Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Alinhamento de textos bilíngues alemão hunsrückisch-português

Marcelo Yuji Himoro
Maria das Graças Volpe Nunes

NILC-TR-13-06

Novembro, 2013

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

O *hunsrückisch* constitui hoje a variedade de alemão mais falada no Brasil. Este trabalho tem como objetivo construir um corpus alinhado bilíngue alemão *hunsrückisch*-português brasileiro, e a partir dele, obter um léxico bilíngue que possa ser utilizado na construção de um sistema de tradução automática estatística (SMT) entre as duas línguas. Apesar do tamanho reduzido do corpus de trabalho, devido principalmente à escassez de material bilíngue, foi encontrada precisão de 81,89% e 84,5% para dois métodos diferentes de alinhamento lexical, valores próximos ao de outros trabalhos existentes na literatura.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AER: Alignment Error Rate

ALMA: Atlas Linguístico-Contatual das Minorias Alemãs na Bacia do Prata

GC: Gale & Church

CGI: Common Gateway Interface

en: inglês

es: espanhol

FAPESP: Fundação de Amparo à Pesquisa do Estado de São Paulo

hrx: alemão *hunsrückisch* ou hunsriqueano

IBM: International Business Machines

ICMC: Instituto de Ciências Matemáticas e Computação

IPOL: Instituto de Investigação e Desenvolvimento em Política Lingüística

MT: Machine Translation

NILC: Núcleo Interinstitucional de Linguística Computacional

PESA: Portuguese-English Sentence Alignment

PEWA: Portuguese-English Word Alignment

pt: português

RBMT: Rule-based Machine Translation

SIL: Summer Institute of Linguistics

SMT: Statistical Machine Translation (TA Estatística)

TA: Tradução Automática

TCA: Translation Corpus Aligner

UFRGS: Universidade Federal do Rio Grande do Sul

UNESCO: United Nations Educational, Scientific and Cultural Organization

USP: Universidade de São Paulo

CAPÍTULO 1: INTRODUÇÃO

1.1. Contextualização

A TA (Tradução Automática), ou MT (*Machine Translation*) em inglês, nasceu no fim da década de 40, impulsionada principalmente pela Guerra Fria e a demanda por traduções rápidas e baratas entre inglês e russo (Martins & Nunes, 2005). Ao longo dos anos, surgiram diversos paradigmas de TA, dentre os quais se destacam principalmente a TA baseada em regras (RBMT: *Rule-based Machine Translation*), utilizada em tradutores automáticos como o Systran[1] e o Apertium[2], e mais recentemente, a TA estatística (SMT: *Statistical Machine Translation*), utilizada pelo Google Translate.

Um RBMT é composto basicamente de um conjunto de regras sintáticas, e um léxico contendo as informações morfológicas, sintáticas e semânticas. A necessidade de conhecimentos linguísticos faz com que sua construção e manutenção sejam muito custosas (Lagarda et al., 2009). Os RBMT costumam ter melhor desempenho em domínios limitados, e essa previsibilidade é justamente o que torna mais simples a correção de erros nesses sistemas (Dove et al., 2012). O RBMT foi a primeira abordagem utilizada em TA, sendo, portanto, uma técnica já bastante madura (Lagarda et al., 2009).

Os SMT, por outro lado, utilizam modelos estatísticos para encontrar as traduções mais prováveis (Lagarda et al., 2009). Sua construção requer corpora paralelos - conjuntos de textos na língua fonte e sua respectiva tradução na língua alvo - suficientemente grandes para que se obtenham resultados satisfatórios. Esses recursos eram até então escassos e extremamente valiosos, mas graças à internet esse cenário está mudando, fazendo com que os SMT emerjam como uma solução bastante viável e pouco custosa, principalmente pelo fato de dispensar a intervenção de um linguista (Dove et al., 2012).

Para a construção de um SMT, é preciso que o corpus esteja alinhado sentencial e lexicalmente, isto é, que haja um mapeamento entre as sentenças e as palavras contidas nos

1 http://www.systransoft.com/

2 http://www.apertium.org/

textos na língua fonte e na língua alvo. O alinhamento pode ser realizado manualmente, ou automaticamente utilizando métodos linguísticos, empíricos ou híbridos. A partir do corpus alinhado, é possível criar um SMT utilizando *toolkits* como o GIZA++ (Och & Ney, 2003) e o Moses (Dyer et al., 2008).

Deve-se ressaltar, no entanto, que não há um consenso sobre a superioridade absoluta de qualquer um dos paradigmas em todos os contextos (Caseli, 2007). Há, inclusive, abordagens híbridas que procuram combinar o melhor de ambas as abordagens (Dove et al., 2012).

1.2. Motivação

Segundo a UNESCO[3] (Organização das Nações Unidas para a Educação, a Ciência e a Cultura), atualmente mais de 6.700 línguas no mundo se encontram em perigo de extinção. Uma delas é o *hunsrückisch* ou *hunsrik*, também chamado de hunsriqueano riograndense, falado por uma minoria alemã no sul do Brasil, principalmente nos estados de Santa Catarina e Rio Grande do Sul. Altenhofen (1996) define-o como "uma variedade supra-regional do alemão falado no sul do Brasil que tem por base um contínuo dialetal formado essencialmente pelo francônio-renano e pelo francônio-moselano, originários de áreas situadas na Renânia Central, e que recebem, no novo meio, uma forte influência do português e de outras variedades em contato." Trazido ao Brasil pelos imigrantes alemães que aqui se instalaram, constitui hoje a variedade de alemão mais falada no país (Altenhofen, 2004).

Não há uma cifra exata de quantas pessoas falam a língua atualmente, já que os censos atuais não coletam dados específicos sobre as línguas de imigração. Segundo Altenhofen (1996), com base em dados do BIRS (Bilingüismo no Rio Grande do Sul) de 1970, haveria, só no Rio Grande do Sul, cerca de 1.386.945 falantes de qualquer variedade do alemão. Em 1996, esse número estaria entre 700.000 e 900.000, dos quais 500.000 seriam falantes de *hunsrückisch*. O *Ethnologue: Languages of the Worl*d (Lewis et al., 2013) da SIL[4] (*Summer Institute of Linguistics*) estima o número de falantes em 3.000.000

3 http://www.unesco.org/

4 http://www.sil.org/

de pessoas dentre os 5.000.000 descendentes de alemães em todo o Brasil. Já segundo o IPOL[5] (Instituto de Investigação e Desenvolvimento em Política Linguística), a cifra englobando os falantes de qualquer variedade de alemão seria bem mais modesta: cerca de 200.000. Apesar de os números serem controversos, o que fica claro é que o alemão encontra-se em recesso no Brasil.

No sentido de preservar a língua, diversas iniciativas vêm sendo criadas. Nos últimos anos, houve um modesto crescimento no número de publicações escritas em *hunsrückisch*. Atualmente, dois dos cronistas mais prolíferos da comunidade são Leonídio Zimmerman, de Biguaçu (SC), e Pio Rambo, de São Sebastião do Caí (RS). Este último é autor dos textos que fazem parte do corpus utilizado neste trabalho. No que diz respeito a trabalhos no meio acadêmico, destaca-se o projeto ALMA[6] (Atlas Linguístico-Contatual das Minorias Alemãs na Bacia do Prata), vinculado ao Instituto de Letras da Universidade Federal do Rio Grande do Sul (UFRGS), cujos trabalhos em andamento são a redação de um atlas linguístico, a proposta de uma grafia supradialetal e a elaboração de um dicionário *hunsrückisch*-português-*hochdeutsch*[7]. A criação de recursos informáticos também poderia ajudar na preservação da língua.

1.3. Objetivos

Este trabalho tem como objetivo construir um corpus alinhado alemão *hunsrückisch*-português. A partir de um corpus pequeno (105 textos bilíngues), pretende-se construir recursos para a obtenção de um léxico bilíngue que possa ser utilizado na construção de um SMT entre as duas línguas. Para isso, é necessário alinhar sentencial e lexicalmente o corpus; ou seja, construir um mapeamento entre as sentenças e as palavras de cada texto que compõe o corpus. O modelo de probabilidade resultante dá origem a um léxico bilíngue.

5 http://www.ipol.org.br/

6 http://www.ufrgs.br/projalma/

7 Hochdeutsch: refere-se ao alemão padrão oficial na Alemanha.

Nota-se que o léxico bilíngue reflete a qualidade do corpus: quanto maior e mais representativo das duas línguas for o corpus, mais fiel será o léxico. Desde já sabemos que o corpus de trabalho é bastante reduzido, mas isso se deve, entre outros fatores, à escassez de material bilíngue. Além da eficácia já comprovada dos SMTs em relação aos tradutores simbólicos, outra razão que nos motivou é a possibilidade de eventualmente expandir o corpus gradativamente, com o uso dos léxicos produzidos em cada passo, e assim atingir um patamar que permita a geração de um SMT entre tais línguas.

1.4. Organização da Monografia

Este trabalho está organizado da seguinte maneira: no capítulo 2, serão apresentados alguns conceitos relevantes para o presente trabalho; no capítulo 3, são descritos os experimentos realizados e os resultados obtidos, e discutidas as limitações deste estudo e sua relevância para trabalhos futuros; finalmente, no capítulo 4, serão apresentadas a conclusão do trabalho e algumas considerações sobre o curso.

CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

2.1. Considerações Iniciais

Este capítulo apresenta alguns conceitos e informações essenciais para o entendimento do presente trabalho. Na Seção 2.2, são apresentados conceitos básicos de TA estatística e trabalhos relacionados. Na Seção 2.3, são apresentadas as ferramentas utilizadas no trabalho. Na Seção 2.4, por fim, são feitas as considerações finais.

2.2. Conceitos Relevantes e Trabalhos Relacionados

2.2.1. Tradução Automática Estatística

A Tradução Automática Estatística (TA estatística) ou SMT (*Statistical Machine Translation*) é um paradigma de tradução empírico, isto é, que utiliza pouca ou nenhuma teoria linguística para realizar a tradução (Specia & Rino, 2002). A ideia por trás da TA estatística é realizar a tradução a partir de dados estatísticos extraídos de corpora bilíngues ou memórias de tradução - bancos de dados contendo frases ou fragmentos de texto bilíngues. Um modo de calcular a probabilidade de uma sentença na língua fonte (*F*) ser traduzida em uma língua alvo (*A*) pode ser dada por uma variante da Regra de Bayes, expressa na equação abaixo (Dorr et al., 2000 apud Specia & Rino, 2002):

$$Pr(A|F) \cong Pr(A) * Pr(F|A)$$

Para extrair esses dados, é preciso que o corpus esteja alinhado sentencial e lexicalmente. Esses conceitos serão explicados com mais detalhes na Seção 2.2.2. A partir desses dados, também é possível construir recursos como gramáticas de tradução e léxicos bilíngues.

Uma das vantagens da TA estatística é o fato de dispensar a necessidade de formulação de regras gramaticais por parte de um linguista, o que a torna uma alternativa pouco custosa em relação a outros sistemas de tradução, além de abranger também particularidades linguísticas, como expressões idiomáticas (Dove et al., 2012). Outra vantagem é a facilidade de se estender tais sistemas, apenas alimentando-os com mais

textos bilíngues. Apesar disso, os resultados produzidos muitas vezes são gramaticalmente incorretos, e para que se obtenham resultados significativos, é necessária uma quantidade muito grande de textos com traduções de boa qualidade (Mateo & Rodríguez, 2012).

2.2.2. Alinhamento de Corpus Paralelo

O alinhamento de um corpus paralelo consiste basicamente em encontrar um mapeamento entre elementos de cada um dos textos na língua fonte e na língua alvo. Um bitexto, ou seja, um texto bilíngue, pode ser alinhado em nível de parágrafos, sentenças ou palavras (dito léxico). As abordagens de alinhamento podem ser classificadas em empírica, linguística e híbrida. Os métodos empíricos são aqueles que não dependem de informações linguísticas, utilizando-se apenas de estatísticas, como a frequência e a distribuição para realizar os alinhamentos (Silva, 2004). Os métodos linguísticos são aqueles que utilizam informações linguísticas, como léxicos, listas de palavras-âncora e etiquetação morfológica (Caseli, 2003). Já os métodos híbridos são aqueles que combinam abordagens empíricas e híbridas (Caseli, 2007).

Nesse contexto, há alguns trabalhos desenvolvidos especificamente para o português brasileiro. Caseli (2003) analisou métodos empíricos, linguísticos e híbridos para o alinhamento sentencial de um corpus português-inglês (pt-en). Silva (2004) analisou métodos empíricos e híbridos para o alinhamento léxico de um corpus português-inglês (pt-en). Caseli (2007), a partir do alinhamento sentencial e lexical de dois corpora - um, português-inglês (pt-en), e outro, português-espanhol (pt-es) -, induziu regras de tradução e léxicos bilíngues. A Tabela 1 mostra as métricas (vide Seção 2.2.4) encontradas nesses trabalhos para o alinhamento de corpora bilíngues entre o português brasileiro e o inglês e o espanhol.

Tabela 1: Métricas encontradas em trabalhos da literatura para o alinhamento léxico de corpora paralelos português brasileiro-inglês e português-brasileiro-espanhol

Corpus	**Precisão**	**Cobertura**	**Medida-F**
pt-en (Silva, 2004)	20,27%	92,93%	33,28%
pt-en (Caseli, 2007)	82,82%	86,38%	84,56%[8]

pt-es (Caseli, 2007)	93,26%	94,42%	93,83%[8]

A seguir, detalhamos alguns dos métodos utilizados neste trabalho.

2.2.2.1. Método Gale & Church

O Método GC (Gale & Church, 1991, 1993) foi um dos primeiros modelos de alinhamento sentencial. É um método empírico cuja ideia principal é a de que o tamanho das sentenças no texto fonte e no texto alvo estão fortemente relacionados: sentenças longas teriam traduções longas e sentenças curtas teriam traduções curtas.

2.2.2.2. Método Translation Corpus Aligner

O TCA (Santos & Oksefjell, 2000) é um método de alinhamento sentencial que utiliza tanto critérios empíricos, como o tamanho das sentenças e a detecção de padrões (cognatos), como também critérios linguísticos, tais como listas de palavras-âncora (pontuação, nomes próprios, etc.). Trata-se, portanto, de um método híbrido.

2.2.2.3. Modelos IBM 1 e 2

Os Modelos IBM 1 e 2 são métodos empíricos de alinhamento léxico. A diferença do Modelo 1 para o Modelo 2 é que, no primeiro, assume-se que todas as conexões entre uma palavra alvo (*A*) e cada uma das palavras da sentença fonte (*F*) são igualmente prováveis; ou seja, a ordem das palavras não importa no cálculo de *Pr(A|F)*. Já o Modelo 2 assume que a probabilidade dessas conexões depende não só da ordem, mas também do tamanho das duas *strings* (Brown et al., 1993).

Nesses modelos, a probabilidade de tradução de uma sentença fonte $F = (F_1, ..., F_{lF})$ em uma sentença alvo $A = (A_1, ..., A_{lA})$ com um alinhamento para cada palavra A_j da sentença alvo para uma palavra F_i da sentença fonte de acordo com a função $a: j \rightarrow i$ é dada por: $p(A,a|F) = \frac{e}{(l_F+1)^{l_A}} \prod_{j=1}^{l_A} t(A_j|F_{a(j)})$ para o modelo IBM 1 e

8 Valores estimados a partir das taxas de erro (AER: *Alignment Error Rate*) de 6,80% e 15,44% respectivamente.

$p(A,a|F)=e\prod_{j=1}^{l_A} t(A_j|F_a(j))a(a(i)|j,l_A,l_F)$ para o modelo IBM 2, onde *t* é a probabilidade de tradução, e *a* é a probabilidade de alinhamento (Koehn, 2010).

2.2.3. Léxicos bilíngues

Um léxico bilíngue é um recurso linguístico que fornece um ou mais equivalentes de uma palavra em uma língua fonte em uma determinada língua alvo (Mann & Yarowsky, 2001). Os léxicos bilíngues, além de serem de grande importância em muitos sistemas de TA, têm papel vital em diversas aplicações, como ferramentas de tradução assistida por computador, alguns métodos de alinhamento de corpora paralelos, recuperação de informação multilíngue, entre outras (Melamed, 1996).

Alguns trabalhos desenvolvidos relacionados à construção automática de léxicos bilíngues são (Wu & Xia, 1994), (Melamed, 1996), (Resnik & Melamed, 1997), (Mann & Yarowsky, 2001) e (Tufiş, 2001 & 2002). No contexto do português brasileiro, pode-se citar o trabalho de Caseli (2007).

2.2.4. Métricas de avaliação

Para avaliar o desempenho dos métodos de alinhamento, costuma-se utilizar três métricas: precisão, cobertura e medida-F. A precisão (Equação 2.1) mostra o número de alinhamentos corretos (*candidatos* ∩ *referência*) em relação aos alinhamentos encontrados (*candidatos*), a cobertura (Equação 2.2) mostra o número de alinhamentos corretos (*candidatos* ∩ *referência*) em relação aos alinhamentos corretos (*referência*), enquanto a medida-F (Equação 2.3), fornece uma média balanceada das duas métricas anteriores.

$$precisão(candidatos|referência)=\frac{|(candidatos \cap referência)|}{|(candidatos)|} \quad (2.1)$$

$$cobertura(candidatos|referência)=\frac{|(candidatos \cap referência)|}{|(referência)|} \quad (2.2)$$

$$medida-F=2\frac{cobertura \times precisão}{cobertura + precisão} \quad (2.3)$$

2.3. Ferramentas utilizadas

Nesta seção, serão brevemente apresentadas as ferramentas utilizadas no trabalho.

2.3.1. TCAalign

O TCAalign é uma ferramenta de alinhamento de textos paralelos em nível sentencial baseada no método TCA (Translation Corpus Aligner) (Hofland, 1996). Foi escrita em Perl por Helena de Medeiros Caseli no contexto do projeto PESA (Portuguese-English Sentence Alignment) (Caseli, 2003), desenvolvido pelo NILC no ICMC-USP.

Em (Caseli, 2007), foram avaliadas a precisão, a cobertura e a medida-F dos alinhamentos realizados pelo TCAalign para dois corpora contendo textos extraídos da revista científica Pesquisa FAPESP[9]: um português-espanhol (pt-es), contendo 1.050.924 *tokens* (504.130 em português e 546.794 em espanhol), e outro português-inglês (pt-en), contendo 1.038.638 *tokens* (504.387 em português e 534.251 em inglês). Os resultados podem ser observados na Tabela 2.

Tabela 2: Valores de métrica encontrados por (Caseli, 2007) no alinhamento de dois corpora utilizando o TCAalign.

Corpus	Precisão	Cobertura	Medida-F
pt-es	93,01%	95,85%	94,41%
pt-en	97,10%	98,23%	97,66%

No entanto, em testes realizados com o corpus de trabalho desse projeto, foram observados muitos alinhamentos incorretos (vide Seção 3.4.1, Tabela 5), mesmo em sentenças que já se encontravam naturalmente alinhadas. Por esse motivo, optou-se pelo uso da ferramenta hunalign, descrita na subseção seguinte.

2.3.2. hunalign

O hunalign (Varga et al., 2005) é uma ferramenta de código aberto para o alinhamento de textos paralelos em nível sentencial, desenvolvida por Dániel Varga do *Media Research Centre*, Departamento de Sociologia e Comunicações da Universidade de

9 http://revistapesquisa.fapesp.br/

Tecnologia e Economia de Budapeste. Escrita em linguagem C++, utiliza o método de Gale & Church (1991) em conjunto com um dicionário fornecido como entrada. Na ausência deste último, a ferramenta gera um dicionário a partir de alinhamentos feitos apenas pelo método GC, e então, alinha os textos novamente utilizando o dicionário gerado.

Em (Varga et al., 2005), foi analisada a precisão e a cobertura de alinhamentos realizados pelo hunalign em quatro corpora: os dois primeiros, contendo uma versão lematizada e uma não lematizada do texto bilíngue inglês-húngaro do romance "Nineteen Eighty-Four" de George Orwell (1984-HE-S e 1984-HE-U respectivamente); um segundo, contendo o texto bilíngue inglês-romeno não lematizado do mesmo romance (1984-REU), e finalmente, outro contendo o texto bilíngue inglês-húngaro lematizado do romance "Cup of Gold: A life of Sir Henry Morgan, Buccaneer, with Occasional Reference to History" de John Steinbeck (CoG-HE-S). Os resultados podem ser observados na Tabela 3.

Tabela 3: Valores de métrica encontrados por Varga et al. (2005) no alinhamento de três corpora utilizando o hunalign.

Corpus	Precisão	Cobertura
1984-HE-S	99,22%	99,24%
1984-HE-U	98,88%	99,05%
1984-RE-U	97,10%	97,98%
CoG-HE-S	97,03%	98,44%

2.3.3. LIHLA

O LIHLA é uma ferramenta de alinhamento de textos bilíngues em nível léxico escrita por Helena de Medeiros Caseli no contexto do projeto PEWA (Portuguese-English Word Alignment) (Caseli, 2007), desenvolvido no ICMC-USP pelo NILC. A partir de um corpus sentencialmente alinhado, a ferramenta realiza o alinhamento em nível léxico utilizando-se de léxicos bilíngues pontuados (ou probabilísticos) gerados pelo pacote de ferramentas NATools (vide Seção 2.3.4) e heurísticas independentes de língua, a fim de encontrar o melhor alinhamento entre palavras ou unidades multipalavra.

2.3.4. NATools

O NATools (Simões & Almeida, 2003) é um pacote de ferramentas para o processamento de corpora bilíngues desenvolvido pelo Departamento de Informática da Universidade do Minho. Inclui um alinhador em nível sentencial e léxico, um gerador de léxicos bilíngues pontuados e uma variedade de outras ferramentas. Neste projeto, foi utilizado apenas como ferramenta auxiliar para gerar os léxicos bilíngues pontuados a serem utilizados com o alinhador léxico LIHLA (vide Seção 2.3.3).

2.3.5. fast_align

O fast_align é uma ferramenta de código aberto para o alinhamento de textos bilíngues em nível léxico que implementa variantes um pouco melhoradas dos modelos de tradução léxica IBM 1 e 2 (vide Seção 2.2.2.3). Faz parte de um pacote maior de ferramentas, o cdec (Dyer et al., 2010), escrito em C++ com a colaboração de diversas pessoas, e que inclui também um decodificador, um *framework* de aprendizado para TA estatística e modelos de predição.

2.3.6. Yawat

O Yawat (Germann, 2008) é uma ferramenta *web-based* para visualização e manipulação de alinhamentos em nível sentencial e léxico escrita por Ulrich Germann da Universidade de Toronto. Implementado como um CGI-Perl no lado do servidor e em JavaScript no lado do cliente, facilita tanto a tarefa de alinhamento manual, possível com poucos cliques de mouse, como a de visualização, já que exibe os alinhamentos em forma de matriz, ou por realce movendo o mouse sobre as palavras.

2.4. Considerações Finais

Neste capítulo, foram apresentados uma breve introdução à TA estatística e ao alinhamento de corpus paralelo, alguns trabalhos relacionados e as ferramentas utilizadas. O capítulo seguinte abordará o desenvolvimento deste trabalho.

CAPÍTULO 3: DESENVOLVIMENTO DO TRABALHO

3.1. Considerações Iniciais

Neste capítulo, serão apresentadas as atividades desenvolvidas neste projeto. Serão descritos os passos seguidos na construção do corpus, no alinhamento e na geração do léxico bilíngue, bem como os resultados obtidos e as limitações deste trabalho.

3.2. Descrição do Problema

Este trabalho tem como objetivo construir um corpus paralelo português-*hunsrückisch* e, a partir do alinhamento sentencial e lexical do mesmo, gerar um léxico bilíngue que possa ser utilizado na construção de um sistema de tradução automática estatística (SMT) entre as duas línguas. A Figura 1 mostra uma visão geral dos passos seguidos no trabalho.

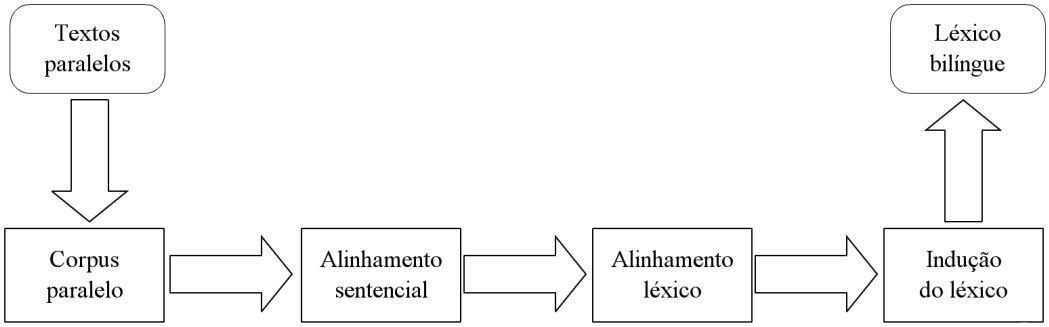

Figura 1: Visão geral dos passos seguidos no projeto.

A construção do corpus paralelo consiste em um pré-processamento dos textos paralelos, cujas etapas são: segmentação, edição, normalização e "tokenização". O corpus, então, é alinhado sentencialmente; isto é, as frases em um sentido são mapeadas a frases no outro sentido. O corpus alinhado sentencialmente é então alinhado lexicalmente; ou seja, os elementos de cada frase (palavras, sinais de pontuação, etc.) em um sentido são mapeados a elementos na frase correspondente no sentido oposto. Desses alinhamentos, obtém-se o léxico bilíngue resultante.

Na Seção 3.3, são descritos os métodos utilizados. Na Seção 3.4, são apresentados os resultados obtidos. Finalmente, na Seção 3.5, é feita uma análise dos resultados encontrados e são discutidas as principais dificuldades encontradas no desenvolvimento deste trabalho.

3.3. Descrição das Atividades Realizadas

3.3.1. Construção do corpus paralelo

Um problema com o qual nos deparamos à hora de reunir textos em *hunsrückisch* para a construção do corpus foi a multiplicidade de grafias. Além das variações naturais existentes na língua, por não haver uma norma para a escrita do *hunsrückisch*, cada autor escreve à sua maneira - uns, com grafias mais próximas do português; outros, mais próximas do alemão padrão. Uma palavra simples como "língua", por exemplo, pode ser escrita de diversas formas, de acordo com a pessoa, o local de origem ou conhecimento do alemão padrão: "sprach", "sprache", "xprach", "schprach", "sproch", "sprooch"...

Dentre os textos disponíveis em *hunsrückisch* com sua respectiva tradução em português, optou-se pelos textos do cronista Pio Rambo. Nascido na cidade de Harmonia (RS) e residente em São Sebastião do Caí (RS), Pio Rambo é eletrotécnico, radialista e músico. Como cronista, escreve em *hunsrückisch* em colunas de jornais do interior do Rio Grande do Sul desde 1995 (Alves Jr., 2013). Pio possui um *blog*[10], o "Língua Alemã Hunsrickisch - Deutsche Hunsrücker", onde publica textos antigos e novos, sempre em português e *hunsrückisch*, e muitas vezes com uma gravação contendo a leitura do texto em voz alta.

A opção por seus textos foi feita baseada nos seguintes critérios:

- estabilidade e maturidade da grafia;
- quantidade de bitextos disponíveis;
- disponibilidade de material digitalizado.

10 http://hunsrickisch.blogspot.com/

Ao todo, foram coletados 105 textos em prosa de tipo crônica. Foram excluídos textos em verso ou letras de música, bem como textos de outros autores quando publicados em outras grafias. Também foram excluídas listas de provérbios e frases feitas onde nem sempre havia uma correspondência direta entre os textos e as traduções, a fim de evitar ruídos. Posteriormente, foram coletadas 614 frases retiradas de postagens sobre gramática, e, por serem da mesma natureza dos demais textos, foram introduzidas no corpus. Alguns textos possuíam trechos sem tradução em português, mas o próprio Pio gentilmente se dispôs a fornecer as traduções faltantes. Da construção do corpus, portanto, sabe-se de antemão que todas as frases possuem um equivalente em ambos os sentidos.

Os textos em português foram verificados para erros utilizando um corretor ortográfico. Pelo pouco tempo disponível, os bitextos não foram revisados manualmente em nenhum dos sentidos. Em seguida, os bitextos foram normalizados e "tokenizados": todas as palavras foram transformadas em minúsculas (os modelos utilizados são *case-sensitive*) e os sinais de pontuação, separados por espaços. Alguns testes foram realizados aplicando *compound splitters* (separadores de palavras compostas) do alemão padrão nos textos em *hunsrückisch*, sem sucesso. O corpus também não pôde ser lematizado nem etiquetado, pois não há lematizadores nem etiquetadores para o *hunsrückisch*. Por fim, os bitextos foram segmentados; isto é, suas frases foram separadas de acordo com o formato de cada ferramenta (algumas utilizam separadores ou quebras-de-linha).

A Tabela 4 apresenta algumas estatísticas do corpus português-*hunsrückisch* (pt-hrx) construído. Um *token* é um símbolo qualquer do texto (uma palavra, um sinal de pontuação, etc.). Um *type* é um *token* único do texto. Um lema[11] é a forma canônica da palavra.

Tabela 4: Estatísticas do corpus pt-hrx construído.

	Português	**Hunsrückisch**
Nº de *tokens*	51.946	56.841
Nº de *types*	6.256	6.461

11 A distinção entre uma raiz e um lema nem sempre é clara. Foi utilizada a implementação de um lematizador de Porter (1997) feita pelo LABIC (ICMC-USP) (Caldas et al., 2001).

Nº de lemas	3.328	- [12]
Nº de sentenças	4.254	4.249

3.3.2 Alinhamento

O alinhamento sentencial foi feito utilizando duas ferramentas: o TCAalign (vide Seção 2.3.1) e o hunalign (vide Seção 2.3.2). Por meio de uma verificação das estatísticas dos alinhamentos e a avaliação do corpus de teste A (vide Seções 3.4.1 e 3.4.2), verificaram-se muitos erros nos alinhamentos gerados pelo TCAalign. Assim, optou-se por manter apenas os alinhamentos gerados pelo hunalign.

A partir do corpus alinhado sentencialmente, foi realizado o alinhamento léxico utilizando duas ferramentas: o LIHLA e o fast_align. Com o NATools, foi gerado o léxico bilíngue pontuado requerido pelo LIHLA. Uma vez gerados os alinhamentos lexicais em ambos os sentidos (pt-hrx e hrx-pt), foi feita a simetrização utilizando o simetrizador incluso no pacote cdec (vide Seção 2.3.5). Os alinhamentos lexicais foram avaliados através dos alinhamentos do corpus de teste B (vide Seção 3.4.3). A Figura 2 mostra um exemplo de alinhamento visualizado com a ferramenta Yawat.

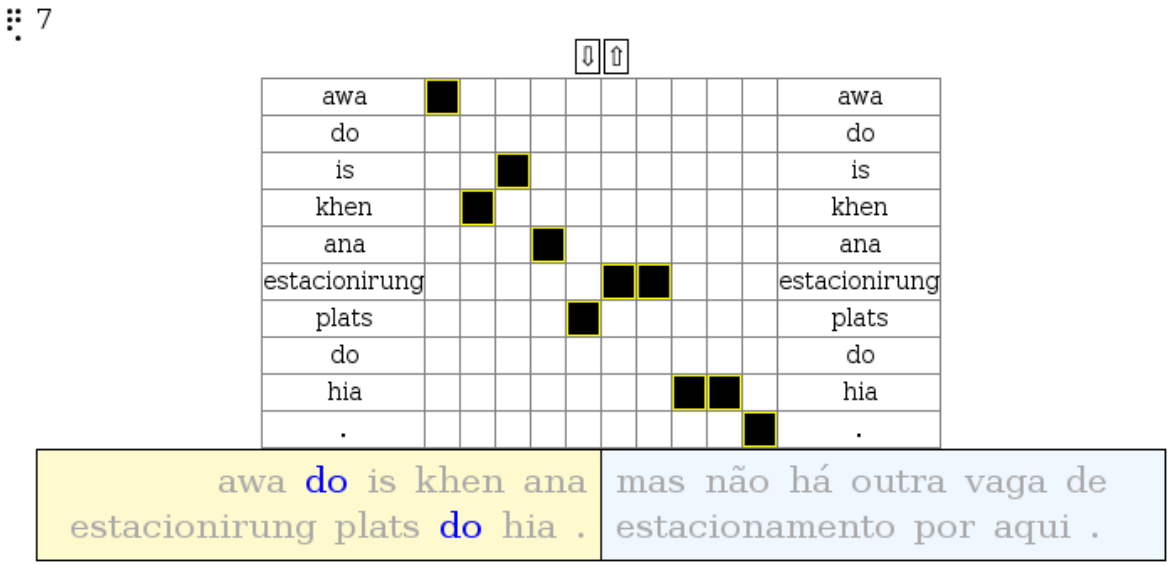

Figura 2: Tela da ferramenta yawat e a representação do alinhamento em matriz de alinhamentos.

12 Não existem lematizadores para o alemão *hunsrückisch*.

3.3.3. Geração do Léxico Bilíngue

Para a geração do léxico bilíngue, procurou-se seguir os passos desenvolvidos no projeto ReTraTos (Caseli, 2007):

P1. Leitura dos exemplos de tradução

1. Criação de um léxico bilíngue para o sentido fonte–alvo

2. Criação de um léxico bilíngue para o sentido alvo–fonte

3. União dos léxicos criados nos passos anteriores

4. Generalização das entradas do léxico bilíngue

5. (opcional) Tratamento de diferenças de gênero ou número

6. Tratamento de multipalavras

No passo P1, a partir dos alinhamentos obtidos, são lidos os exemplos de tradução. No passo 1, para cada palavra na língua fonte são procuradas as equivalências na língua alvo, e calculadas suas respectivas frequências de ocorrência. No passo 2, o mesmo é feito no sentido contrário. Assim, é criado um léxico no sentido fonte-alvo, e outro no sentido alvo-fonte. No passo 3, verifica-se se, para uma determinada palavra na língua fonte, o equivalente de maior frequência na língua alvo é também válido no sentido contrário; isto é, se o equivalente na língua alvo também tem como equivalente de maior frequência a palavra da língua fonte. Caso não sejam, as entradas são marcadas com "LR" (*left-right*, indicando que ela só é válida no sentido fonte-alvo) ou "RL" (*right-left*, indicando que ela só é válida no sentido alvo-fonte). Os passos 4 a 6 não puderam ser realizados, pois o corpus de trabalho não foi etiquetado morfossintaticamente (não existem etiquetadores para o *hunsrückisch*).

O léxico gerado, assim como no projeto ReTraTos, segue o formalismo utilizado pelo tradutor automático Apertium (Figura 3). A seção *alphabet*, utilizada para definir o alfabeto, *sdefs*, para definir os símbolos existentes no léxico, e *pardefs*, para definir os paradigmas, não foram utilizadas, e portanto, mantidas vazias no léxico. Cada entrada é

demarcada pelo elemento “e”, que contém um elemento “p” (par), por sua vez formado por dois elementos: “l”, indicando *left*, ou seja, a fonte; e “r”, indicando *right*, ou seja, o alvo.

```
<?xml version="1.0" encoding="UTF-8"?>
<dictionary>
   <alphabet/>
      <sdefs></sdefs>
      <pardefs></pardefs>
      <section id="main" type="standard">
         ...
         <e>
            <p>
               <l>gud</l>
               <r>bem</r>
            </p>
         </e>
         …
      </section>
</dictionary>
```

Figura 3: Estrutura do léxico gerado seguindo o formalismo do tradutor automático Apertium.

3.4. Resultados Obtidos

3.4.1. Estatísticas dos Alinhamentos

Nesta seção, são apresentadas as estatísticas dos alinhamentos encontrados pelas ferramentas utilizadas.

A Tabela 5 mostra estatísticas do alinhamento sentencial realizado pelo TCAalign e pelo hunalign. Como pode ser observado, o TCAalign encontrou grande quantidade (cerca de 51,51%) de alinhamentos do tipo 1:0 e 0:1. Um alinhamento 1:0 indica que uma determinada unidade (no caso, uma sentença) do texto na língua fonte não possui um equivalente no texto na língua alvo. Já um alinhamento 0:1 é justamente o contrário: uma determinada unidade do texto na língua alvo não possui equivalente no texto na língua fonte. Dada a natureza conhecida do corpus - aproximadamente a mesma quantidade de frases em ambos os sentidos e nenhuma omissão, ao menos a nível sentencial -, sabe-se de antemão que se trata de alinhamentos incorretos. Já o hunalign, não encontrou nenhum

alinhamento dos dois tipos. Portanto, optou-se por utilizar o corpus alinhado sentencialmente pelo hunalign para fazer os alinhamentos léxicos. A avaliação dos alinhamentos sentenciais será apresentada na Seção 3.4.2.

Tabela 5: Estatísticas dos alinhamentos sentenciais gerados pelo TCAalign e pelo hunalign.

Tipo	**Ferramentas**			
	TCAalign		**hunalign**	
1:0	1.321	25,94%	-	-
0:1	1.302	25,57%	-	-
1:1	1.674	32,87%	4.249	99,46%
1:2	406	7,97%	14	0,33%
2:1	389	7,64%	9	0,21%
Total	5.092	100%	4.272	100%

A Tabela 6 apresenta os alinhamentos léxicos encontrados pelo LIHLA e pelo fast_align. Observa-se que o fast_align encontrou maior variedade de tipos de alinhamentos em comparação ao LIHLA. A avaliação dos alinhamentos léxicos será apresentada na Seção 3.4.3.

Tabela 6: Estatísticas dos alinhamentos lexicais gerados pelo LIHLA e pelo fast_align.

Tipo	**Ferramentas**			
	LIHLA		**fast_align**	
0:1	5.321	8,86%	3.141	6,41%
1:0	9.553	15,91%	4.644	9,48%
1:1	42.263	70,4%	34.071	69,55%
2:1	1.837	3,06%	2.564	5,23%
1:2	835	1,4%	2.564	5,23%
2:2	36	0,06%	893	1,82%
2:3	6	0,01%	78	0,16%
2:4	-	-	6	0,01%

2:5	-	-	2	< 0,01%
3:1	112	0,19%	601	1,23%
3:2	3	< 0,01%	140	0,28%
3:3	3	< 0,01%	15	0,03%
3:4	-	-	1	< 0,01%
1:3	53	0,09%	154	0,31%
4:1	3	< 0,01%	72	0,15%
1:4	7	0,01%	10	0,02%
4:2	-	-	14	< 0,01%
4:3	-	-	1	< 0,01%
5:1	-	-	9	0,02%
1:5	-	-	4	< 0,01%
5:2	-	-	1	< 0,01%
5:3	-	-	1	< 0,01%
Total	60.032	100%	48.986	100%

3.4.2. Avaliação do alinhamento sentencial

No caso dos alinhamentos sentenciais, as estatísticas de alinhamento obtidas (Seção 3.4.1, Tabela 5) já forneciam indícios de que muitos dos alinhamentos obtidos pelo TCAalign eram incorretos, uma vez que, da construção do corpus, sabe-se que a maior parte dos alinhamentos corretos é do tipo 1:1 (ou seja, uma frase em *hunsrückisch* era mapeada a uma frase em português), e que há poucos alinhamentos do tipo n:m, com $n \neq m$, e n e/ou $m > 1$, e nenhum alinhamento do tipo 0:1 ou 1:0.

A fim de comparar o desempenho do TCAalign e do hunalign, foi criado o corpus de teste A, formado por aproximadamente 2,5% do corpus original não alinhado e mantida a proporção de textos e frases provenientes de exemplos gramaticais existente no corpus original. A Tabela 7 apresenta alguns dados estatísticos desse corpus.

Tabela 7: Estatísticas do corpus de teste A.

	Português	**Hunsrückisch**
Nº de *tokens*	1.414	1.617
Nº de *types*	564	565
Nº de lemas	482	-
Nº de sentenças	124	124

Para esse corpus, foram calculadas três métricas: precisão, cobertura e medida-F (Caseli, 2003). A Tabela 8 mostra os resultados obtidos. Os resultados mostram que o desempenho do hunalign foi muito superior ao do TCAalign. Os valores de 100% nas três métricas indicam que o método de alinhamento acertou todos os alinhamentos em comparação aos alinhamentos de referência. Esses altos valores são pouco usuais e podem ser explicados pelo tamanho reduzido do corpus de teste A.

Tabela 8: Valores encontrados para as métricas no alinhamento sentencial do corpus de teste A.

	TCAalign	**hunalign**
Precisão	79,7%	100%
Cobertura	84,13%	100%
Medida-F	82,49%	100%

3.4.3. Avaliação do alinhamento lexical

No caso dos alinhamentos lexicais, foi criado o corpus de teste B alinhado sentencialmente, contendo aproximadamente 2,5% do corpus original e mantida a proporção de textos e frases provenientes de exemplos gramaticais existente no corpus original. A Tabela 9 fornece mais detalhes a respeito do corpus de teste.

Tabela 9: Estatísticas do corpus de teste B.

	Português	**Hunsrückisch**
Nº de *tokens*	1.393	1.586

Nº de *types*	557	552
Nº de lemas	477	-
Nº de sentenças	123	123

Para esse corpus, foram calculadas três métricas: precisão, cobertura e medida-F (Caseli, 2003). A Tabela 10 mostra os resultados obtidos. Analisando os valores da tabela, nota-se que o fast_align apresentou um desempenho ligeiramente superior ao do LIHLA, considerando o equilíbrio entre cobertura e precisão (medida-F).

Tabela 10: Valores encontrados para as métricas no alinhamento lexical do corpus de teste B.

	LIHLA	**fast_align**
Precisão	84,5%	81,89%
Cobertura	63,33%	78,9%
Medida-F	72,4%	80,37%

Na avaliação, foram considerados os alinhamentos parcialmente corretos no alinhamento multipalavra. Apenas a título de ilustração, a Tabela 11 mostra a porcentagem de alinhamentos multipalavra corretos considerando alinhamentos parcialmente e totalmente corretos. Apesar de o desempenho do fast_align ter sido superior ao do LIHLA, a quantidade de alinhamentos multipalavra corretos ainda é muito baixa.

Tabela 11: Desempenho do LIHLA e do fast_align no alinhamento multipalavra.

	LIHLA	**fast_align**
Parcialmente corretos	10,56%	33,69%
Totalmente corretos	7,6%	10,08%

3.4.4. Léxico bilíngue

Nesta seção, são apresentados alguns exemplos de entradas do léxico gerado. Devido ao pouco tempo disponível, não foi possível realizar nenhum tipo de avaliação, ou mesmo filtragem das entradas do léxico, exceto pelo número de ocorrências.

A Figura 4 ilustra três exemplos de entrada do léxico; uma de cada tipo. As duas primeiras são equivalentes da palavra “sim” em *hunsrückisch*. A diferença entre elas está em que “ia” é a palavra de maior frequência em ambos os sentidos. A palavra “aham”, menos frequente, não pode ser generalizada como equivalente para “sim” em todos os casos, portanto, recebendo a marcação “LR”. O mesmo ocorre com a palavra “ieda”: “cada” é o equivalente de maior frequência. Como o equivalente de “qualquer” em *hunsrückisch* é “ieda”, mas o mesmo não se pode dizer do equivalente de “ieda” em português, a entrada recebe uma marcação “RL”.

```
<e>
   <p>
      <l>ia</l>
      <r>sim</r>
   </p>
</e>
```

```
<e r="LR">
   <p>
      <l>aham</l>
      <r>sim</r>
   </p>
</e>
```

```
<e>
   <p>
      <l>ieda</l>
      <r>cada</r>
   </p>
</e>
```

```
<e r="RL">
   <p>
      <l>ieda</l>
      <r>qualquer</r>
   </p>
</e>
```

Figura 4: Alguns exemplos de entradas válidas em ambos sentidos, LR e RL.

Como há muitos alinhamentos multipalavra incorretos e muitos deles só ocorrem uma única vez, é possível eliminar entradas como a da Figura 5 - a palavra “nommo” junto com uma vírgula foi considerada erroneamente uma unidade multipalavra e alinhada incorretamente (“nommo” significa “de novo” ou “novamente” em português) - simplesmente impondo um valor mínimo de ocorrência. Algumas entradas como a da Figura 6 - “ich” significa simplesmente “eu” - ainda permanecem por causa de alinhamentos incorretos repetidos várias vezes, dada a frequência com que ocorre “eu vou” em português.

```
<e r="LR">
   <p>
      <l>,+nommo</l>
      <r>de</r>
```

```
   </p>
</e>
```

Figura 5: Exemplo de entrada multipalavra incorreta de frequência baixa.

```
<e r="RL">
   <p>
      <l>ich</l>
      <r>eu+vou</r>
   </p>
</e>
```

Figura 6: Exemplo de entrada multipalavra incorreta de frequência alta.

Foram encontradas também inconsistências na grafia do autor, gerando algumas entradas duplicadas, como pode ser observado na Figura 7.

```
<e r="LR">
   <p>
      <l>iun</l>
      <r>rapaz</r>
   </p>
</e>
```

```
<e r="LR">
   <p>
      <l>iung</l>
      <r>rapaz</r>
   </p>
</e>
```

```
<e>
   <p>
      <l>iunn</l>
      <r>rapaz</r>
   </p>
</e>
```

Figura 7: Entradas duplicadas geradas por inconsistências na grafia.

3.5. Dificuldades, Limitações e Trabalhos Futuros

A principal dificuldade encontrada foi na construção do corpus: a multiplicidade de grafias em textos de diferentes autores e a dificuldade de se automatizar uma uniformização dessas grafias a fim de se obter mais bitextos fez com que o corpus de trabalho tivesse um tamanho muito restrito. Não existe até a data uma grafia supradialetal robusta e experimentada para representar de maneira satisfatoriamente uniforme as variedades do *hunsrückisch*.

Outra dificuldade encontrada é o fato de não haver ferramentas como lematizadores, etiquetadores e *compound splitters* (separadores de palavras compostas) específicos para o *hunsrückisch*. Os testes com os *compound splitters* existentes para o alemão padrão não se mostraram satisfatórios. Isso nos remete novamente ao problema da

não existência de uma grafia aceita pelos falantes da língua. Como já mencionado na Seção 1.2, há projetos em andamento no meio acadêmico nesse sentido.

Infelizmente, em razão do curto tempo de duração de um trabalho de conclusão de curso, não foi possível realizar nenhum tipo de avaliação ou filtragem das entradas do léxico obtido. O tamanho do corpus de trabalho, bem como algumas inconsistências nos textos originais, também influenciaram na quantidade e na qualidade das entradas. Uma pré-edição dos bitextos poderia colaborar para aumentar a corretude das entradas, uma vez que possivelmente ajudaria também a aumentar a precisão dos alinhamentos.

Um possível trabalho futuro poderia ser a geração de um léxico bilíngue pontuado português-*hunsrückisch* a partir de um corpus contendo textos de diferentes autores, a fim de comparar as diferenças de grafia para uma mesma palavra. Dessa forma, seria possível estabelecer correspondências entre diferentes grafias, e aproximar a grafia de textos de diferentes autores de maneira semiautomática, obtendo assim mais bitextos para a construção de um SMT.

3.6. Considerações Finais

Este capítulo abordou os objetivos do trabalho, apresentou a metodologia utilizada desde a construção do corpus até a geração do léxico bilíngue, bem como os resultados obtidos e as dificuldades encontradas. O capítulo a seguir discute as conclusões e as considerações sobre o curso.

CAPÍTULO 4: CONCLUSÃO

4.1. Contribuições

A proposta deste trabalho foi construir um corpus paralelo *hunsrückisch*-português, alinhá-lo sentencial e lexicalmente, e gerar um léxico bilíngue que pudesse ser utilizado na construção de um SMT entre as duas línguas.

Apesar do corpus de tamanho reduzido (105 textos), a precisão obtida nos alinhamentos lexicais foi próxima da encontrada em outros trabalhos da literatura. No entanto, pelo fato de haver muitos alinhamentos multipalavra incorretos, muitas das entradas do léxico bilíngue gerado são incorretas ou apenas parcialmente corretas. Algumas inconsistências na grafia do autor geraram entradas duplicadas, ressaltando ainda mais a dificuldade de se trabalhar com línguas não normativizadas e a necessidade urgente de uma grafia unificada para o *hunsrückisch*. Um pré-processamento dos bitextos poderia ter evitado parte dos problemas, mas isso requereria um tempo considerável, não usual ao escopo e à duração de um trabalho de conclusão de curso.

A respeito das contribuições pessoais ao aluno, o trabalho contribuiu para o desenvolvimento de habilidades de pesquisa, e também, para um primeiro contato com a área pesquisada.

REFERÊNCIAS

Altenhofen, C. V. Hunsrückisch in Rio Grande do Sul. Ein Beitrag zur Beschreibung einer deutschbrasilianischen Dialektvarietät im Kontakt mit dem Portugiesischen. Stuttgart: Steiner, 1996.

Altenhofen, C. V. A constituição do corpus para um "Atlas Lingüístico-Contatual das Minorias Alemãs na Bacia do Prata". In: Martius-Staden-Jahrbuch, São Paulo, n. 51, p. 135-165, 2004.

Alves Jr., O. D. Parlons hunsrückisch: Dialecte allemand du Brésil. L'Harmattan, 2013.

Brown, P. F.; Della Pietra, V. J.; Della Pietra, S. A.; Mercer, R. L. The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. 19, 2 (June 1993), p. 263-311, 1993.

Caldas Junior, J.; Imamura, C. Y. M.; Rezende, S. O. Avaliação de um Algoritmo de Stemming para o Língua Portuguesa. In the Proceedings of the 2nd Congress of Logic Applied to Technology, Vol. 2, pp. 267–274, 2001.

Caseli, H. M. Alinhamento sentencial de textos paralelos português-inglês. Dissertação (Mestrado em Ciências de Computação), Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, 2003.

Caseli, H. M. Indução de léxicos bilíngues e regras para a tradução automática. Dissertação (Doutorado em Ciências de Computação), Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, 2007.

Dorr, B. J; Jordan P. W.; Benoit, J. W. A Survey of Current Paradigms in Machine Translation. In M. Zelkowidtz (ed), Advances in Computers, Vol. 49, p. 1-68. Academic Press, London, 2000.

Dove, C.; Loskutova, O., de la Fuente, R. What's Your Pick: RbMT, SMT or Hybrid?. In: Proceedings of The Tenth Biennial Conference of theAssociation for Machine Translation in the Americas, 2012.

Dyer, C., Lopez, A., Ganitkevitch, J., Weese, J., Ture, F., Blunsom, P., Setiawan, H., Eidelman, V., Resnik, P.. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models. In: Proceedings of ACL, July, 2010.

Dyer, C.; Muresan, S.; Resnik, P. Generalizing Word Lattice Translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2008.

Gale, W. A.; Church, K. W. Identifying word correspondences in parallel texts. In: Proceedings of the 4th DARPA Speech and Language Workshop. Pacific Grove, CA:[s.n.], p. 152–157, 1991.

Germann, U. Yawat: Yet Another Word Alignment Tool. In: Proceedings of the ACL-08: HLT Demo Session, p. 20-23, 2008.

Hofland, K. A program for aligning English and Norwegian sentences. In: HOCKEY, S.; IDE, N.; PERISSINOTTO, G. (eds.). Research in Humanities Computing. Oxford: Oxford University Press. p. 165-178, 1996.

Koehn, P. Statistical Machine Translation, Cambridge University Press, 2010.

Lagarda, A. L.; Alabau, V.; Casacuberta, F.; Silva, R.; Díaz-De-Liaño, E. Statistical Post-Editing of a Rule-Based Machine Translation System. In Proceedings of NAACL HLT., pp. 217–220. Boulder, Colorado, 2009.

Lewis, M. P.; Gary F. S., Charles D. F. Ethnologue: Languages of the World, Seventeenth edition. SIL International, 2013. Disponível em <http://www.ethnologue.com/language/hrx>. Acessado em 26/10/2013.

Mann, G. S.; Yarowsky, D. Multipath translation lexicon induction via bridge languages. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01). Association for Computational Linguistics, Stroudsburg, PA, USA, 1-8, 2001.

Martins, R. T.; Nunes, M. G. V. Noções Gerais de Tradução Automática. NILC-TR-05-12, NOTAS DIDÁTICAS DO ICMC-USP (No.68), Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional, NILC - ICMC-USP, 2005.

Mateo, C. G.; Rodríguez, M. A. The Galician Language in the Digital Age: O Idioma Galego na Era Dixital. Springer (ed), 2012.

Melamed, I. D. Automatic construction of clean broad-coverage translation lexicons. In: Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996). Montreal, Canada: [s.n.], p. 125–134, 1996.

Och, F. J.; Ney, H. A Systematic Comparison of Various Statistical Alignment Model. Computational Linguistics Volume 29 Issue 1, p. 19-51, 2003.

Porter, M. F. An algorithm for suffix stripping. In Readings in information retrieval, Karen Sparck Jones and Peter Willett (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 313-316, 1997.

Resnik, P.; Melamed, I. D. Semi-automatic acquisition of domain-specific translation lexicons. In: ANLP. [S.l.: s.n.], p. 340–347, 1997.

Santos, D.; Oksefjell, S. An evaluation of the Translation Corpus Aligner, with special reference to the language pair English-Portuguese. In: Proceedings of the 12th "Nordisk datalingvistikkdager". Trondheim, Departmento de Lingüística, NTNU. p.191-205, 2000.

Silva, A. M. P. Alinhamento lexical de textos paralelos português-inglês. Dissertação (Mestrado em Ciências de Computação). Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos, 2004.

Simões, A., Almeida, J. J. NATools - A Statistical Word Aligner Workbench. Revista da SEPLN - Sociedade Española para el Procesamiento del Lenguaje Natural 31, p. 217-226, 2003.

Specia, L.; Rino, L. H. M.. Introdução aos Métodos e Paradigmas de Tradução Automática. NILC-TR-02-04, Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional, NILC - ICMC-USP, 2002.

Tufiş, D. A cheap and fast way to build useful translation lexicons. In Proceedings of the 19th international conference on Computational linguistics - Volume 1 (COLING '02), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 1-7, 2002.

Tufiş, D.; Barbu, A. Advances in Automation, Multimedia and Modern Computer Science, WSES, Press, p. 156-172, 2001.

Varga, D.; Németh, L.; Halácsy, P.; Kornai, A.; Trón, V.; Nagy, V. Parallel corpora for medium density languages. In Proceedings of the RANLP 2005, p. 590-596, 2005.

Wu, D.; Xia, X. Learning an English-Chinese lexicon from parallel corpus. In: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-1994). Columbia, MD: [s.n.], p. 206–213, , 1994.