues em relação ao vocabulário do domínio, realizou-se uma intersecção do dicionário português → inglês com as listas anotadas utilizadas para a ampliação dos dicionários morfológicos na terceira etapa do processo (Seção 4.3). Optou-se pela utilização do dicionário português → inglês nesta intersecção devido à lista de referência ser composta por um número maior de entradas.

A intersecção da lista com o dicionário resultou em um conjunto composto por 11.766 entradas (comuns à ambas as duas fontes). Esta lista corresponde a 58,75% das palavras do dicionário português → inglês (composto por 20.026 entradas) e a 15,07% da lista de referência (composta por 78.076 entradas). A partir dos valores encontrados pôde-se perceber que, apesar da abrangência do dicionário em relação à lista poder ser considerada baixa (15,07%), a maior parte das entradas do dicionário relacionam-se ao domínio (58,75%). Este resultado vai de encontro ao objetivo de extrair traduções para vocabulários específico do domínio.

Durante a análise das entradas dos dicionários bilíngues pôde-se perceber, ainda, a repetição de entradas ocasionada por erros durante a anotação morfológica. Assim, algumas entradas tem sua forma plural registrada tanto graficamente (palavra escrita) quanto por intermédio de anotação morfológicas (através de rótulos). No entanto, estima-se que este tipo de erro não tenha grande impacto durante a tradução pois a entrada selecionada dependerá do anotador morfológico utilizado (apenas uma alternativa será selecionada).

### 7.2.2 Palavras Simples

A Tabela 7.2 apresenta a quantidade de palavras simples classificadas em cada uma das três categorias consideradas (Seção 7.1.3). A categorização é apresentada de acordo com o conjunto total de palavras simples (soma dos conjuntos de ambos os juízes).

Tabela 7.2 – Avaliação de palavras simples.

| **Categoria** | **Número de Palavras** |
|---|---|
| Válidas | 324 (81,00%) |
| Parcialmente-Válidas | 26 (6,50%) |
| Não-Válidas | 50 (12,50%) |
| **Total** | 400 |

Conforme a Tabela 7.2, para a amostra selecionada a precisão do processo na extração de equivalências bilíngues foi de 81% (classificadas como *Válidas*). Das demais entradas avaliadas, 6,5% foram categorizadas como parcialmente válidas (PV), indicando erros em informações morfológicas associadas (problemas relacionados à etapa de anotação morfológica) e 12,5% foram consideradas como completamente erradas (Não-Valias).

Uma análise conduzida sobre as entradas classificadas como Parcialmente-Válidas revelou que muitas dessas foram assim classificadas devido a falhas na identificação expressões multipalavras em casos de tradução do tipo *1:n*, nos quais a expressão multipalavras correspondente foi parcialmente reconhecida, ou seja, apenas parte da expressão foi considerada. Como exemplo deste tipo de erro pode-se apresentar a palavra "*constant*" (inglês), por exemplo, que deveria ter sido reconhecida como "*symbolic constant*" para que a tradução fosse considerada como correta (Válida). Das 26 entradas classificadas nesta categoria (PV), 10 apresentaram esta característica. Vale ressaltar que a amostra de palavras simples foi coletada considerando o idioma fonte (inglês), ou seja, seus respectivos equivalentes bilíngues (português) podem ser compostos por mais de uma palavra.

### 7.2.3 Expressões Multipalavras

A Tabela 7.3 apresenta a quantidade de expressões multipalavras classificadas em cada uma das três categorias consideradas. Assim como para as palavras simples, foi considerado o conjunto completo de expressões multipalavras avaliadas (união dos conjuntos de ambos os juízes).

Tabela 7.3 – Classificação de expressões multipalavras.

| **Categoria** | **Quantidade** |
|---|---|
| Válidas | 39 (39%) |
| Parcialmente-Válidas | 41 (41%) |
| Não-Válidas | 20 (20%) |
| **Total** | 100 |

Conforme apresentado na Tabela 7.3, considerando-se apenas as entradas válidas como corretas, pode-se observar uma precisão de 39% na definição de equivalentes bilíngues para expressões multipalavras. Na tabela, pode-se observar, ainda, que o número de entradas classificadas como Parcialmente-Válidas (41) é maior do que o número de entradas classificadas como Não-Válidas (incorretas).

Uma investigação dos conjuntos de expressões multipalavras classificadas como Parcialmente-Válidas demonstrou que muitas dessas foram assim classificadas devido às entidades definida como seus prováveis equivalente encontrarem-se incompletas, ou seja, as expressões multipalavras identificadas como suas possíveis traduções encontram-se parcialmente identificadas (assim como ocorrido na avaliação das palavras simples). A Tabela 7.4 apresenta exemplos deste tipo de problema. No total, foram identificados 25 casos (das 41 entradas classificadas como PV) nos quais este problema foi responsável por tornar a equivalência inválida.

Tabela 7.4 – Exemplos de problemas relacionados a expressões multipalavras.

| **Palavra** | **Tradução** |
|---|---|
| "guia de uso diário" | "*guide*" |
| "abrir menu" | "*open* " |
| "novo alarme" | "*alarm*" |
| "menu ajuda" | "*help*" |

A Tabela 7.4 apresenta exemplos de entradas do dicionário que foram alinhadas a partes de suas equivalências corretas. A entrada "*abrir menu*", por exemplo, foi alinhada a palavra "*open*", que constitui uma parte da expressão "*open menu*", uma das possíveis equivalência corretas da expressão "*abrir menu*". Sendo assim, sentenças paralelas nas quais as entradas "*abrir menu*" e "*open*" co-ocorrem foram avaliadas demonstrando que, na maioria dos casos, a palavra "*open*" era seguido por "*menu*", indicando uma falha no reconhecimento da expressão.

Como apresentado na literatura [39] e previamente discutido na Seção 4.4, durante o alinhamento léxico, a ferramenta *Giza++* estabelece equivalências do tipo *n:1*, ou seja, a ferramenta reconhece apenas expressões multipalavras (não previamente identificadas) na sentença fonte, realizando o alinhamento dessas apenas com palavras simples da sentença alvo. Assim sendo, busca-se realizar a identificação e união (dos elementos formadores) deste tipo de construção linguística antes do alinhamento léxico, mais precisamente na etapa de análise morfológica (Seção 4.3).

Baseando-se no tipo de erro encontrado nas expressões avaliadas, bem como na característica do alinhamento conduzido pela ferramenta *Giza++* e na necessidade de identificação prévia das expressões multipalavras, pôde-se constatar que muitos dos erros encontrados na definição de equivalências bilíngues para expressões multipalavras proveem de problemas na etapa de identificação deste tipo de construção linguística, durante a análise morfológica. Pôde-se perceber, ainda, que apesar de muitas entradas terem sido consideradas como incorretas do ponto de vista de tradução, grande parte dessas encontram-se corretas do ponto de vista do alinhamento léxico.

Por fim, constatou-se que o problema de identificação de expressões multipalavras aconteceu principalmente no corpus de língua inglesa, indicando uma baixa precisão da ferramenta utilizada

em relação ao tipo de corpus empregado (ruidoso). Como trabalhos futuros, ferramentas específicas para a extração de multipalavras como o *mwetoolkit* [42] serão experimentadas.

### 7.2.4 Desempenho Geral do Processo

Por fim, uma avaliação geral do sistema foi realizada levando em consideração o conjunto completo de entradas avaliadas (palavras simples e expressões multipalavras). A quantidade de entradas classificadas em cada uma das categorias é apresentada na Tabela 7.5.

Tabela 7.5 – Classificação geral da amostra.

| **Categoria** | **Quantidade** |
|---|---|
| Válidas | 363 (72,6%) |
| Parcialmente-Válidas | 67 (13,4%) |
| Não-Válidas | 70 (14%) |
| **Total** | 500 |

Conforme apresentado na Tabela 7.5, para a amostra avaliada (composta por 500 entradas), obteve-se uma precisão de 72,6%. Vale ressaltar que para este cálculo foram consideradas apenas as entradas classificadas como Válidas (V). Outro fato que pode ser observado na tabela é a proximidade entre o número de entradas classificadas como Parcialmente-Válidas (PV) e Não-Válidas (NV). Levando-se em consideração que várias das entradas consideradas como Parcialmente-Válidas (PV) foram assim classificadas devido a problemas na identificação de expressões multipalavras (como previamente apresentado) ou de anotação morfológica, pode-se perceber que o número de entradas classificadas como completamente erradas (Inválidas) é baixo (14%).

Como dicionário bilíngues de referência (*golden standads*) não encontravam-se disponíveis para uma avaliação ou comparação, buscou-se estimar o desempenho do processo proposto com base nos resultados apresentados em trabalhos relacionados. Desses, aquele que mais se assemelha ao presente trabalho (tanto no processo utilizado quando na avaliação conduzida) é o de Caseli e Nunes [12] (apresentado na Seção 5.1), selecionado como base de comparação. Dos demais trabalhos Ha *et ali.* [14] e Zhang [53] priorizam a avaliação da extração terminológica enquanto Tiedemann [49] não avalia o material coletado e Tiedemann [47] possui enfoque no alinhamento léxico.

A Tabela 7.6 apresenta a precisão obtida pelo processo de Caseli e Nunes [12] e pelo processo proposto neste trabalho. Vale ressaltar que, apesar dos processos serem muito similares, o conjunto de ferramentas utilizadas ao longo do processo difere, assim como o domínio do corpus empregado. Logo, apenas uma comparação estimativa pode ser realizada.

Tabela 7.6 – Comparação estimativa com trabalhos relacionados.

| **Trabalho** | **Palavras Simples** | **Exp. Multipalavras** | **Total** |
|---|---|---|---|
| Caseli e Nunes [12] | 83% | 38% | 75% |
| Processo proposto | 81% | 39% | 72,6% |

Como pode ser observado na Tabela 7.6 os resultados obtidos pelo processo proposto neste trabalho são bastante próximos aos encontrados no trabalho de Caseli e Nunes [11] apresentando uma uma precisão menor em relação as palavras simples (diferença de 2%) e uma pequena vantagem (1%) em relação a identificação de expressões multipalavras.

Apesar da comparação ser apenas estimativa, pode-se perceber que o desempenho do processo proposto encontra-se próximo ao encontrado por trabalhos da literatura. Uma comparação mais adequada com o trabalho de Caseli e Nunes [12], utilizando o mesmo corpus e processo de avaliação, encontra-se listada como um dos trabalhos futuros.

Uma segunda avaliação a ser conduzida em trabalhos futuros é da qualidade dos dicionários bilíngues produzidos. Como previamente mencionado, pôde-se observar que algumas das entradas foram desnecessariamente repetidas devido a problemas durante a análise morfológica e a ruídos resultantes do processo de conversão dos documentos para texto plano. Uma técnica de filtragem que apresentou bons resultados na eliminação deste tipo de entradas foi a intersecção dos dicionários com listas de terminologia do domínio, esta técnica será experimentada nos próximo passos do trabalho.

# 8. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho foi proposto um processo para a extração automática de vocabulários multilíngues a partir corpus paralelos. Esse teve como principal objetivo auxiliar na resolução de problemas encontrados na utilização de serviços de tradução de máquina durante reuniões de equipes multilíngues, dos quais destacam-se as traduções inconsistentes.

O processo proposto baseia-se em um conjunto de passos comuns aos trabalhos encontrados na literatura, diferenciando-se desses na forma pela qual realiza a etapa de análise morfológica (terceira etapa). Nessa, os dicionários morfológicos utilizados para a atribuição de informações morfológicas e identificação de expressões multipalavras são ampliadas com listas de palavras anotadas com entradas extraídas a partir do corpus bilíngue utilizado por ferramentas extratoras de terminologia (*ExATOlp* e *TTC TermSuite*). Esta abordagem teve como objetivo aumentar a abrangência dos dicionários morfológicos em relação ao domínio do corpus.

Para que o processo e as técnicas propostas pudessem ser testado e avaliados, um corpus paralelo bilíngue foi construído, considerando os idiomas português e inglês. Esse foi construído a partir de manuais de softwares, domínio relacionado aos objetivos do projeto de pesquisa no qual este trabalho encontra-se inserido. O corpus foi então processado (pelo processo proposto) gerando dois vocabulários bilíngues formatados de acordo com o formalismo dos dicionários da ferramenta *Apertium*, sendo um na direção *português* $\rightarrow$ *inglês* e outro na direção *inglês* $\rightarrow$ *português*.

Devido a falta de padrões de referência (*golden standads*), os dicionários produzidos foram manualmente avaliados (intrinsecamente) por dois juízes humanos e mensurados de acordo com a métrica de precisão. Ao final, obteve-se uma precisão de 81% para palavras simples, 39% para expressões multipalavras e 72,5% para a amostra completa (união das palavras simples e expressões multipalavras). Ainda que tenham sido comparados apenas de forma estimativa (devido a diferenças no processo e no domínio do corpus) com os resultados apresentados por Caseli e Nunes [12], esses valores demonstraram-se concisos com os encontrados na literatura.

Como principal problema encontrado, destacaram-se os erros na identificação de expressões multipalavras, nos quais apenas parte dessas eram reconhecidas e alinhadas a suas correspondências bilíngues. Este tipo de erro foi responsável por grande parte das equivalências terem sido classificadas como incorretas. Sendo assim, novos experimentos utilizando ferramentas específicas para a extração deste tipo de estrutura (*mwetoolkit [42], por exemplo*) serão realizados em etapas futuras do trabalho.

Por fim, destaca-se que o presente trabalho gerou um processo para a extração de vocabulário bilíngue e dois dicionários bilíngues para os idiomas português e inglês. Apesar do vocabulário não abranger todas as áreas do domínio escolhido (documentação de software), o processo proposto encontra-se preparado para utilização em novos corpus que venham a ser construídos ao longo do projeto.

## 8.1 Relevância ao Projeto

Em relação ao projeto no qual este trabalho encontra-se inserido, as principais contribuições desse foram os recursos linguísticos coletados e gerados, além do próprio processo proposto.

Quanto aos recursos linguísticos, os vocabulários coletados podem ser empregados tanto na ampliação dos dicionários do processo proposto quanto em outras etapas do projeto como na construção de corpus de nível mais técnico (na forma sementes de busca, por exemplo), na construção de ontologias, entre outros.

Os padrões técnicos (Engenharia de *Software*), por sua vez, podem ser utilizados como base (corpus) para a extração de padrões sintáticos a serem empregados na identificação de expressões multipalavras, por exemplo.

O vocabulário extraído pode ser utilizado tanto nas ferramentas de tradução quanto em funcionalidades de auxílio a escrita como, por exemplo, corretores ortográficos e auto-complementação (*autocomplete*).

O processo de extração, principal contribuição deste trabalho, foi proposto com o objetivo de possibilitar a extração do vocabulário multilíngue presente nos corpora construídos no decorrer das atividades do projeto.

## 8.2 Publicações e Participações em Eventos

Em relação a publicações, este trabalho deu origem a publicação de um artigo, intitulado "Extração de vocabulário multilíngue a partir de documentação de *software*", no evento Ontobras 2012 e um relatório técnico intitulado "*Real-Time Machine Translation for Software Development Teams*", disponível em [19].

O trabalho foi apresentado, ainda, em dois *workshops* sendo um relacionado do próprio projeto, e o outro referente ao projeto *Camaleon* (também relacionado a extração de vocabulário multilíngue).

## 8.3 Trabalhos Futuros

No decorrer do desenvolvimento deste trabalho, puderam ser identificados alguns pontos que poderiam ser melhor investigados ou então melhorados através da aplicação de técnicas diferentes das adotadas. Os principais pontos identificados são apresentados nas próximas seções.

### 8.3.1 Ampliação dos Dicionários Multilíngues

Como apresentado na Seção 6.4.1 os dicionários morfológicos padrão da ferramenta *Apertium* [15] foram ampliados com listas de palavras e expressões multipalavras geradas por *softwares* extratores de terminologia como, por exemplo, o *ExATOlp* [29] e o *TTC TermSuite* [44].

No entanto, o procedimento de ampliação realizado foi bastante simples não tendo sido aproveitados recursos de otimização tais quais os *paradigmas*, que além de diminuírem o tamanho do arquivo permitem que palavras para as quais o modelo de declinação seja conhecido, sejam complementadas com informações retiradas do paradigma definido a partir de palavras similares.

Neste sentido poderiam ser propostos métodos nos quais, baseado no conjunto de informações morfossintáticas fornecidas pelo extrator, fossem buscados paradigmas que pudessem ser empregados sobre cada nova entrada.

### 8.3.2 Avaliação das Ferramentas Empregadas

Como apresentado no Capítulo 4 para cada etapa do processo foram levantadas ferramentas capazes de implementá-las sendo que, para este trabalho, as mesmas foram selecionadas de acordo com testes bastante limitados.

Uma das consequências dos critérios de seleção adotados foi a detecção precária das palavras do vocabulário pelo *TTC TermSuites* que impactou negativamente na identificação das expressões multipalavras e, consequentemente no desempenho do processo.

Sendo assim, além da busca por ferramentas alternativas para a extração do vocabulário, uma avaliação das ferramentas propostas para cada seção de modo experimental pode auxiliar na escolha da opção mais adequada para o tipo de cenário encontrado.

### 8.3.3 Avaliação Extrínseca dos Dicionários

Neste trabalho, apenas uma avaliação intrínseca foi realizada sendo que esta não apresentou algumas das informações necessárias para que a pergunta de pesquisa pudesse ser respondida de forma definitiva.

Assim sendo, propõem-se a realização de uma avaliação extrínseca, na qual os dicionários extraídos a partir dos documentos sejam utilizados, em conjunto com a ferramenta *Apertium* [15] utilizados e avaliados na tradução de sentenças.

### 8.3.4 Comparação do Processo com Trabalhos de Referência

Como previamente mencionado, a comparação realizada na Seção 7.2.4 com o trabalho de Caseli e Nunes [12] foi apenas estimativa devido à diferenças no domínio dos corpus utilizados por ambos os trabalhos e pela falta de uma padronização do processo de avaliação.

Assim sendo, propõem-se como atividade futura a utilização de ambos os processos sobre um mesmo corpus bilíngue e sua avaliação através de um mesmo processo de avaliação, utilizando os mesmo juízes e amostras de mesmo tamanho.

# REFERÊNCIAS BIBLIOGRÁFICAS


[1] Audy, J. L. N.; Prikladnicki, R. "Desenvolvimento Distribuído de Software: Desenvolvimento de Software com Equipes Distribuídas". Série Livros Didáticos Campus-SBC, 2007, 211p.

[2] Baroni, M.; Bernardini, S. "Bootcat: Bootstrapping corpora and terms from the web." In: LREC, 2004, pp. 1313–1316.

[3] Berber Sardinha, T. "Lingüística de corpous: histórico e problemas". In: *DELTA*, 2000, vol. 16, pp. 323–367.

[4] Bird, S.; Klein, E.; Loper, E. "Natural Language Processing with Python". O'Reilly Media, 2009, 1 ed., 504p.

[5] Brown, P. F.; Lai, J. C.; Mercer, R. L. "Aligning sentences in parallel corpora". In: Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1991, pp. 169–176.

[6] Calefato, F.; Lanubile, F. "Using frameworks to develop a distributed conferencing system: an experience report", *Softw. Pract. Exper.*, vol. 39–15, october 2009, pp. 1293–1311.

[7] Calefato, F.; Lanubile, F.; Conte, T.; Prikladnicki, R. "Assessing the impact of real-time machine translation on requirements meetings: A replicated experiment". In: 6th Int'l Symposium on Empirical Software Engineering and Measurement (ESEM'12), 2012, pp. 19–20.

[8] Calefato, F.; Lanubile, F.; Prikladnicki, R. "A controlled experiment on the effects of machine translation in multilingual requirements meetings". In: Global Software Engineering (ICGSE), 2011 6th IEEE International Conference on, 2011, pp. 94 –102.

[9] Casacuberta, F.; Civera, J.; Cubel, E.; Lagarda, A. L.; Lapalme, G.; Macklovitch, E.; Vidal, E. "Human interaction for high-quality machine translation", *Commun. ACM*, vol. 52, october 2009, pp. 135–138.

[10] Caseli, H. M. "Alinhamento sentencial de textos paralelos português-inglês", Dissertação de Mestrado, ICMS-USP, 2003, 119p.

[11] Caseli, H. M. "Indução de léxicos bilíngües e regras para a tradução automática", Tese de Doutorado, ICMCUSP, São Paulo, Brazil, 2007, 186p.

[12] Caseli, H. M.; M.G.V., N. "Automatic induction of bilingual lexicons for machine translation". In: *International Journal of Translation*, 2007, vol. 19, pp. 29–43.

[13] Caseli, H. M.; Ramisch, C.; Graças Volpe Nunes, M.; Villavicencio, A. "Alignment-based extraction of multiword expressions", *Language Resources and Evaluation*, vol. 44, 2010, pp. 59–77.

[14] Fan, X.; Shimizu, N.; Nakagawa, H. "Automatic extraction of bilingual terms from a chinese-japanese parallel corpus". In: Proceedings of the 3rd International Universal Communication Symposium, 2009, pp. 41–45.

[15] Forcada, M. L.; Ginestí-Rosell, M.; Nordfalk, J.; O'Regan, J.; Ortiz-Rojas, S.; Pérez-Ortiz, J.; Sánchez-Martínez, F.; Ramírez-Sánchez, G.; Tyers, F. "Apertium: a free/open-source platform for rule-based machine translation", *Machine Translation*, vol. 25, 2011, pp. 127–144.

[16] Gale, W. A.; Church, K. W. "A program for aligning sentences in bilingual corpora". In: Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1991, pp. 177–184.

[17] Gomes, F.; Pardo, T.; Caseli, H. M. "Visualtca: Uma ferramenta visual on-line para alinhamento sentencial de textos paralelos". In: Anais do XXVII Congresso da Sociedade Brasileira de Computação - V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL), 2007, pp. 1729–1732.

[18] Gornostay, T.; Gojun, A.; Marion, W.; Ulrich, H.; Morin, E.; Daille, B.; Blancafort, H.; Sharoff, S.; Claude, M. "Terminology extraction, translation tools and comparable corpora: Ttc concept, midterm progress and achieved results". In: CREDISLAS 2012: Workshop on Creating Cross-language Resources for Disconnected Languages and Styles co-located with LREC 2012, 2012, pp. 35–39.

[19] Hilgert, L.; Calefato, F.; Lanubile, F.; Prikladnicki, R.; Vieira, R.; Finatto, M. J.; Termignoni, S. "Real-time machine translation for software development teams", Relatório Técnico, PUCRS, 2012.

[20] Hofland, K. "A program for aligning english and norwegian sentences". In: *Research in humanities computing*, Hockey, S.; Ide, N.; Perissinotto, G. (Editores), Oxford: Oxford University Press, 1996, pp. 165–168.

[21] IEEE Computer Society. "Software engineering body of knowledge (swebok)". Capturado em: "http://www.swebok.org/", Dezembro 2012.

[22] IEEE Computer Society. "Ieee standard glossary of software engineering terminology", *IEEE Std 610.12-1990*, 1990, pp. 1–84.

[23] IEEE Computer Society. "Ieee standard for software project management plans", *IEEE Std 1058-1998*, 1998, pp. 1–28.

[24] IEEE Computer Society. "Ieee standard for software quality assurance plans", *IEEE Std 730-2002 (Revision of IEEE Std 730-1998)*, 2002, pp. 1 –10.

[25] IEEE Computer Society. "Ieee standard for software configuration management plans", *IEEE Std 828-2005 (Revision of IEEE Std 828-1998)*, 12 2005, pp. 1 –30.

[26] IEEE Computer Society. "Systems and software engineering – vocabulary", *ISO/IEC/IEEE 24765:2010(E)*, 2010, pp. 1–418.

[27] Jacquemin, C.; Bourigault, D. "Term Extraction adn Automatic Indexing". Oxford [u.a.]: Oxford University Press, 2003, cap. 33, pp. 599–615.

[28] Koehn, P. "Statistical Machine Translation". New York, NY, USA: Cambridge University Press, 2010, 1 ed., 446p.

[29] Lopes, L.; Vieira, R.; Fernandes, P.; Couto, G. "Exatolp: extraction of language resources from portuguese corpora". In: International Conference on Computational Processing of the Portuguese Language - PROPOR, 2012, pp. 45–47.

[30] Lopes, L.; Vieira, R.; Finatto, M. J.; Martins, D. "Extracting compound terms from domain corpora", *Journal of the Brazilian Computer Society*, vol. 16, 2010, pp. 247–259.

[31] Lutz, B. "Linguistic challenges in global software development: Lessons learned in an international sw development division". In: Proceedings of the 2009 Fourth IEEE International Conference on Global Software Engineering, 2009, pp. 249–253.

[32] Ma, X.; Liberman, M. Y. "Bits: A method for bilingual text search over the web", *Proceedings of Machine Translation Summit VII*, 1999, pp. 538–543.

[33] Manning, C. D.; Schütze, H. "Foundations of Statistical Natural Language Processing". Cambridge, MA, USA: MIT Press., 1999, 1 ed., 620p.

[34] Melamed, I. D. "A geometric approach to mapping bitext correspondence". In: Conference on Empirical Methods in Natural Language Processing, 1996, pp. 1–12.

[35] Moore, R. C. "Fast and accurate sentence alignment of bilingual corpora". In: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, 2002, pp. 135–144.

[36] Morin, E.; Prochasson, E. "Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora". In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, 2011, pp. 27–34.

[37] Nakatsuka, M.; Yasunaga, S.; Kuwabara, K. "Extending a multilingual chat application: Towards collaborative language resource building". In: Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on, 2010, pp. 137 –142.

[38] Nazar, R. "Parallel corpus alignment at the document, sentence and vocbulary levels". In: *Procesamiento del Lenguaje Natural*, 2011, vol. 47, pp. 129–136.

[39] Och, F. J.; Ney, H. "A systematic comparison of various statistical alignment models", *Comput. Linguist.*, vol. 29–1, Mar 2003, pp. 19–51.

[40] Padró, L.; Stanilovsky, E. "Freeling 3.0: Towards wider multilinguality". In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012), 2012, pp. 2473–2479.

[41] Picht, H. "Corpora como ponto de partida para a extração de dados terminológicos". In: *Cadernos de Tradução. A Terminologia em Foco*, Instituto de Letras - UFRGS, 2004, pp. 67–77.

[42] Ramisch, C.; Villavicencio, A.; Boitet, C. "mwetoolkit: a Framework for Multiword Expression Identification". In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), 2010, pp. 662–669.

[43] Resnik, P.; Melamed, I. D. "Semi-automatic acquisition of domain-specific translation lexicons". In: Proceedings of the fifth conference on Applied natural language processing, 1997, pp. 340–347.

[44] Rocheteau, J.; Daille, B. "Ttc termsuite: A uima application for multilingual terminology extraction from comparable corpora". In: 5th International Joint Conference on Natural Language Processing, 2011, pp. 9–12.

[45] Sinclair, J. "Corpus and Text - Basic Principles". Oxford: Oxbow Books, 2005, cap. 1, pp. 1–16.

[46] Sommerville, I. "Software Engineering". Harlow, England: Addison-Wesley, 2010, 9 ed., 792p.

[47] Tiedemann, J. "Recycling translations: Extraction of lexical data from parallel corpora and their application in natural language processing", Tese de Doutorado, Uppsala University, Uppsala, Sweden, 2003, 142p.

[48] Tiedemann, J. "News from opus - a collection of multilingual parallel corpora with tools and interfaces". In: *Recent Advances in Natural Language Processing (vol V)*, Nicolov, N.; Bontcheva, K.; Angelova, G.; Mitkov, R. (Editores), Amsterdam/Philadelphia: John Benjamins, 2009, pp. 237–248.

[49] Tiedemann, J. "Parallel data, tools and interfaces in opus". In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Calzolari, N.; Choukri, K.; Declerck, T.; Doğan, M. U.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S. (Editores), 2012, pp. 2214–2218.

[50] Varga, D.; Németh, L.; Halácsy, P.; Kornai, A.; Trón, V.; Nagy, V. "Parallel corpora for medium density languages". In: Recent Advances in Natural Language Processing (RANLP 2005), 2005, pp. 590–596.

[51] Yamashita, N.; Inaba, R.; Kuzuoka, H.; Ishida, T. "Difficulties in establishing common ground in multiparty groups using machine translation". In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2009, pp. 679–688.

[52] Yamashita, N.; Ishida, T. “Effects of machine translation on collaborative work”. In: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work, 2006, pp. 515–524.

[53] Zhang, Y.; Nivre, J. “Transition-based dependency parsing with rich non-local features”. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2, 2011, pp. 188–193.