UNIVERSIDADE DO SUL DE SANTA CATARINA

CARLOS EDUARDO DA SILVA

UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTOS EM CORPORA PARALELO PARA AUXÍLIO NA PESQUISA ACADÊMICA EM ESTUDOS DA TRADUÇÃO: UM ESTUDO DE CASO

Florianópolis

2014

CARLOS EDUARDO DA SILVA

UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTOS EM COPORA PARALELO PARA AUXÍLIO NA PESQUISA ACADÊMICA EM ESTUDOS DA TRADUÇÃO: UM ESTUDO DE CASO

Monografia apresentada ao Curso de pós-graduação *Lato Sensu* em Engenharia e Projetos de Software da Universidade do Sul de Santa Catarina como requisito parcial à obtenção do título de Especialista.

Orientador: Prof. Aran Bey Tcholakian Morales, Dr.

Florianópolis

2014

CARLOS EDUARDO DA SILVA

UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTOS EM COPORA PARALELO PARA AUXÍLIO NA PESQUISA ACADÊMICA EM ESTUDOS DA TRADUÇÃO: UM ESTUDO DE CASO

Esta monografia foi julgada adequada à obtenção do título de Especialista em Engenharia e Projetos de Software e aprovada em sua forma final pelo Curso de pós-graduação *Lato Sensu* em Engenharia e Projetos de Software da Universidade do Sul de Santa Catarina.

Florianópolis, fevereiro de 2014.

_____

Prof. e orientador Aran Bey Tcholakian Morales, Dr.
Universidade do Sul de Santa Catarina

_____

Profª. Vera Rejane N. Schuhmacher, Dra.
Universidade do Sul de Santa Catarina

_____

Profª. Maria Inés Castiñeira, Dra.
Universidade do Sul de Santa Catarina

Dedico este trabalho a minha esposa Danielle Amanda pelo apoio dado a mim e pela paciência nos finais de semana nos quais tive aula.

AGRADECIMENTOS

Eu gostaria de expressar a minha gratidão a todos que me acompanharam ao longo desta Especialização e que, de alguma forma, contribuíram para finalizar este estudo. Eu gostaria de agradecer de maneira especial as seguintes pessoas:

A minha esposa, **Danielle Amanda**, pelo seu amor, ajuda e incentivo para que eu nunca desistisse frente aos desafios encontrados no caminho;

Ao professor **Dr. Aran B. T. Morales**, por ter aceitado a orientação deste projeto e pela confiança depositada em mim;

A professora **Dra. Maria Inés Castiñeira**, pela avaliação e correção pontual desta monografia;

Ao curso de Curso de Pós-graduação *Lato Sensu* em Engenharia e Projetos de Software na pessoa da professora **Dra. Vera Schuhmacher**, pela paciência e o interesse no sucesso de todos os alunos do curso, o que foi de grande incentivo para todos;

Agradeço, também, aos meus colegas que tive durante a especialização, que muito me ajudaram a chegar até o final desta caminhada. Além disso, agradeço a todos os professores do curso que sempre se preocuparam em passar um conteúdo de qualidade, atual e relevante para a área de Engenharia de Software;

Agradeço a UNISUL por oferecer um bom ambiente de ensino e professores capacitados com conhecimentos científicos e mercadológicos da área de Engenharia de Software.

Nunca se deve engatinhar quando o impulso é voar.
(Helen Keller).

RESUMO

Apoiada em referências teóricas da Mineração de Textos, Processamento de Linguagem Natural e Estudos da Tradução com Base em Corpus, esta pesquisa teve por objetivo realizar um mapeamento da área de mineração de textos e discutir sua aplicação em um corpus paralelo para o estudo acadêmico de textos traduzidos do inglês para o português. Isso foi feito com o uso de algoritmos da mineração de textos aplicados a um corpus paralelo, ou seja, uma grande coleção de textos bilíngues (originais e suas respectivas traduções) pareados a nível sentencial. Assim, foi realizado um estudo de caso sobre o sistema COPA-TRAD (Corpus Paralelo de Tradução) a fim de mostrar como se deu a construção do corpus paralelo utilizado por este sistema e também como as técnicas de mineração de textos foram aplicadas nesse contexto. Com isso, espera-se auxiliar pesquisadores de Estudos da Tradução na investigação de fenômenos tradutórios que ocorrem em um texto traduzido uma vez que técnicas de mineração de textos foram utilizadas para extrair informações úteis desta coleção de textos em formato digital, dispensando o uso da etiquetação (*tagging*), manual ou automática, dos textos.

Palavras-chave: Mineração de Textos. Corpus Paralelo. Estudos da Tradução.

ABSTRACT

This study reviews Text Mining as an interdisciplinary field and discusses its application on a parallel corpus for the academic study of translated texts in the linguistic pair English-Portuguese. This was possible using Text Mining algorithms applied on a parallel corpus, i.e., a large collection of bilingual texts (originals and their translations) aligned on sentence level. The parallel corpus here envisaged is COPA-TRAD (*Corpus Paralelo de Tradução*). A case study about COPA-TRAD is presented to illustrate the application of text mining techniques on a parallel corpus. In addition, the key steps involved in the development of such technology are described. As a result, this study expects to provide a technological solution to support researchers from Translation Studies in the investigation of translational phenomena that occur in a translated text. Text mining techniques were used to extract useful information from the texts comprising COPA-TRAD database, avoiding the need of manual or automatic text labeling (or tagging), as well as, basic text filtering.

Keywords: Text Mining. Parallel Corpus. Translation Studies.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Este capítulo introdutório faz uma contextualização do cenário atual no que diz respeito ao crescente volume de dados digitais com o objetivo de localizar a área da mineração de textos. Em um segundo momento, alguns desafios da mineração de textos são apresentados dando especial ênfase a idiomas distintos do inglês. Na subseção seguinte, sugere-se a aplicação de técnicas de mineração de textos para auxiliar na investigação fenômenos tradutórios. Discutem-se, também, alguns dos desafios implícitos à aplicação da mineração de textos em um corpus paralelo para a tradução. Por último, são explicitados os objetivos gerais desta pesquisa e a organização estrutural da mesma.

1.1 CONTEXTUALIZAÇÃO DO CENÁRIO

A mineração de textos (do termo em inglês *text mining* ou ainda *text analytics*) é um "campo multidisciplinar que envolve a área de recuperação de informações, análise de textos, extração de informações, descoberta por agrupamento (*clustering*), categorização de textos, visualização de dados, banco de dados, aprendizado de máquina e mineração de dados" (TAN, 1999, p. 65). Este campo multidisciplinar ocupa-se "com a descoberta computacional de informações, novas ou previamente desconhecidas através da extração automática de informações de diferentes fontes textuais" (HEARST, 2003). A aplicação da mineração de textos (MT) vem despertando o interesse de instituições públicas e privadas. Este crescente interesse deve-se ao fato de que o volume de dados em formato digital tornou-se uma *commodity*[1] muito valorizada. Tal fenômeno se deve a, pela menos, dois aspectos fundamentais: o primeiro diz respeito às tecnologias de armazenamento de informação, que estão cada vez mais baratas, e o segundo aspecto seria o fato de que o volume de dados é abundante principalmente se considerarmos a transição de dados analógicos para digitais como, por exemplo, fotografias e músicas (GROSSMAN, 2012, p. 166-170).

Conforme observado por Grossman em seu livro *The Structure of Digital Computing: from mainframes to Big Data* (2012, p. 23-43), tal realidade começou a se concretizar na Era dos Computadores Pessoais (1980-2000), ou seja, quando houve um processo maciço de popularização dos computadores. A Era da Web (1995-2015), por sua

[1] Segundo SICULAR (2014), commodity é: 1 - algo útil que pode ser comercializado ou gerar alguma vantagem comercial. 2 - Um artigo que pode ser negociado ou comercializado. 3 - Uma vantagem ou benefício.

vez, vem contribuindo para a geração de dados através de serviços de mensagem eletrônica, redes sociais, blogs, jornais, comércio eletrônico, entre outros recursos disponíveis na Internet. Além disso, desde 2005 começamos a experimentar a Era dos Dispositivos Conectados a Internet (2005-2025) – Ou Internet das Coisas (*Internet of Things*) – período no qual uma gama de dispositivos (aparelhos celular, tablets, câmeras de segurança, carros, vídeo games, etc.) estão conectados a Internet, enviando e recebendo informações, bem como, utilizando soluções na nuvem. A passagem destes períodos importantes na tecnologia, leva o autor a concluir que estamos a caminho da Era do Conhecimento, na qual o volume de dados armazenados (nas eras antecessoras) será utilizado (GROSSMAN, 2012, p. 23-43).

Esta sociedade altamente conectada a Internet, através de tantos dispositivos eletrônicos, vem gerando um alto volume de dados que alimenta o *Big Data*. Manyika e outros (2011) informam que o *Big Data* é uma grande massa de dados que cresce de forma muito veloz sendo que pode ser analisada para gerar informação útil. Essa é uma tendência global e crescente.

Há anos atrás, a maior preocupação estava centrada em *como* armazenar dados, pois era algo muito caro. Nos últimos anos, contudo, o armazenamento de dados deixou de ser um problema crítico haja vista que em "1994 um disco rígido de um *terabyte* custava em torno de um milhão de dólares, em 2003 passou a custar três mil dólares, já em 2008 o valor era de duzentos e cinquenta dólares e em 2011 um disco de dois terabytes custava em torno de cem dólares" (GROSSMAN, 2012, p. 164).

Com a redução do custo de equipamentos para armazenar dados, consequentemente, um número muito maior de informações passou a ser guardada. De acordo com o documento intitulado "*Demystifying Big Data: A Practical Guide To Transforming The Business of Government*", elaborado *pela TechAmerica Foundation's Federal Big Data Commission* (2012), segundo dados liberados pela IBM, em 2011, foi gerado uma quantidade surpreendente de 1.8 zetabytes de informações em todo o mundo, ou seja, "o volume de dados gerado equivale a 200 bilhões de filmes em HD, sendo que cada filme tem duração de 2 horas e que uma pessoa levaria exatos 47 milhões de anos de forma interrupta para assistir a todos estes filmes". O mesmo relatório estima que esta quantidade de dados gigantesca duplique a cada ano.

O alto volume de dados desperta o interesse de empresas, instituições públicas e cientistas, pois, pode ser explorado de várias maneiras. Por exemplo, no mercado de ações, notícias em tempo real juntamente com a cotação da bolsa de valores podem servir de insumo para predizer, entre outras possibilidades, os movimentos do valor das ações, bem como,

estabelecer relações entre as notícias dos jornais e a cotação da bolsa de valores (FUNG et al., 2005, p. 1). Além disso, governos podem utilizar esta grande quantidade de dados para analisar como está a opinião dos cidadãos sobre determinado assunto ou para saber se a população está satisfeita ou não com determinado programa ou ação governamental. Este alto volume de dados pode, ainda, auxiliar sistemas de recomendação a determinar se estabelecimentos comerciais, como restaurantes, são bem avaliados pelo público ou não (KWOK e YU, 2013, p. 84).

Já no âmbito acadêmico, como por exemplo, na medicina ou biologia, um alto volume de dados coletados a partir de artigos científicos e textos especializados da área, podem auxiliar na descoberta de uma nova droga para tratar uma doença (KRALLINGER et al., 2005, p. 439). De acordo com Altman e outros (2008, p. 2), há uma grande demanda na utilização da mineração de textos na biologia, pois existe a necessidade de "traduzir as informações de um texto para um formato mais computável a fim de realizar ligações cruzadas com as informações de um banco de dados biológico".

Indo mais além, uma análise de um grande volume de dados pode, inclusive, predizer certas tendências e acontecimentos futuros: Pavlyshenko (2013) conduziu um estudo no qual analisou um grande volume de dados da rede social Twitter com o objetivo de descobrir qual seria o nome do filho do casal real. Com os resultados obtidos, foi possível predizer de forma consistente que o nome do filho do casal real Kate e William seria George e esta tendência foi confirmada.

1.2 O DESAFIO DA MINERAÇÃO DE TEXTOS

Este assunto é pertinente, mas ao mesmo tempo lança um macro desafio: Ao passo que a mineração de dados é independente da língua, a mineração de textos baseia-se em um volume de dados em formato textual em linguagem natural (TAN, 1999, p. 70). Chen (2001, p. 18) estima que 80% de toda a informação disponível na Internet está na forma de textos. Estes textos em linguagem natural, nos mais variados gêneros e idiomas, (CHEN, 2001, p. 34) estão disponíveis em formato não-estruturado (ver seção 2.3). A natureza heterogênea dos dados não-estruturados constitui-se em um desafio para a mineração de textos e para o Processamento de Linguagem Natural (PLN). Textos em linguagem natural são difíceis de lidar no processamento automático, pois apresentam ambiguidades lexicais, semânticas, sintáticas e pragmáticas. Conforme observado por Tan (1999, p. 70), "o métodos

de análise semântica são recursos computacionais caros e geralmente funcionam na ordem de poucas palavras por segundo". Além disso, o alto volume de textos em formato digital não está em um idioma único e por este motivo "línguas diferentes têm problemas diferentes e um vocabulário ativo bem diferente: na língua inglesa se utiliza apenas cerca de 800 palavras enquanto a língua alemã possui cerca de 4000 palavras no vocabulário ativo" (SCHNEIDER, 2001, p. 4). Os mecanismos computacionais de análise textual quando processam textos bilíngues ou multilíngues precisam fazer primariamente a identificação do idioma de cada texto para depois poder aplicar os algoritmos corretos.

1.3 A MINERAÇÃO DE TEXTOS, LINGUÍSTICA COMPUTACIONAL E TRADUÇÃO

Esta breve contextualização auxilia na compreensão da importância de extração de conhecimento útil de um alto volume de dados. Aqui é necessário chamar atenção para a "extração de conhecimento útil" e muitas vezes novo (isto é, que não se conhecia até então), pois este é o foco principal da mineração de textos (MOONEY e BUNESCU, 2005, p. 1).

Deste modo, conforme já mencionado, várias áreas podem beneficiar-se desta descoberta de conhecimento (MCDONALD e KELLY, 2012, p. 3). O conceito de extração de conhecimento útil pode valer tanto para a mineração de textos quanto para a mineração de dados (WITTEN, 2005, p. 2). A diferença é que a Mineração de Dados "pode ser caracterizada como a extração de informações úteis implícitas e previamente desconhecidas dos dados", estas informações implícitas podem ser extraídas com a utilização de técnicas automáticas da mineração de dados, que lidam em grande parte com *dados estruturados* (WITTEN, 2005, p. 2). Por sua vez, na Mineração de Textos (MT), "a informação a ser extraída está claramente declarada no texto" de natureza *não estruturada*, assim "o empenho da mineração de textos está em trazer informações de um texto que sejam adequadas para um computador sem o contato humano" (WITTEN, 2005, p. 2).

Além disso, um termo talvez menos conhecido para este processo de extração de conhecimento em textos é o de KDT ou *Knowledge Discovery in Texts* que, por sua vez, diferencia-se de KDD ou *Knowledge Discovery in Database*, que está relacionado à mineração de dados e a área de *business intelligence* (BI) (SILVA FILHO, 2009, p. 1).

Com vistas a exploração desses mecanismos de coleta de informação provenientes de textos, há um ramo conhecido como Linguística Computacional que "é a área de

conhecimento que explora as relações entre a linguística e a informática, tornando possível a construção de sistemas com capacidade de reconhecer e produzir informação apresentada em linguagem natural" (OTHERO, 2006, p. 342)". Othero (2006, p. 342) divide a linguística computacional em duas subáreas, a saber: "Linguística de Corpus e Processamento de Linguagem Natural (PLN)".

A Linguística de Corpus lida com o estudo de corpora eletrônico, ou seja, a análise sistemática de grandes coleções de textos em formato digital em linguagem natural. Oliveira (2009, p. 49) define um corpus (plural corpora) linguístico como "coleções de textos que ocorrem naturalmente na língua, organizadas sistematicamente para representar áreas de uso da língua e das quais podemos extrair informações" sendo que estes corpora estão em formato eletrônico de modo que sejam analisáveis por ferramentas computacionais (OTHERO, 2006, p. 342).

Já a PLN está envolvida com o "estudo da linguagem voltado para a construção de softwares, aplicativos e sistemas computacionais específicos" que tenham a capacidade de "interpretar e gerar informações em linguagem natural" (OTHERO, 2006, p. 343). Nesse sentido, Witten (2005, p. 2) afirma que a mineração de textos tende a "compreender toda a área de Processamento de Linguagem Natural e vai muito mais além". Além disso, a pesquisadora Hearst (1999, p. 4) em seu texto seminal *Untangling Text Data Mining*, ao localizar a mineração de textos informa que já existe um campo comprometido com a mineração de textos que é a linguística computacional com base em corpus. No entanto, Hearst (1999, p. 4) faz uma ressalva ao dizer que as técnicas utilizadas na linguística computacional são estritas, pois são aplicadas somente para satisfazer suas próprias necessidades e não se aplicam a um público mais amplo. Apesar do uso de técnicas de MT serem de uso estrito na linguística computacional, este estudo pretende mostrar que existe uma via de mão dupla entre estas duas áreas.

A linguística de corpus contribui para a tradução com base em corpus (ver seção 2.1), uma área do campo disciplinar de Estudos da Tradução (ET). No entanto "a maioria dos corpora disponíveis são monolíngues e em geral servem as necessidades da linguística" (KENNY, 2009, p. 59). Kenny (2009, p. 59) enfatiza que "no entanto, estudiosos da tradução, podem ter necessidades diferentes, por exemplo, em corpora que contém dados em mais de uma língua". Baker (1995, p. 224) assume que é animador o desenvolvimento de técnicas baseadas em corpus no campo da terminologia e tradução automática e que o uso de corpus vem a contribuir para áreas aplicadas de ET como, por exemplo, a formação de tradutores e a

crítica da tradução e em sentido mais amplo, ao fornecer uma explicação mais satisfatória dos fenômenos tradutórios.

Conforme observado em uma pesquisa conduzida no portal de periódicos da CAPES[2] em Janeiro de 2014, existem poucos estudos que utilizam a mineração de textos para áreas como a linguística, literatura e/ou tradução. Ao buscar pelas palavras-chave "mineração de textos + linguística" somente três resultados foram encontrados. Para as palavras-chave "mineração de textos + tradução" e "mineração de textos + literatura" nenhum resultado foi encontrado. Os resultados encontrados para a palavra chave "mineração de texto + linguística" foram de artigos apresentando uma aplicação da mineração de textos e sistemas de recomendação para ambientes virtuais de ensino.

1.4 PROBLEMATIZAÇÃO

Na área acadêmica de Estudos da Tradução, a tradução automática, corpora e memória de tradução são assuntos pesquisados que envolvem o amplo uso de recursos computacionais (HARTLEY, 2009, p. 106). No entanto, a aplicação específica da mineração de textos em conjunto com um corpus para investigar fenômenos tradutórios que ocorrem em um texto traduzido a partir de uma língua fonte não é algo tão comum, especialmente para o português.

Com base em Baker (1995, p. 230-235), acredita-se que a mineração de textos pode auxiliar como ferramenta na investigação das diferenças de um texto traduzido (ou alvo) em relação a um texto original (ou fonte). Pode auxiliar também na identificação de traços e recursos estilísticos do tradutor ou sua "impressão digital" deixada em um texto traduzido, inserções e omissões textuais que não se encontram no texto original. Pode, ainda, auxiliar a identificar quais foram as soluções adotadas por um tradutor em face de um "termo intraduzível", ou seja, algum elemento cultural que não existe na língua alvo.

Os itens acima mencionados podem ser compreendidos de uma maneira mais abrangente com o auxílio de um corpus paralelo (ver seção 2.1), uma grande quantidade de textos digitalizados em mais de uma língua, tornando-se ainda mais úteis quando alinhados a nível de sentença ou palavra e de modo pesquisável (OLOHAN, 2004, p. 24). Os textos que

[2] http://www.periodicos.capes.gov.br/

constituem um determinado corpus são de domínios específicos e podem compreender alguns tipos de texto tais como literários, acadêmicos, técnicos, entre outros.

A análise automática de textos em mais de uma língua constitui-se em um desafio extra, pois em uma língua como o inglês um texto pode ser analisado de forma mais rápida e completa em virtude do largo suporte disponível - ferramentas computacionais de análise e estudos conduzidos por pesquisadores, entre outros. Contudo, para uma análise de um texto em português o suporte e a disponibilidade de ferramentas e corpora não é tão rico como no caso da língua inglesa.

1.5 OBJETIVOS

Esta pesquisa visa realizar um mapeamento da área de mineração de textos, suas técnicas, soluções disponíveis e possíveis aplicações em um corpus paralelo bilíngue, para fornecer suporte ao estudo e investigação acadêmica de fenômenos tradutórios no campo disciplinar de Estudos da Tradução. Um estudo de caso de um sistema e corpus (ver capítulo 3) desenvolvido pelo próprio autor[3] durante seu mestrado será apresentado como um ponto de partida para a exemplificação e teste de aplicabilidade dos objetivos aqui propostos.

A revisão de literatura e o estudo de caso visam mostrar a possibilidade de aplicação da mineração de textos em conjunto com um corpus paralelo, na pesquisa de fenômenos tradutórios. O objeto utilizado no estudo de caso possui o nome de COPA-TRAD (Corpus Paralelo de Tradução) e foi desenvolvida durante o mestrado no Programa de Pós-Graduação em Letras e Literatura da Universidade Federal de Santa Catarina durante o período de 2011 a 2013. Esta solução permite que o usuário investigue as práticas de tradutores profissionais através da identificação de padrões tradutórios relacionados a um determinado elemento ou padrão linguístico.

Na versão atual do COPA-TRAD, técnicas de mineração de textos relacionadas ao processamento inicial dos textos inseridos foram utilizadas para a extração de *tokens*, alinhamento de sentenças, elaboração de estatísticas, indexação e recuperação de informações. Outras técnicas podem ser aplicadas para a criação de um sistema mais inteligente na extração de informações úteis para o campo da tradução.

---

[3] Disponível em: http://www.copa-trad.ufsc.br

Com isso, espera-se contribuir para uma área que movimenta um mercado de bilhões de dólares[4] e oferecendo ferramentas que auxiliem de forma mais profunda no entendimento dos fenômenos tradutórios e como tradutores profissionais lidaram com certos problemas tradutórios.

1.6 ORGANIZAÇÃO DA PESQUISA

Neste capítulo foi contextualizado a necessidade e a aplicação da mineração de textos em algumas áreas. Em seguida alguns desafios, problemas e os objetivos foram explicitados.

No capítulo 2 será realizada a revisão bibliográfica que servirá de arcabouço teórico para a pesquisa aqui proposta, além de definir e esclarecer melhor os termos e técnicas envolvidas na pesquisa de corpus para a tradução e a mineração de textos.

Em seguida, no capítulo 3, que diz respeito ao método, será apresentado um estudo de caso mostrando a aplicação prática dos objetivos aqui propostos e na última parte do capítulo serão considerados alguns protótipos em desenvolvimento.

Por último, na conclusão, serão recapitulados os principais pontos desta pesquisa, aplicações, contribuições para a área e possíveis assuntos para pesquisas futuras.

[4] Com base na estimativa para 2014. Disponível em http://www.gala-global.org/translation-global-business.

2 REVISÃO DE LITERATURA

Neste capítulo os dois objetos de estudo principais desta pesquisa são fundamentados de maneira teórica. Em primeiro lugar será detalhado e discutido o assunto de corpus e por sua vez o que vem a ser um corpus paralelo. Na segunda etapa um estudo detalhado dos principais tópicos da mineração de textos será conduzido com vistas a lançar um arcabouço teórico para fundamentar o estudo de caso e as sugestões de pesquisa.

2.1 ENTENDIMENTO DE UM CORPUS PARALELO

Em Processamento de Linguagem Natural (PLN) ou na mineração de textos um corpus ou *dataset* (ver seção 1.3) é um recurso primordial, pois serve de insumo para realizar análises textuais e também pode ser utilizado como um corpus de treinamento para melhorar a acuracidade de um algoritmo de mineração. Existem várias definições de corpus e meios de descrevê-los, isso é plenamente observável principalmente em PLN / Linguística Computacional onde existe uma área dedicada unicamente ao estudo de corpus que é chamada de Linguística de Corpus, conforme foi discutido na Seção 1.3.

No Brasil os corpora mais conhecidos fazem parte do projeto NILC (Núcleo Interinstitucional de Linguística Computacional) da Universidade de São Paulo em São Carlos[5]. No entanto, o foco deste trabalho está em um corpus paralelo e mais especificamente dentro da área Estudo da Tradução com Base em Corpus, uma área que prima mais pela pesquisa qualitativa do que quantitativa sendo que ambas as formas de pesquisas não são excludentes e sim complementares.

No campo de Estudos da Tradução (ET) temos para a língua portuguesa três principais corpora com alinhamento paralelo. O primeiro é o COMPARA de origem portuguesa. O segundo é o projeto COMET mantido pela Universidade de São Paulo. E o terceiro, e o mais novo de todos, é o COPA-TRAD que será abordado nesta pesquisa.

Segundo Simões (2011, p. 68) um "texto paralelo (ou bitexto) é um texto numa língua juntamente com a sua tradução numa outra língua", assim uma grande coleção de tais textos paralelos são chamados de corpora paralelos. É importante frisar que o corpus paralelo pode ter um texto original e várias traduções em línguas diferentes (corpus paralelo

[5] Maiores informações disponíveis em: http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm

multilíngue) ou um texto original e sua respectiva tradução em uma única língua (corpus paralelo bilíngue), neste último caso os textos alinhados são chamados de paralelos ou bitexto. Um ponto a observar e mencionado por Simões (2011, p. 68) é que apesar de a definição habitual de corpora paralelo não incluir o alinhamento de tais textos "é nossa convicção de que estes recursos são especialmente úteis quando alinhado a nível de frase", assim o autor passa a adotar o termo corpora paralelo para designar textos que são alinhados somente a nível de frase. Esta visão também é endossada por Mcenery e Xiao (2007, p.2-3), eles argumentam que um corpus é paralelo somente se contiver textos originais e suas respectivas traduções em paralelo. Mais adiante os mesmos autores informam que para um corpus paralelo ser útil, uma etapa importante é o alinhamento (ou pareamento) do texto original com as suas respectivas traduções, ou seja, um *link* ou relacionamento entre uma frase ou palavra de um texto original com sua respectiva frase ou palavra do texto fonte.

Textos em paralelo são úteis para um levantamento de conhecimentos linguísticos e são especialmente empregados na tradução automática, na construção de dicionários de tradução, entre outros (KUMANO e HIRAKAWA, 1994, p. 76). Shi e outros (2006, p. 443) acrescentam que um corpora paralelo bilíngue pode ser utilizado ainda na recuperação de informações e desambiguação lexical de sentido. Caseli e Nunes (2004, p. 1) dizem que os textos paralelos "são fontes ricas de conhecimento linguístico, isso porque a tradução de um texto para uma outra língua pode ser entendida como uma anotação detalhada do significado do texto original". Além disso, essas autoras destacam que um corpus paralelo alinhado possui muitas aplicações como para a tradução humana e automática e a recuperação de informações entre línguas diferentes.

Caseli e Nunes (2004, p. 1) definem dois tipos de alinhamentos de textos originais e suas respectivas traduções. O primeiro é o alinhamento sentencial, no qual o alinhamento é realizado a nível de sentenças, sendo que se o processo for automatizado o algoritmo deve identificar as sentenças no texto traduzido que correspondem a uma ou mais sentenças no texto original. Isso é importante destacar, pois dependendo das escolhas de um tradutor o alinhamento pode apresentar variações sendo que a relação de 1 para 1 fica prejudicada. Muitas vezes ocorrem omissões ou junções de frases no texto traduzido, assim um alinhador automático precisa identificar o mais próximo possível estas nuanças. Outro tipo de alinhamento informado pelas autoras é o alinhamento a nível lexical, o qual é realizado através de palavras.

Além disso, é possível acrescentar a esta lista, o alinhamento manual que se refere ao processo de alinhamento quando realizado por um humano, algo lento, mas que pode

garantir maior qualidade do produto final. Por último, temos o alinhamento assistido/supervisionado ou semi-automático quando o alinhamento é feito de forma automática e posteriormente conferido por um humano.

A exemplo do corpus monolíngue, os corpora paralelos passam por uma rigorosa seleção de textos de um domínio específico. Geralmente cada corpus é composto por textos de apenas um domínio como jornalísticos, redes sociais, revistas, textos literários, textos acadêmicos entre outros.

2.2 MOTORES DE BUSCA

Esta seção visa discutir de maneira resumida o assunto de motores de busca, pois tais mecanismos podem ser utilizados em conjunto com um corpus, especialmente no projeto aqui proposto. A ideia é utilizar um motor de busca para fazer o papel da técnica de recuperação de informações, parte do sistema de mineração de textos.

Os motores de busca (ou *search engines*), conforme se pode deduzir pelo próprio nome, são mecanismos para realizar buscas. Para fazer uma analogia, podemos comparar com serviços amplamente conhecidos na Internet como o Google, Yahoo ou o Bing da Microsoft, estes são caracterizados como motores de busca, embora de grande porte com um alto nível de complexidade e sistemas de computação distribuída (BRIN e PAGE, 1998, p. 107-108).

Um dos motores de busca disponível é o Sphinx Search[6] ou simplesmente Sphinx que é uma suíte de ferramentas para indexar, consultar e disponibilizar informações de uma fonte de dados para o usuário. Conforme observado por Curioso e outros (2010, p. 369-370), o Sphinx provê uma busca rápida, eficiente e relevante, pois um dos segredos desta agilidade é que o Sphinx cria e mantém um índice próprio desacoplado do Banco de Dados sem causar *overload* neste último.

O Sphinx é um sistema que pode indexar uma massa de dados estruturados ou semi-estruturados para realizar pesquisas *Full-Text* em um índice invertido. Conforme observado por Santos e Nunes (2011, p.2) um índice invertido "é constituído por um vocabulário, conjunto distinto de palavras-chave, onde [sic] cada palavra guarda um apontador para o início de uma lista invertida que armazena um conjunto de referências para as tuplas da base de dados". Uma diferença no índice do Sphinx é que o vocabulário de

[6] Site oficial do Sphinx Search: http://sphinxsearch.com/

palavras-chave, ou seja, cada palavra-chave é traduzida para sua forma em 32bit baseada na função polinomial CRC32[7] isto mantém o índice enxuto e pequeno.

As pesquisas realizadas pelo Sphinx podem ser conduzidas com base em alguns algoritmos disponíveis, sendo que o padrão é o SPH_RANK_PROXIMITY_BM25[8]. Todas as informações que devem fazer parte do índice são definidas pelo próprio pesquisador, em um arquivo de configuração. O Sphinx Search é *Open Source* e foi criado por Andrew Aksyonoff sendo que foi projetado para trabalhar e ser integrado de maneira descomplicada com o MySQL, no entanto suporta também outros bancos de dados assim como, documentos em XML.

Os autores Curioso e outros (2010, p. 369) fornecem em um capítulo dedicado as pesquisas *Full-Text* em seu livro, um detalhamento do motor de busca Sphinx. Com base nestes autores, o Quadro 1, lista e detalha as principais ferramentas disponíveis no Sphinx Search.

Quadro 1 – Ferramentas do Sphinx (CURIOSO et al., 2010, p. 369-370) .

Ferramenta	Descrição
Indexer	O programa responsável por criar os índices a partir de fontes de dados como, por exemplo, o MySQL.
Search	Um programa em linha de comando que procura um termo diretamente no índice. Geralmente este comando é utilizado para realizar testes no índice.
Searchd	Este é um *deamon* (servidor) que fornece para os clientes a funcionalidade de busca, gerenciado as requisições, buscas no índice, e o retorno dos resultados da busca. O searchd também possibilita a criação de índices distribuídos.
SphinxQL	Uma linguagem SQL para buscas.
Sphinxapi	Uma biblioteca de API clientes para serem implementadas em várias linguagens como PHP, Perl, Python ou Ruby.
Spelldump	Uma ferramenta linha de comando para a extração de itens do dicionário do ispell ou MySpell (Open Office) para a customização do índice.
Indextool	Uma ferramenta da linha de comando para checar a consistência dos índices.

[7] Para mais informações sobre o CRC32 visite: http://www.accuhash.com/what-is-crc32.html

[8] Para maiores informações visite: http://sphinxsearch.com/docs/2.0.5/boolean-syntax.html

Existem outros motores de busca. Um que merece ser destacado é o Lucene[9] mantido pelo projeto Apache. O Lucene apresenta algumas desvantagens em relação ao Sphinx, como por exemplo, o índice tende a crescer em tamanho muito rápido já que as palavras-chave são armazenas sem serem traduzidas para um *hash* como o CRC32. O Lucene não será abordado neste projeto.

2.3 TÓPICOS EM MINERAÇÃO DE TEXTOS

A Internet, em virtude da grande quantidade de dados, é considerada por muitos pesquisadores como uma grande base de dados. No entanto, os dados contidos neste "grande repositório" são de natureza heterogênea ou seja, este grande volume de informações estão representados em diferentes formatos como, por exemplo, HTML, XML, PDF, Bancos de Dados[10], Tabelas entre outros. A natureza de tais dados está classificada em três grandes categorias, a saber:

- Dados não-estruturados – textos livre, TXT, PDF.
- Dados semi-estruturados – HTML, XML.
- Dados estruturados – Tabelas em um Banco de Dados, RDF.

Na mineração de textos o foco está em dados em sua grande parte de natureza não-estruturados e também semi-estruturados. Dados não-estruturados, conforme observado, são textos corridos escritos em linguagem natural sem nenhum tipo de marcação (*tags*). Textos semi-estruturados por sua vez podem estar em HTML e XML, conforme Bhatia e outros (2011, p. 443) observa, "boa parte das informações na web estão em formato semi-estruturado por causa da estrutura de um documento em HTML, links e redundâncias". De acordo com Hotho e outros (2001, p. 5) a mineração de textos ou Descoberta de conhecimento de bases de dados textuais (KDT – *Knowledge discovery from text*) lida com a análise automática de textos e utiliza técnicas tais como Recuperação de Informações (IR – *Information Retrival*), Extração de Informações (IE - *Information Extraction*), Processamento

[9] Site oficial do Lucene: http://lucene.apache.org/core/

[10] Com relação a Bancos de Dados a linha de pesquisa em Dados na Web preocupa-se entre outros assuntos em como buscar, extrair, analisar e disponibilizar informações de banco de dados na Internet que muitas vezes podem ser acessados somente através de web forms, o que constitui um desafio.

de Linguagem Natural (PLN), técnicas e métodos de KDD (*Knowlege Discovery in database*), mineração de dados, Aprendizagem de Máquina (*Machine Learning* – ML) e estatística. Assim a mineração de textos "consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente para objetivos específicos" (ARANHA e PASSOS, 2006, p. 1).

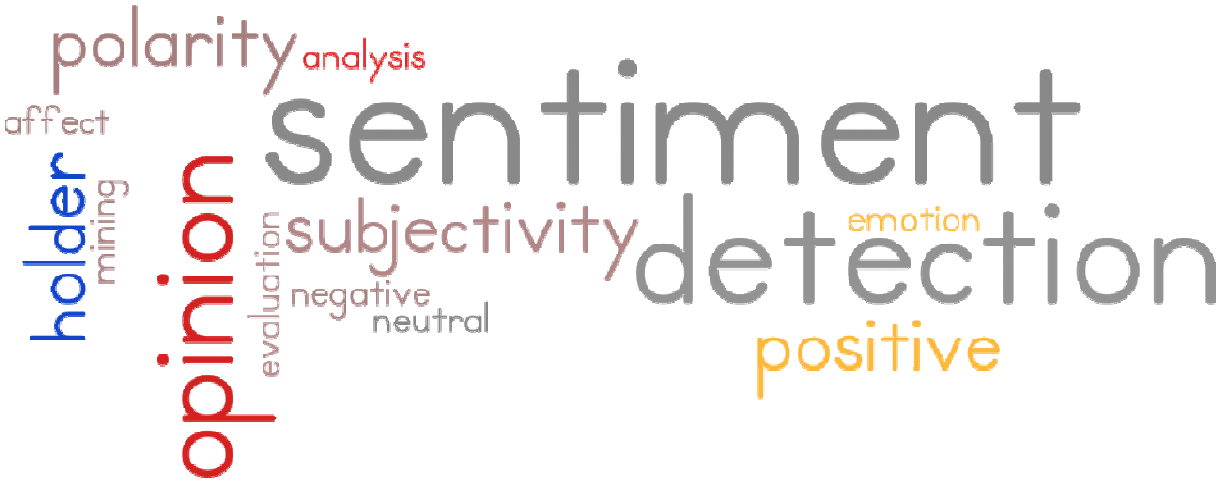
Scholtes (2009, p. 3), acrescenta que o estudo da mineração de textos "envolve o desenvolvimento de vários cálculos matemáticos e estatísticos, técnicas de linguística e reconhecimento de padrões" estas áreas embora pareçam distintas veem a contribuir para uma "análise automática de informações não-estruturadas assim como possibilitar a extração de dados altamente relevantes e de qualidade", isto possibilita que uma análise textual não fique limitada apenas a procura de palavras chaves.

A mineração de textos não é um bloco monolítico que oferece por si só um único jeito para processar, analisar textos e gerar conhecimento. Neste sentido, é interessante considerar a mineração de textos como um termo guarda-chuva que abriga uma variedade de tecnologias e caminhos diferentes para chegar a um determinado resultado (SAIKRISHNA, 2012, p. 225). Abaixo da mineração de textos existem várias subáreas que preocupam-se em realizar análises em uma variedade de domínios específicos. Com base na organização da área da mineração de textos e pesquisa realizada pelo autor, as seguintes subáreas são sugeridas:

- Análise de Opiniões (*Sentiment Analysis*);
- Análise de E-mails (*E-mail Analytics*);
- Análise de Uso/Comportamento (*Log Analytics*);
- Detecção de Fraudes (*Fraud Detection*);
- Análise de Redes Sociais (*Social Network Analysis*);
- Sistemas de Recomendação (*Recommendation Systems*);
- Resumo de Textos (*Text Summarization*);
- Categorização de Textos (*Text Categorization*);
- Extração de Padrões (*Pattern Extraction*);
- Detecção de Tendências (*Trend Detection*);
- Extração de Informações (*Information Extraction*);
- Recuperação de Informações (*Information Retrieval*).

Entre as subáreas mencionadas acima existem outras e dentro de cada subárea pode haver tópicos específicos. A título de exemplo em Análise de Opiniões vários assuntos podem ser pesquisados, conforme observado na Figura 1.

Figura 1 – Tópicos em Análise de Opiniões. Fonte: http://sentilab.sabanciuniv.edu



Hotho e outros (2001, p. 5) elencam as técnicas mais comuns utilizadas pela mineração de textos e que podem auxiliar a compreender melhor o que vem a ser a mineração de textos de maneira prática.

A Recuperação de Informações lida com o processo de encontrar documentos que possuam “respostas para perguntas e não somente respostas por si só”, deste modo, a área de recuperação de informações lida com o processamento de informações e a recuperação de dados. Com relação ao processamento de linguagem natural os autores Hotho e outros (2001, p. 5) dizem que o principal objetivo é alcançar um melhor entendimento da linguagem natural para ser utilizada por computadores. Nesta etapa uma análise linguística pode ser utilizada para o processamento do texto. Em seguida temos a área de extração de informações cujo principal objetivo é a extração de informações específicas de uma massa de dados textual.

Um ponto primordial mencionado pelos autores é que para “minerar uma grande quantidade de dados é necessário primeiro realizar um pré-processamento de toda a massa de dados textual para em seguida armazenar em alguma estrutura de dados” (HOTHO et. al, 2001, p. 6). Os autores argumentam que este é o modo mais apropriado para utilização e processamento futuro, ou seja, armazenar o texto de forma estruturada é melhor do que utilizar arquivos de textos simples. A parte de codificação dos textos é algo a ser mencionado,

pois para evitar problemas no reconhecimento de caracteres de uma determinada língua é necessário que a codificação seja a mesma em todos os textos.

2.4 PROCESSAMENTO LEXICAL

Conforme mencionado no parágrafo anterior, o mais indicado é que o texto esteja armazenado de forma estruturada como, por exemplo, em um banco de dados. Para isso é necessário realizar um pré-processamento do texto para adequar a certa estrutura definida pelo escopo e as necessidades do projeto que está sendo conduzido.

A primeira ação geralmente empregada na etapa de pré-processamento de um texto é a de *tokenization*, nesta etapa um texto é quebrado em unidades textuais com significados, estas unidades podem ser apenas uma palavra, um conjunto de palavras ou até uma frase, estas unidades textuais são denominadas de *tokens*.

Na etapa de pré-processamento textual Hotho e outros (2001, p. 7), destacam três importantes etapas: filtragem, lematização e *stemming*. Com relação à filtragem, um exemplo prático encontra-se na Seção 3.1.3, a qual apresenta o processo de identificação e remoção de *stopwords*, palavras que ocorrem com muita frequência ou rara frequência ao longo de um texto são removidas, pois elas não diferenciam bem entre documentos.

A lematização e *stemming* são dois processos que podem ser confundidos, pois parecem ser algo muito parecido. A lematização é o processo de passar uma palavra para o masculino e singular. Hotho e outros (2001) destacam ainda que "formas verbais são passados para sua forma infinitiva e substantivos para o singular". Existem vários algoritmos que facilitam o processo de lematização principalmente para o Inglês, mas também temos para o português, geralmente adaptações do algoritmo de língua inglesa. Para Hotho e outros (2001), *Stemming* é o processo que reduz uma palavra a sua forma mais básica e primitiva, sufixos, afixos e gerúndios são removidos. Após este processo "cada palavra é representada por sua *stem*".

Os pesquisadores De Lucca e Nunes (2002, p. 14) explicam de maneira clara a diferença entre lematização e *stemming*:

> Lematização difere fundamentalmente de stemming. Enquanto lematização existe puramente no contexto lexicográfico, stemming não. Lematização é, pois, a representação da palavra através de seu masculino singular, adjetivos e substantivos e infinitivos (verbos), apenas no contexto da lexicografia. Stemming é a retirada de

> sufixos do radical, enquanto stem é o radical. Assim, as estruturas são distintas, embora eventualmente possam ser graficamente semelhantes.

O algoritmo mais utilizado para transformar palavras em *stem* é o de Porter (1980) sendo que este mesmo algoritmo já foi adaptado para ser utilizado na língua portuguesa como é o caso do PTStemmer[11].

A indexação é a segunda etapa após o pré-processamento e oferece suporte para definir que somente palavras selecionadas são utilizadas para descrever certo documento. Silva (2002, p. 30) argumenta que a utilização de índices são importantes na "análise de informações em textos, os índices são peças importantes, pois eles são uma forma de validar o desempenho e a precisão da recuperação da informação".

Uma forma para construir índices pode ser através de algoritmos próprios e específicos desenvolvidos por pesquisadores de um projeto ou através de motores de busca. Conforme discutido na seção 2.2, um destes motores de busca que fornece um sistema para criar índices é o Sphinx Search, um sistema que pode indexar uma massa de dados heterogênea (Banco de Dados e XML) para realizar pesquisas *Full-Text* em seu índice com base em alguns algoritmos disponíveis pelo próprio sistema, todas as informações que devem fazer parte do índice são definidas pelo próprio pesquisador em um arquivo de configuração.

2.5 PROCESSAMENTO LINGUÍSTICO

Dependendo do projeto que está sendo desenvolvido um processamento linguístico mais apurado pode ser utilizado para "melhorar informações disponíveis sobre os termos" (HOTHO et. al, 2001, p. 9). Deste modo algumas técnicas podem ser úteis para conduzir um processamento linguístico. Hotho e outros (2001, p. 9) citam três etapas importantes:

- **Marcação das Partes do Discurso (*Part-of-speech tagging* - POS)** – define se uma palavra é um verbo, substantivo, adjetivo, ou seja, marcar/etiquetar (i.e. *tagging*) cada termo com sua respectiva classe gramatical. Muitas vezes esta marcação é realizada de forma manual ou com modelos probabilísticos. Manning e Schütze (2000, p. 345) destacam os modelos mais conhecidos, a saber:

[11] O PTStemmer pode ser encontrado em https://code.google.com/p/ptstemmer/

- Marcação baseada no Modelo de Markov, utilizado para identificação de padrões novos ou ocultos a partir de outros padrões conhecidos. Manning e Schütze (2000, p. 345), destacam que no modelo de Markov uma sequência de tags em um texto é verificada como uma Cadeia de Markov, onde *X* é uma sequência de variáveis aleatórias em um conjunto de tempo finito *t*, a cadeia é constituída de duas partes:

  - **Horizonte Limitado** (dependência da marcação anterior):

    $$P(X_{i+1} = t^j | X_1, \ldots, X_i) = P(X_{i+1} = t^j | X_i)$$

  - **Invariante de Tempo** (sem mudança no decorrer do tempo):

    $$P(X_{i+1} = t^j | X_i) = P(X_2 = t^j | X_1)$$

Manning e Schütze (2000, p. 345), explicam que uma marcação *s* de uma palavra depende da marcação anterior (horizonte limitado) e esta dependência não muda de acordo com o tempo (invariante de tempo). O autor fornece um exemplo: Caso um verbo finito tem a probabilidade 0.2 de ocorrer depois de um pronome no começo de uma frase, então esta probabilidade não vai mudar na medida em que novas frases são marcadas.

Manning e Schütze (2000) destacam outros modelos para realizar marcação automática de palavras como o algoritmo Viterbi, marcadores com base em trigramas, árvores de decisão, redes neurais, etc. Para o português temos marcadores específicos como o Aelius[12] e o Etiquetador Tree-Tagger[13].

- **Segmentação Textual:** com base no tipo de tarefa, são exigidos segmentos que possuem granularidades diferentes. Na segmentação textual são produzidos segmentos ou unidades significativas que evolvam uma ideia ou um conceito básico do texto (MAZIERO et. al, 2001, p. 7), tais unidades são conhecidas como sintagmas nominais ou verbais. Um analisador de segmentação textual para o português é o DiZer (*Discourse Analyser for Brazilian Portuguese*)[14].

---

[12] Disponível em: http://sourceforge.net/projects/aelius/

[13] Disponível em: http://www2.lael.pucsp.br/corpora/etiquetagem/

[14] http://www2.lael.pucsp.br/corpora/etiquetagem/

- **Desambiguação Lexical de Sentido (*Word Sense Disambiguation* – WSD):** A desambiguação lexical de sentido é necessária, pois uma palavra ou frase pode ter vários sentidos diferentes em relação a certo contexto. Palavras isoladas dadas fora de contexto constituem-se um problema para serem interpretadas segundo Manning e Schütze (2000, p. 229). Cada significado de uma palavra pode ser armazenado para constituir um dicionário de apoio à desambiguação (HOTHO et. al, 2001, p. 9). A desambiguação lexical de sentido pode ser utilizada como um fator chave para o sucesso de um bom tradutor automático, pois "um dos principais problemas é a ambiguidade lexical que ocorre quando da multiplicidade de opções, durante a seleção de uma palavram na língua alvo (LA), para traduzir uma palavra da língua-fonte (LF)" (SPECIA E NUNES, 2004, p. 1).

- **Parser:** Neste processo uma frase ou texto é analisada para produzir anotações sintáticas para "encontrar a relação de uma palavra em uma frase com todas as outras e a sua função na frase (i.e., sujeito, objeto, etc.)" (HOTHO et. al, 2001, p. 9).

2.6 OBSERVAÇÕES FINAIS

Neste capítulo foram abordados os conceitos e o quadro teórico que fundamentam o estudo de caso, apresentado a seguir. Os três principais assuntos foram descritos: Corpus Paralelo, Motores de Busca e Mineração de Textos. Outros dois fatores chave cobertos neste capítulo foram o de pré-processamento textual e o processamento linguístico.

No próximo capítulo será abordado o estudo de caso que visa a mostrar de maneira prática os itens abordados neste capítulo para satisfazer os objetivos gerais deste projeto.

3 ESTUDO DE CASO

O foco principal deste capítulo está no estudo de caso do COPA-TRAD para mostrar como esta ferramenta funciona e como as técnicas de mineração de textos foram empregadas. Conforme já discutido, o COPA-TRAD é um corpus paralelo que atualmente possui mais de dois milhões de palavras bilíngues e que foi desenvolvido durante o mestrado deste autor na Universidade Federal de Santa Catarina. Este corpus pode servir de massa de dados para a condução de projetos mais abrangentes que envolvem a mineração de textos para extrair informações úteis que auxiliem pesquisadores principalmente no campo de tradução, linguística computacional e a linguística *per se*. A aplicação de técnicas de mineração de textos em um corpus paralelo pode auxiliar na descoberta de informações e respostas para explicar certos fenômenos que ocorrem em um texto traduzido. Deste modo a mineração de textos pode auxiliar na pesquisa acadêmica de tradução. No final do capítulo será apresentado alguns protótipos que estão sob implementação e visam utilizar de maneira mais abrangente a aplicação da mineração de textos em um corpus paralelo e alinhado de tradução.

3.1 APRESENTAÇÃO DO COPA-TRAD

O COPA-TRAD é o primeiro corpus que compreende textos originais e traduzidos de gêneros como Literatura Infantil e Fantasia (um mercado que movimenta milhões de dólares) entre outros gêneros que serão detalhados mais a frente. Esta ferramenta computacional foi desenvolvida pelo autor com base nos estudos de Fernandes (2004).

O principal objetivo do COPA-TRAD é atender a necessidade de pesquisadores de Estudos da Tradução para investigar fenômenos tradutórios como, por exemplo, a tradução de colocações, termos culturalmente marcados, omissões na tradução, uso de vocabulário mais rico ou mais pobre na tradução, entre outros pontos. Um segundo público que o COPA-TRAD pretende atingir são os tradutores profissionais que podem tirar vantagens de uma ferramenta de corpus para encontrar soluções para certos problemas tradutórios que ocorrem na hora de traduzir um texto para o português.

Atualmente o COPA-TRAD está constituído de 24 textos em inglês e suas respectivas traduções em português o que dá um total de 48 textos completos e alinhados. É necessário salientar que todos os textos são de livros. No sistema gerenciador do COPA-

TRAD tem ainda mais 8 textos aguardando moderação para integrar a coleção do corpus. A ideia é de que os textos sejam distribuídos em cada subcorpus sendo que atualmente a maioria dos textos concentra-se em apenas um subcorpus. O COPA-TRAD é constituído de 5 subcorpus distribuídos por domínios específicos listados na Figura 2.

Figura 2 – Os cinco subcorpus que constituem o COPA-TRAD.

Conforme podemos observar, o primeiro subcorpus é o COPA-LIJ é constituído de textos da literatura infantil, estes textos englobam gêneros tais como fantasia, fábulas, contos de fada, etc. O segundo subcorpus é o COPA-TEL que é constituído de textos de literatura em geral como poesia, biografias, ficção, romances, entre outros. A maioria dos textos deste corpus está em domínio público (i.e., sem copyright). O terceiro subcorpus é o COPA-MDT que tem como objetivo específico receber textos científicos em Estudos da Tradução. O quarto subcorpus é o COPA-RAC um subcorpus que possui somente os resumos encontrados em artigos científicos, o motivo disso é que estes resumos geralmente estão em formato bilíngue. Por último temos o COPA-TEJ que são textos acadêmicos de ordem jurídica, existem muitos destes textos em formato bilíngue por causa da tradução juramentada.

Na versão de produção o COPA-TRAD possui 5 ferramentas para acesso as informações que foram mineradas. A seguir, estão listadas estas ferramentas:

- COPA-CONC – Um concordanciador paralelo.
- MONO-CONC – Um concordanciador monolíngue.
- WORDLIST – Um painel para consulta de frequências de palavras-chave.

- COPA-BUILDER – Uma ferramenta para auxiliar o usuário a criar um corpus descartável[15].
- COPA-STATS – Um painel para acesso a informações de ordem estatística.

Na versão em desenvolvimento alguns protótipos estão sendo desenvolvidos como um módulo para auxiliar o usuário no alinhamento automático de textos, um módulo para um futuro corpus multimodal que abrigue coleções de sons, imagens e vídeos, entre outros.

O COPA-TRAD foi construído em três grandes etapas seguindo um modelo sugerido por Fernandes (2004, p. 74) e adaptado para este projeto. O primeiro estágio é o de Design do Corpus no qual foi feito a parte de análise, prototipagem e definições de como selecionar e preparar os textos, definir os domínios e gêneros. O segundo grande estágio foi o de Construção do Corpus na qual foi realizado a parte de implementação de todo o sistema e configuração das tecnologias envolvidas, mesmo esta parte tendo sido executada por uma única pessoa algumas metodologias ágeis foram empregadas, como por exemplo, entregar partes utilizáveis do sistema, etc. O último grande estágio é o chamado de Processamento do Corpus em que o sistema é colocado em funcionamento, assim como os módulos visuais para a consulta do usuário final ao corpus. Cada uma destas etapas é descrita nas próximas três subseções.

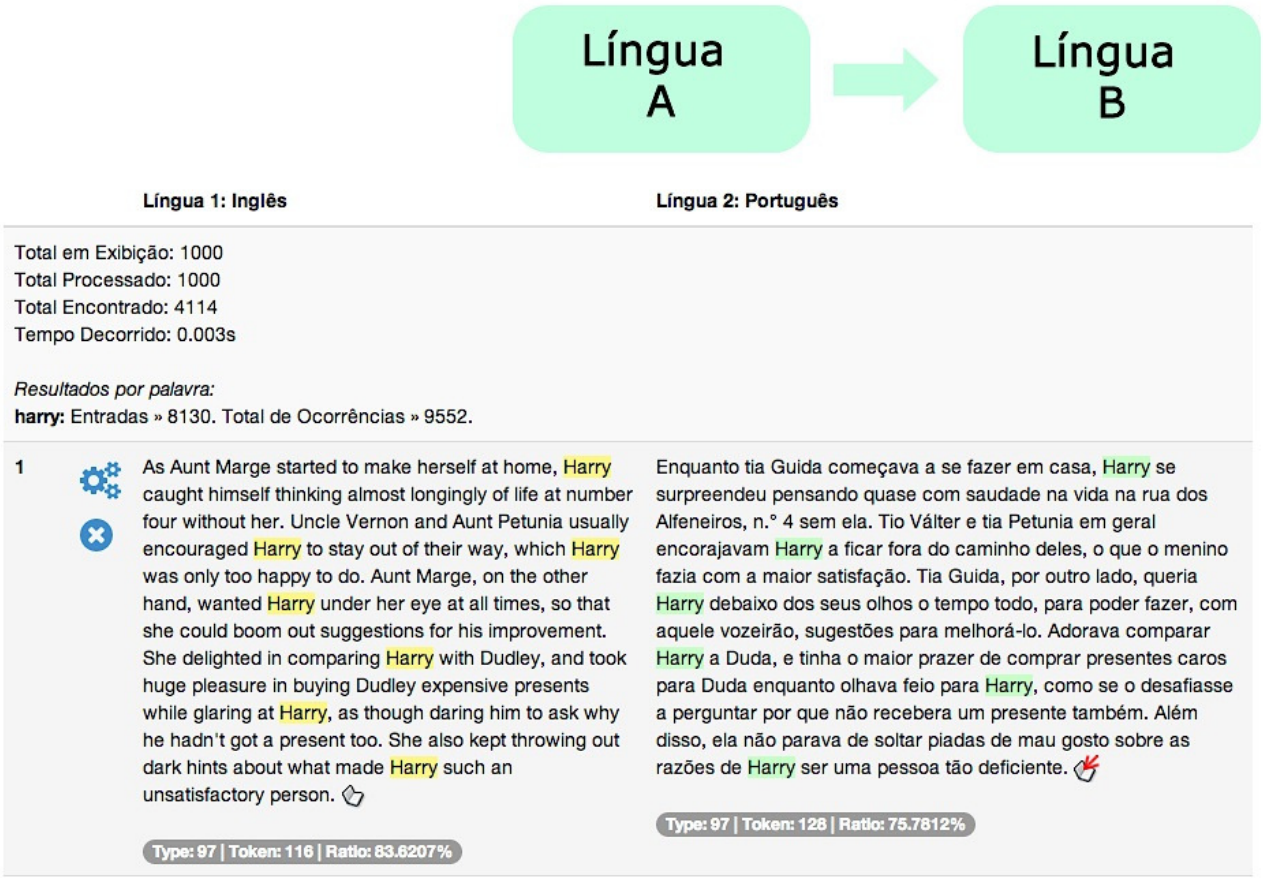
3.1.1 Design do Corpus

Os primeiros procedimentos adotados foram a parte de pesquisa, levantamento de requisitos, modelagem e elaboração do projeto, ou seja, as etapas usuais envolvidas no processo de análise de um software. Outro ponto abordado nesta etapa é sobre qual tipo de corpus seria construído para suprir as necessidades da pesquisa em Estudos da Tradução. Com base em Olohan (2004, p. 24), que indica o tipo de corpus mais adequado para a pesquisa em tradução, o tipo de corpus escolhido foi o paralelo, pois é constituído de textos originais e suas respectivas traduções alinhadas a nível sentencial, a Figura 3 ilustra um exemplo.

---

[15] Corpus descartável são coleções de textos que não fazem parte do corpus principal e geralmente são utilizados para realizar uma pesquisa ou investigação específica sendo que depois não são mais necessários.

Figura 3 – Exemplo de uma concordância bilíngue.



Uma das principais aplicações de um corpus paralelo é proposta por Baker (1995) a qual diz que um corpus paralelo permite estabelecer, de forma objetiva, como tradutores solucionam problemas tradutórios na prática da tradução.

Um problema levantado nesta parte de projeto foi com relação aos direitos autorias, pois textos com copyright não podem ser publicados na íntegra na Internet. Assim, com base em um estudo realizado pelo autor, optou-se por manter o COPA-TRAD fechado com os seguintes níveis de usuário:

- Administrador / Moderador – Os usuários que tem privilégios para gerenciar o corpus, executar os serviços de processamento e extração e alinhamento, além de moderar os textos submetidos por usuários.
- Pesquisador UFSC – Usuários de um pequeno grupo de pesquisa TraCor (tradução e corpora) que possuem privilégios de acesso a todos os módulos, assim como submeter um texto através de um painel especifico e visualizar todos os textos do corpus incluindo os textos com direitos autorais.

- Usuário Visitante – Demais usuários que possuem acesso às ferramentas de concordância e somente aos textos em domínio público ou que não possuem direitos autorais.

A seleção de textos foi realizada seguindo os critérios definidos no projeto e que compreendessem os domínios de cada subcorpus. Os textos foram selecionados, em sua maioria digitalizados com uma scanner OCR, passados para dois arquivos texto (um para cada língua) alinhados em nível de sentença de forma automática e revisados manualmente para garantir qualidade. Em seguida os textos foram salvos no formato UTF-8 sem BOM[16]. UTF-8 foi escolhido por fornecer um melhor suporte a caracteres de padrão universal no padrão Unicode. Outro ponto em favor do UTF-8 é que na implementação e configuração do motor de busca Sphinx, todos os códigos do padrão Unicode de todas as línguas modernas foram mapeados para o Sphinx reconhecer. Assim e conforme já testado em um protótipo o COPA-TRAD suporta textos inclusive no alfabeto cirílico, árabe, entre outros.

Com o módulo de submissão de textos os próprios usuários podem enviar textos através de um painel construído de maneira intuitiva para facilitar a inserção de textos no corpus.

3.1.2 Construção do Corpus

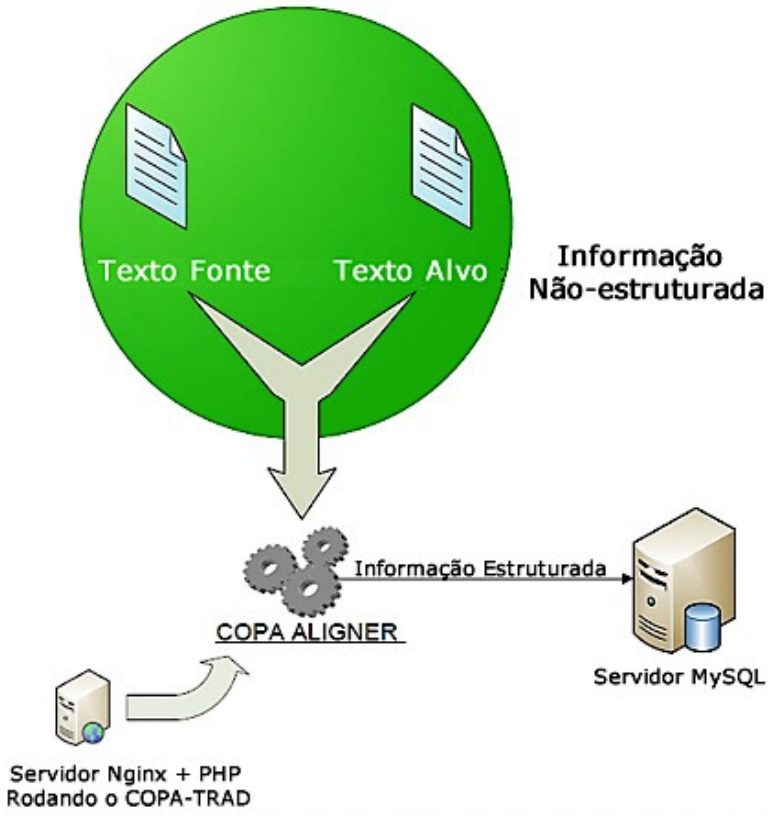
Esta parte focou principalmente na implementação do sistema no qual foi adotado um modelo incremental de desenvolvimento. O sistema foi desenvolvido 100% para estar na Internet por isso foi adotado a linguagem PHP no lado do servidor e no lado cliente a usual tríade HTML + JavaScript e CSS. No lado do servidor foi utilizado o framework em PHP CodeIgniter[17], personalizado pelo autor para fornecer suporte a *namespaces*. Além disso, todo o código de processamento do corpus foi construído como bibliotecas complementares deste framework. Atualmente na versão de desenvolvimento está sendo elaborado um framework próprio inspirado no Symfony[18] que seja enxuto para garantir maior segurança e desempenho

[16] BOM (Byte order mark) http://en.wikipedia.org/wiki/Byte_order_mark
[17] O CodeIgniter pode ser encontrado em: http://ellislab.com/codeigniter
[18] Informações sobre o Symfony http://symfony.com/

no COPA-TRAD. Nesta etapa o banco de dados também foi modelado e colocado em funcionamento (MySQL – InnoDB) assim como a configuração e teste do motor de busca e pequenas adaptações em um arquivo *htaccess* simples para adequar o servidor Web Nginx ao COPA-TRAD.

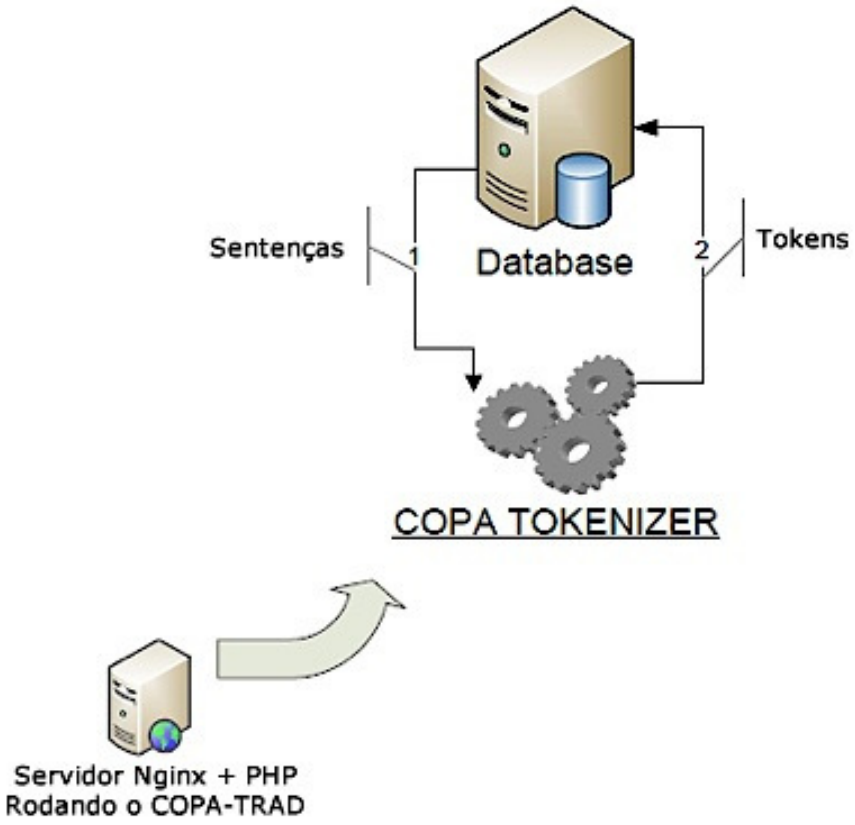
Conforme mencionado os módulos de processamento foram construídos em um formato reutilizável de bibliotecas. Estes módulos dão suporte a parte de processamento, alinhamento e extração lexical dos textos enviados. O primeiro módulo a ser mencionado é o COPA ALIGNER que é responsável por extrair as sentenças dos arquivos em formato txt e alinhar no banco de dados criando uma relação de um para um com a frase do texto fonte e texto alvo Figura 4. Como este módulo pode ser ativado por um usuário administrador através do painel web, o módulo possui um recurso de interoperabilidade para continuar em execução no servidor mesmo se o usuário perder a conexão com a Internet ou fechar o navegador. Nesta parte características previamente cadastradas são relacionadas a cada tupla, como nome do autor, idioma, variante do idioma, gênero, etc.

Figura 4 – Diagrama arquitetural do COPA ALIGNER.



O segundo módulo é o COPA-TOKENIZER que é responsável por extrair os *tokens* de todas as sentenças presentes no banco de dados (originais e traduções), dois algoritmos são utilizados dependendo do tipo de língua que a sentença está (Inglês ou Português). A tabela de sentenças é consultada no banco de dados e frase a frase os *tokens* são extraídos, computados e em seguida armazenados novamente em uma tabela específica. Em caso do *token* já estar presente na tabela, o mesmo é incrementado em uma lista de frequência. Por último cada *token* é relacionado a tupla específica da tabela que contém a sentença processada, conforme pode ser observado na Figura 5.

Figura 5 – Diagrama arquitetural do COPA TOKENIZER.



Na versão de produção as palavras não foram normalizadas e lematizadas, sendo que a única normalização que ocorre é em palavras estrangeiras presentes em textos do idioma inglês. Através de expressões regulares estas palavras são limpas. Conforme pode ser observado no Quadro 2.

Quadro 2 – Caracteres que são normalizados em palavras de origem estrangeira no idioma inglês.

A	ŠŒŽšœžŸ¥µÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖØÙÚÛÜÝßàáâãäåæçèéêëìíîïðñòóôõöø ùúûüýÿ
B	SOZsozYYuAAAAAAACEEEEIIIIDNOOOOOOUUUUYsaaaaaaaceeeeiiiionoooooouu uuyy

Conforme observado, a parte de normalização, limpeza e lematização não foram conduzidas, pois o objetivo inicial do corpus estava focado principalmente em pesquisas de ordem qualitativas. A exceção ocorreu somente para textos na língua inglesa que continham palavras provenientes de outros idiomas as quais tiveram que ser normalizadas. O resultado disso deve-se ao fato de que as palavras estrangeiras, em relação ao idioma inglês, não estavam de acordo com as regras gramaticais do inglês. Isto constituía um problema dentro do corpus e na geração de estatística dos *types* e *tokens,* pois uma palavra com acentuação gráfica poderia ser contada como duas.

3.1.3 Processamento do Corpus

Esta última etapa compreende basicamente a colocação em funcionamento de todo o sistema desenvolvido. Primariamente um texto é enviado por um usuário cadastrado que por sua vez é moderado por outro usuário administrador, com o texto aprovado uma mensagem de aprovação é enviada para a pessoa que submeteu o texto. Em um segundo momento o administrador é redirecionado para um painel em que é possível executar os serviços de processamento textual (i.e., COPA-ALIGNER e COPA-TOKENIZER).

Nesta etapa, conforme observado, uma pré-limpeza de todo o texto já é realizada, pois sinais diacríticos, tabulações, entre outros elementos são removidos. É interessante observar este ponto, pois no COPA-TRAD esta etapa é realizada duas vezes. Na primeira vez um grande bloco textual é quebrado em parágrafos e armazenado em banco de dados, sendo que na segunda etapa os parágrafos são quebrados em tokens, que por sua vez vão para o banco de dados também.

A filtragem e remoção de palavras comuns ou muito recorrentes como as *stopwords*, não foi realizada para o COPA-TRAD, pois até mesmo as palavravas mais comuns podem revelar certos comportamentos ou certos padrões recorrentes em um texto, como por exemplo, o traduzido. Para a versão em desenvolvimento uma tabela para armazenar uma lista de

palavras sem stopwords entre outros pontos será construída. Com todas as informações no banco de dados é possível rodar o serviço para geração de estatística.

O processo de indexação é automático, sendo executado por uma regra definida no *cron* o agendador de tarefas no Linux. O serviço de indexação do corpus é executado uma vez ao dia em um horário de baixo uso do sistema. Nesta etapa o Indexer consulta o banco de dados, de acordo com as regras definidas no arquivo de configuração e rotaciona os índices para adicionar as novas palavras-chave assim como outras informações específicas, como tipo de língua, identificador da tupla no banco de dados etc. Após todo este processo o texto está disponível para consulta e análise no sistema para todos os usuários.

3.2 APLICAÇÃO PRÁTICA DO COPA-TRAD

Conforme observado, o COPA-TRAD é uma solução constituída das seguintes partes: um sistema com uma interface intuitiva para que o usuário final possa realizar consultas e ter acesso as informações de maneira rápida e eficiente. Um sistema responsável por realizar todo o processamento textual e transformar informações não estruturada em estruturada; por último o corpus é armazenado em banco de dados MySQL que por sua vez é indexado por um motor de busca. A primeira ferramenta que foi apresentada é o COPA-CONC um concordanciador paralelo bilíngue que permite buscas a partir do texto fonte para o texto alvo quanto do texto alvo para o texto fonte. Na versão em produção o COPA-CONC possui um pequeno painel com alguns filtros para realizar buscas no corpus. Como o motor de busca é baseado no Sphinx, todos os caracteres coringas, buscas exatas, entre outros procedimentos de filtragem, estão disponíveis. O funcionamento é bastante parecido com um serviço de busca na Internet, com a diferença que é necessário definir quais serão as línguas que a serem mostradas no lado esquerdo e direito da tela. O mecanismo é bastante simples e intuitivo, dispensando o uso de manuais. Os caracteres coringas que podem ser utilizados no filtro de busca são:

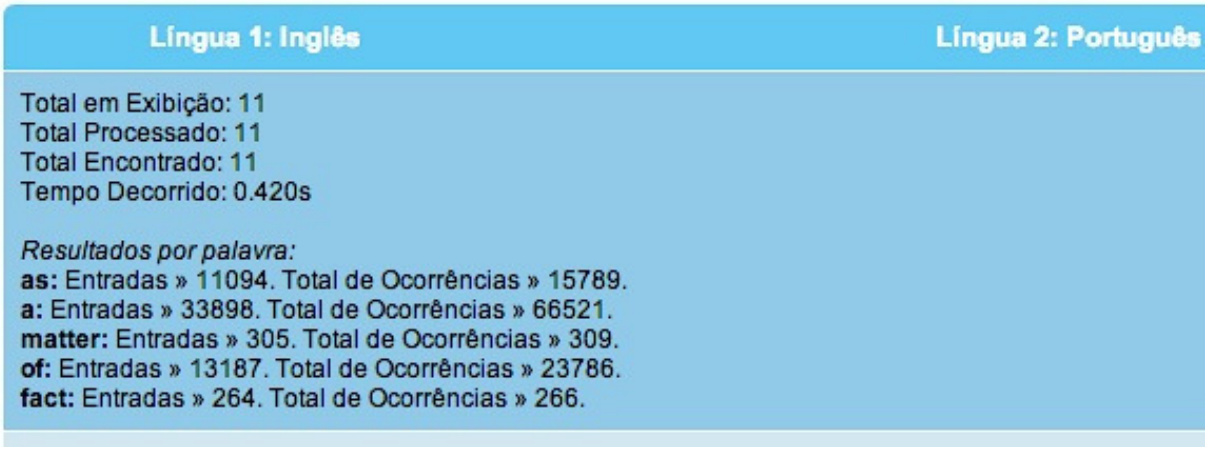
- Operadores Booleanos.
    - Operador AND como em “after & him”.
    - Operador OR como em “after | him”.
    - Operador NOT como em “after -him".
    - Operadores booleanos compostos como para estudar alguns verbos de ação como em: “(Harry & jumped) | (Harry & ran)”.

- Operador de frase exata: este operador é útil para procurar palavras exatas em um texto e pode ser largamente empregado na investigação da tradução de colocações como em “by the book” aqui as aspas duplas fazem parte do filtro.
- Operador de Início: este operador é utilizado para marcar palavras que começam em uma frase sendo que ocorrências subsequentes são ignoradas. O operador de iníncio de frase é o caractere de acento circunflexo como em “^take your time”.
- Operador de Final: este operador é utilizado para marcar palavras que terminam em uma frase, sendo que as ocorrências antes deste ponto são desconsideradas. O sinal é o cifrão, como por exemplo em “leave off$”.

O COPA-TRAD também possui um mecanismo inteligente para mostrar palavras que não foram traduzidas, com um método de *highlight*. Este é um recurso visual que pode mostrar se nomes próprios, de lugar, objetos ou palavras de origem latinas foram traduzidos ou não.

No COPA-CONC existem outros recursos para auxiliar o tradutor, como possiblidade de tirar entradas nas ocorrências exibidas na tela e que não são interessantes para o pesquisador. Possibilidade para imprimir todas as ocorrências, além de consultar meta-informações sobre uma determinada entrada. Outro ponto a ser mencionado é o recurso de consulta em máquinas de tradução famosas como o Google Translate e o Bing Translator, o sistema possui acesso a API do Bing da Microsoft e através de uma parceria com o Google Acadêmico o sistema tem permissão para consultar o Google Translate. Assim, ao clicar para traduzir um termo, uma tela é aberta onde é mostrado a versão original do texto selecionado, a versão da tradução oficial e logo abaixo a versão traduzida pelo Google Translate e o Bing Translator. Todas as traduções assim como o texto original são acompanhados de seus respectivos *type*, *tokens* e *ratio*. No topo da listagem é possível consultar também alguns dados estatísticos sobre a pesquisada realizada conforme pode ser observado na Figura 6:

Figura 6 – Dados estatísticos sobre a pesquisa realizada do usuário são mostrados em tempo real.



Na versão em produção do COPA-CONC é possível ainda, delimitar a busca por um subcorpus específico assim como por idioma.

Dando prosseguimento a próxima ferramenta é o MONO-CONC um concordanciador monolíngue que mostra os resultados pesquisados na tela no formato KWIC (ou Key Word in Context). Neste formato de exibição a palavra pesquisada fica ao centro e o texto anterior e posterior a esta palavra é mostrado em cada lado. O MONO-CONC pode ser utilizado especialmente para investigar as colocações e como uma palavra se relaciona com suas vizinhas ou seja estudar se um conjunto de palavras formam um certo agrupamento padronizado para representar um significado.

A próxima ferramenta é chamada de WORDLIST que é uma lista de frequência de palavras, a qual pode ser pesquisada por idioma e por ordem alfabética. Existe também uma nuvem de palavras que pode exibir de forma visual as palavras que ocorrem 600 ou mais vezes em todo o corpus, conforme exemplificado na Figura 7 a seguir:

Figura 7 – Palavras que ocorrem com mais frequência no idioma inglês.

Quando uma determinada palavra é clicada, uma pequena tela surge e mostra um exemplo de uma frase extraída do corpus na qual palavra escolhida está presente, assim como a sua frequência.

Em seguida temos a ferramenta CORPUS-BUILDER que conforme discutido auxilia o usuário a criar um corpus do tipo descartável. Existem dois campos textuais para o usuário colocar o seu texto em cada lado, sendo que estes campos são numerados para facilitar no trabalho de alinhamento. O CORPUS-BUILDER possui uma série de filtros para realizar a encontrar padrões no texto, tais filtros são bastante limitados, mas podem servir para uma pesquisa de pequeno porte. A limitação existe, pois o Sphinx não é utilizado, somente é utilizado um JavaScript e expressões regulares para realizar a busca.

Outra ferramenta disponível é o COPA-STATS que é a parte relacionada a mostrar estatísticas e gráficos relacionados a informações de ordem quantitativa. Com ela é possível criar gráficos e consultar informações de cada língua e também de textos específicos. Finalmente temos um módulo de submissão de textos para facilitar e acelerar o processo de inserção de textos no corpus. Neste módulo são listados todos os textos que um usuário enviou, assim como é possível consultar o resultado da aprovação ou não por parte dos moderadores sobre um texto que foi enviado. Existe ainda um formulário para inserir novos textos assim como para prover meta-informações de um texto e outro formulário responsável pela edição que pode ocorrer várias vezes enquanto um texto não é moderado e aprovado. Depois do texto aprovado a edição é bloqueada.

3.3 CONSIDERAÇÕES FINAIS

Foi apresentado neste capítulo um estudo de caso da ferramenta COPA-TRAD e como as técnicas de mineração de textos podem ser utilizadas para extrair informações úteis que auxiliam na investigação de fenômenos tradutórios. Parte do mecanismo interno e processamento do sistema também foram descritas. No próximo capítulo algumas conclusões são realizadas sobre o trabalho aqui proposto.

4 CONCLUSÃO

Este capítulo de fechamento objetiva recapitular brevemente as principais discussões deste trabalho assim como revisitar os objetivos propostos. Alguns problemas em aberto assim como algumas sugestões de pesquisa futura são apresentados. Por último é discutido as contribuições deste trabalho.

4.1 BREVE RECAPITULAÇÃO

O trabalho aqui proposto tem como objetivo principal realizar um mapeamento da área de mineração de textos e de algumas tecnologias utilizadas nesta área. A ideia foi mostrar que a aplicação de técnicas e soluções da mineração de textos pode ser utilizada para auxiliar na investigação de fenômenos tradutórios. Em seguida na parte de revisão um mapeamento da mineração de textos foi apresentado assim como a definição sobre o que é um corpus paralelo e também uma subseção dedicada ao que vem ser um motor de busca. No estudo de caso foi mostrado como todas estas tecnologias interagem e cooperam entre si para realizar uma investigação sobre um determinado assunto. O módulo web também foi descrito para mostrar como um usuário final pode interagir com as ferramentas disponíveis.

4.2 PROBLEMAS EM ABERTO E SUGESTÕES PARA PESQUISA FUTURA

Embora o assunto da área de processamento de linguagem natural estar de certa forma solidificado, a mineração de textos constitui-se um terreno novo (KAO e POTEET, 2005, p. 1). Neste sentido alguns desafios existem como internacionalização dos processos de mineração em textos e sobre a qualidade dos textos que são utilizados como massa de dados para pesquisa. Como exemplo e problemática é possível citar o Twitter em que nem sempre textos são apresentados, mas também links, caracteres especiais para formar desenhos, etc. Isto pode dificultar, por exemplo, um sistema de análise de opiniões na análise de alguns tweets. No entanto, como o volume de informações/tweets é muito grande, este problema muitas vezes não é relevante.

Voltando a atenção para o COPA-TRAD existem desafios também. Como por exemplo, na definição de como os dados podem ser apresentados para o usuário, assim como formas mais inteligentes de recursos da mineração de textos podem ser aplicados na

descoberta de conhecimentos novos. Outro ponto a ser mencionado é que ferramentas e técnicas de aferição precisam ser desenvolvidas para garantir a integridade e a confiabilidade de todo o sistema. Um ponto chave é com relação ao alinhamento automático de textos, pois como o corpus objetiva auxiliar na pesquisa qualitativa, a avaliação manual dos alinhamentos é primordial e isto pode tornar mais lento a disponibilização de novos textos no corpus. Neste sentido o autor propõe o uso de um algoritmo híbrido de alinhamento de textos (estatístico e lexical), adaptado para trabalhar com o par linguístico inglês e português. Em uma escala de qualidade do alinhamento variando de 0 a 1, foi possível obter uma qualidade de alinhamento entre 0.7 e 0.8 (dependendo do texto que está sendo processado). Para isso uma lista com cerca de 25 mil palavras alinhadas no par linguístico inglês-português já foi criada e testes iniciais em um protótipo estão sendo realizados com o algoritmo de alinhamento, até o momento os resultados têm-se mostrados satisfatórios.

4.3 CONTRIBUIÇÃO DESTE TRABALHO

Este trabalho possui uma contribuição de ordem teórico-prática, pois foram mostrados conceitos teóricos relativos a mineração de textos e como os mesmos podem ser aplicados em um projeto real. Outro ponto a ser mencionado diz respeito a proposta de aplicação de técnicas de mineração de textos em um corpus paralelo de tradução. Com estes recursos, um pesquisador pode investigar fenômenos tradutórios como, por exemplo, o uso e tradução de colocações, expressões idiomáticas, prosódia semântica, padrões linguísticos, nomes próprios e de lugar e tantos outros assuntos que constituem possíveis fenômenos passíveis de investigação. A ferramenta que foi construída está disponível para a comunidade acadêmica interessada a fim de que possa servir de apoio à investigação e pesquisa relacionada a corpus paralelo, entre outras aplicações.

REFERÊNCIAS

ALTMAN, R. B. *et al*. **Text mining for biology - the way forward: opinions from leading scientists**. Genome Biology, 9 (2), 2008, p. 3-7.

ARANHA, C.; Passos, E. A Tecnologia da Mineração de Textos. **RESI – Revista Eletrônica de Sistemas de Informação**, n. 2, p. 1 – 8, 2006.

BAKER, M. **Corpora in Translation Studies. An Overview and Suggestions for Future Research**. Target, 7(2), 1995, p. 223-243.

BHATIA, A. *et al*. Analysis of Pattern Recognition (text mining) with web Crawler. **International Transactions in Applied Sciences**, v. 03. N. 03. P.441-456, set. 2011.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. **Computer Networks and ISDN Systems**, v. 30, n. 1-7, p. 107–117, 1998.

CASELI, H.M; NUNES, M.G.V. Corpus paralelo e corpus paralelo alinhado: propriedades e aplicações. **Estudos Linguísticos**, v. 33, Taubaté, p. 1-6, 2004.

CHEN, H. **Knowledge Management Systems**: a text mining perspective. Tucson: University of Arizona, 2001.

CURIOSO, A. *et al*. **Expert PHP and MySQL**. Indianapolis: Wiley Publishing, 2010.

DE LUCCA, J. L.; NUNES, M. G. V. **Lematização versus Stemming**. São Carlos: NILC – ICMP-USP, 2002. 16 p. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional).

DEMYSTIFYING Big data: A Practical Guide To Transforming The Business of Government. Washington: TechAmerica Foundation, 2012. Disponível em: http://www-304.ibm.com/industries/publicsector/fileserve?contentid=239170
Acessado em 01/02/2014, às 20hs.

FERNANDES, L. **Brazilian Practices of Translating Names in Children's Fantasy Literature: a corpus-based study**. Universidade Federal de Santa Catarina, Florianópolis, 2004.

FUNG, G. P. C. *et al*. **The Predicting Power of Textual Information on Financial Markets**. IEEE Intelligent Informatics Bulletin, 5(1), 2005, p. 1-10.

GROSSMAN, R. L. **The Structure of Digital Computing From Mainframes to Big Data**. Illinois: Open Data Press LLC, 2012.

HARTLEY, T. **Technology and Translation**. The Routledge Companion to Translation Studies, 2009, p. 106-127.

HEARST, M. **Untangling text data mining**. In: ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 37, 1999. College Park. p. 3-10.

HEARST, M. **What Is Text Mining?**. 17 out. 2003. Disponível em: http://people.ischool.berkeley.edu/~hearst/text-mining.html. Acesso em 08 jun. 2014.

HOTHO, A. *et al*. **A Brief Survey on Text Mining**. In: LDV Forum, 2005, p. 19-62.

KAO, A; POTEET, S. **Text Mining and Natural Language Processing** – Introduction for the Special Issue. SIGKDD Explorations Newsletter, v. 7, n. 1, 2005, p. 1-2.

KRALLINGER, M. *et al*. **Text-mining approaches in molecular biology and biomedicine**. Drug Discovery Today, 10 (6), 2005, p. 439-445.

KENNY, D. **Corpora**. In: Routledge Encyclopedia of Translation Studies. London: Routledge, 2009. V. 2, p. 59-63.

KUMANO, A.; HIRAKAWA, H. **Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistic Information**. In: Proceedings of International Conference on Computational Linguistics, Kyoto, 1994, p. 76-81.

KWOK, L.; YU, B. **Spreading Social Media Messages on Facebook: An Analysis of Restaurant Business-to-Consumer Communications**. Cornell Hospitality Quarterly, 54 (1), 2013, p. 84-94.

MANNING, D. C.; SCHÜTZE, H. **Statistical Natural Language Processing**. Massachusetts: The MIT Press, 2000.

MANYIKA, J. Et. al. Big data: **The next frontier for innovation, competition, and productivity**. 01 mai. 2011. Disponível em: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Acesso em 08 jun. 2014.

MAZIERO, G. E. Et al. **Identificação automática de segmentos discursivos: o uso do parser PALAVRAS**. São Carlos: NILC-ICMP-USP, 2007. 23p. (Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional).

MCDONALD, D.; KELLY, U. **The Value and Benefits of Text Mining**. Disponível em: http://www.jisc.ac.uk/sites/default/files/value-text-mining.pdf. Acessado em 01/05/2014, às 13hs.

MCENERY, A. M.; XIAO, R. Z. **Parallel and comparable corpora: What are they up to?**. Incorporating Corpora: Translation and Linguist. Londres, 2007.

MOONEY, R. J.; BUNESCU, R. **Mining Knowledge from Text Using Information Extraction**. SIGKDD Explorations, 7 (1), 2005.

OLIVEIRA, L. P. **Linguística de Corpus: Teoria, Interfaces e Aplicações**. Matraga, 16 (24), 2009, p. 48-76.

OLOHAN, M. **Introducing Corpora in Translation Studies**. London and New York: Routledge, 2004.

OTHERO, G. A. **Linguística Computacional: uma breve introdução**. Letras de Hoje, 41 (2), 2006, p. 341-351.

PAVLYSHENKO, B. **Can Twitter predict royal baby's name?** 23 jul. 2013. Disponível em: http://bpavlyshenko.blogspot.com.br/2013/07/can-twitter-to-predict-royal-baby-name.html. Acesso em: 05 jan. 2014.

PORTER, M. F. **An Algorithm for Suffix Stripping**. v. 14, n. 3, p. 130-137, 1980.

SAIKRISHNA, V. Et al. String Matching and its Applications in Diversified Fields. **IJCSI International Journal of Computer Science Issues**, v. 9, n. 1, p. 219-226, 2012.

SANTOS, A.; NUNES, S. **Abordagens para a pesquisa por palavra-chave em bases de dados estruturadas**. In: INForum 2011. 2011, p. 1-6.

SCHNEIDER, M. O. **Processamento De Linguagem Natural (PLN)**. Pontifícia Universidade Católica De Campinas. 2001.

SCHOLTES, J. **Text Mining**: The next step in Search Technology. In: DESI-III Workshop, Barcelona, 2009, p. 1 – 22.

SHI, L. et al. **A DOM Tree Alignment Model for Mining Parallel Data from the Web**. Proceedings of Meeting of the Association for Computational Linguistics (ACL), p.489-496, 2004.

SICULAR, S. **The Era of Data**. 03 de jun. 2014. Disponível em: http://blogs.gartner.com/svetlana-sicular/the-era-of-data/ Acesso em: 08 jun. 2014.

SILVA, E. M. **Descoberta de Conhecimento com o uso de Text Mining**: Cruzando o Abismo de Moore. 2002. 174f. Dissertação (Mestrado em Informática) – Curso de Pós-graduação Stricto Sensu em Informática, Universidade Católica de Brasília, Brasília, 2002.

SILVA FILHO, L. A. **Mineração De Regras De Associação Utilizando KDD E KDT: Uma Aplicação Em Segurança Pública**. 2009. Dissertação (Mestrado em Informática) – Programa de Pós-Graduação em Ciência da Computação. Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará, Pará, 2009.

SPECIA, L.; NUNES, M. G. V. **Um Modelo para a Desambiguação Lexical de Sentido na Tradução Automática**. In: WTDIA - Workshop de Teses e Dissertações em Inteligência Artificial - XVII SBIA, São Luis, 2004.

TAN, A. **Text Mining: The state of the art and the challenges**. In: Proceedings of the Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD99, 3, 1999. Beijing.

SILVA, V. A.; ANDRADE, L. H. C. Etinobotânica Xucuru: espécies místicas. Biotemas, Florianópolis, v. 15, n. 1, p. 45-57, 2002.

WITTEN, I. H. Text mining. **Practical handbook of internet computing**. Boca Raton: Chapman & Hall/CRC Press, 2005.

ANEXOS

ANEXO A – Formulário de Registro de Patente do COPA-TRAD

INPI

PEDIDO DE REGISTRO DE PROGRAMA DE COMPUTADOR

protocolo

INPI INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL PROTOCOLO GERAL
16/05/2012 017120000491 15:20 RESC
0000271202187537

IDENTIFICAÇÃO DO PEDIDO (Para uso do INPI)

Número do Pedido | Protocolo, Data e Hora

DADOS DO AUTOR DO PROGRAMA

Nº de Autores: 2 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF*: 807.832.529-00
Nome: LINCOLN PAULO FERNANDES
Nome Abreviado, pseudônimo ou sinal convencional (se houver):
Data de Nascimento: | Nacionalidade: BRASILEIRA
Endereço: RODOVIA JOÃO PAULO, 710, T1, 301-B
Cidade: FLORIANÓPOLIS | UF: SC | País: BRASIL
CEP: 88.030-300 | Telefone: 4833048936 | FAX:
E-mail: lincoln.fernandes@ufsc.br

DADOS DO TITULAR DOS DIREITOS PATRIMONIAIS

Nº de Titulares: 1 — Se mais de um, preencha a "Continuação", com todos os dados solicitados neste Quadro. Date e assine.

CPF/CNPJ*: 83899526000182
Nome/Razão Social: UNIVERSIDADE FEDERAL DE SANTA CATARINA
Nome abreviado, pseudônimo ou sinal convencional (se houver): UFSC
Data de Nascimento: | Nacionalidade/Origem:
Endereço: CAMPUS UNIVERSITÁRIO, SN, CP 476, TRINDADE
Cidade: FLORIANÓPOLIS | UF: SC | País: BRASIL
CEP: 88.040-900 | Telefone: 4837219628 | FAX:
E-mail: dit@reitoria.ufsc.br

☐ SIM, este Titular é Pessoa Jurídica. Caso afirmativo, assinale a melhor classificação:

☐ Órgão Público ☐ Sociedade co Intuito não Econômico ☐ Microempresa ☐ Software House
☒ Instituição Pública de Ensino ou Pesquisa ☐ Instituição Privada de Ensino ou Pesquisa ☐ Outras

ENDEREÇO PARA CORRESPONDÊNCIA E CONTATO (Preencha apenas o necessário)

Toda correspondência será enviada para: ☐ O Procurador ou ☒ O Titular acima ou
☐ Escaninho nº | ☐ Representação INPI em: | ☐ O Endereço abaixo:

Nome:
Endereço:
Cidade: | UF: | País:
CEP: | Telefone: | FAX:
E-mail:

Modelo I (folha 1/2) E

ANEXO B – Publicação no diário oficial do pedido de patente

Regime de Guarda: Sigilo Até 23/04/2022
Procurador: PAULO AUGUSTO MALTA MOREIRA - CPF:66320844604

Processo: 13279-5 **080**
Título: ALFA
Titular: UNIVERSIDADE FEDERAL DE VIÇOSA - CPF/CNPJ:25944455000196
Criador: ANDRÉ FERNANDO DE OLIVEIRA
Linguagem: VBA, VISUAL BASIC
Campo de Aplicação: FQ-16
Tipo de Programa: FA-01
Data da Criação: 28/02/2010
Regime de Guarda: Sigilo Até 18/05/2022
Procurador: PAULO AUGUSTO MALTA MOREIRA - CPF:66320844604

Processo: 13280-4 **080**
Título: THOTAU/THOTAU - SISTEMA INTEGRADO DE GESTÃO EMPRESARIAL
Titular: ORION SISTEMAS LTDA - CPF/CNPJ:03005347000115
Criador: EDSON TEIXEIRA MARQUES , EDUARDO BARBOSA DE SOUZA
Linguagem: OBJECT PASCAL
Campo de Aplicação: AD-05, AD-08, AD-09, FN-05, FN-06
Tipo de Programa: AT-03
Data da Criação: 07/11/2008
Regime de Guarda: Sigilo Até 18/05/2022
Procurador: Não informado ou inexistente

Processo: 13281-6 **080**
Título: COPA-TRAD: CORPUS PARALELO DE TRADUÇÃO
Titular: UNIVERSIDADE FEDERAL DE SANTA CATARINA - CPF/CNPJ:83899526000182
Criador: CARLOS EDUARDO DA SILVA , LINCOLN PAULO FERNANDES
Linguagem: JAVASCRIPT, PHP
Campo de Aplicação: CO-03
Tipo de Programa: FA-01, GI-01, GI-08, UT-01
Data da Criação: 01/06/2011
Regime de Guarda: Sigilo Até 16/05/2022
Procurador: Não informado ou inexistente

Processo: 13282-1 **080**
Título: MATA ATLÂNTICA, O BIOMA ONDE EU MORO
Titular: UNIVERSIDADE FEDERAL DE SANTA CATARINA - CPF/CNPJ:83899526000182
Criador: ANA BEATRIZ BAHIA SPINOLA BITTENCOURT, CRISTINA VALÉRIA SANTOS, EMÍLIO TAKASE , MATHEUS BASSI BLANK GONÇALVES
Linguagem: FLASH
Campo de Aplicação: ED-01
Tipo de Programa: ET-01, ET-02
Data da Criação: 20/01/2012
Regime de Guarda: Sigilo Até 16/05/2022
Procurador: Não informado ou inexistente

Processo: 13283-3 **080**
Título: D-1 DISTRIBUTION ONE
Titular: ALCIRO MARCOS ORLAMÜNDER - CPF/CNPJ:68401361915
Criador: ALCIRO MARCOS ORLAMÜNDER
Linguagem: 4GL, JAVA, PROGRESS, VISUAL BASIC
Campo de Aplicação: AD-05, AD-08, AD-10, AD-11, FN-06
Tipo de Programa: AP-01, AP-02, AP-03, IA-01, IA-02
Data da Criação: 01/07/2001

Regime de Guarda: Sigilo Até 17/05/2022
Procurador: ALCIRO MARCOS ORLAMÜNDER - CPF:68401361915

Processo: 13329-1 **080**
Título: DOMINIO ESCRITA FISCAL VERSÃO 04
Titular: DOMÍNIO SISTEMAS LTDA. - CPF/CNPJ:02825945000178
Criador: ADRIANO DIAS, ADRIANO FRANCISCO, ALESSANDRA TEREZINHA DA SILVA, ALEXANDRE DE ALMEIDA, ALEXANDRE NIERO, ALEXANDRE ROBERTO LEMES MARTINS, ALINE CORREA RAMOS, ALISSON DE VILLA GERONIMO, ALISSON DOS SANTOS SILVA, ANDERSON FELISBERTO MANOEL, ANDERSON RICARDO DOS SANTOS RODRIGUES, ANDERSON SILVESTRI FERRO, ANTONIO JOSÉ VIEIRA JUNIOR, ANTONIO MARCOS DE OLIVEIRA, BRUNO BRISTOT LOLI, CAMILA MOTTA WOSNIESKI, CARLA EYNG, CESAR EDUARDO FRANCO ISE COLONETTI, CIRILO PINTER COLOMBO, CLEVERSON REINERT, DANGELO ROSSO ZANETTE, DANIEL DE MEDEIROS BOFF, DAVI GONÇALVES, DIEGO GOMES ANTONELI, DIEGO MACHADO MEDEIROS, DIEGO MARIANI DE MELO, DIEGO MARTINS DA ROCHA, EDGAR SOUZA DA CRUZ, EDIVALDO LUCIO, EVERSON NERI FRANCELINO, FAGNER LEANDRO DE SOUZA, FELIPE CORAL SASSO, FELIPE ORTMEYER HENRIQUE DA SILVA, FERNANDA D AGOSTIN, FERNANDO NAZARIO PIZZETTI, FLARIS BARRETO MARTINHAGO, GABRIEL GUADANHIM GENEROSO, GUILHERME FRANCISCO DE SOUZA, GUILHERME TEODORO DE OLIVEIRA, GUSTAVO GRIGGIO DE SOUZA, HEMERSON BEZ BIROLO, HENRIQUE COLOMBO GUINZAIN, HENRIQUE PIAZZA LUCIANO, HERLON HILBERT, HERON POTRIKUS CRESTANI, IURI SONEGO CARDOSO, JAISSON RODRIGUES DEMBOSKI, JEFERSON LUIZ BATISTI, JESSICA RONCONI DONDOSSOLA, JULIANA GUADANHIM GENEROSO, JULIANO MARQUES, LEONARDO BENEDET, LUANA GASPAR SOARES, LUCAS VITORINO GONÇALVES, MARCELO DEHON BATISTA DE PRA, MARCIO DAGOSTIM DE CASTRO, MARCONDES DE BORBA, MARIANA ANTONIO SARTORI, MARIANI COLONETTI, MARIANNA SANTOS SAGGIORATO, MARILIA TEIXEIRA PIRES, MARINA KURTZ SCHMIDT, MARITY EYNG NUERNBERG, MATHEUS MEDEIROS ANACLETO, MELISSA DA PAZ TEIXEIRA, MICHAEL CELSO BITENCOURT, PAULA CRISTINA VIEIRA RONSANI, PAULO HENRIQUE ELI, PAULO ROBERTO DABOIT MILANEZ, RAFAEL CECHINEL SILVESTRI, REGINALDO DAROLT, RENAN ROSSO DA SILVA, RICHARDSON PICININI CORREIA, ROBERTO MENDES GARCIA, ROBERTO VEFAGO CAROLLI, ROGERIO BRUM HERMANY, ROGERIO DAMACENO DE FARIAS, SAMUEL LODETTI GHELLERE, SIMONE PEREIRA DA CUNHA, SUELEN JUVENCIO DAMAZIO, TALINE FELTRIN DE SOUZA, TAMARA JOSEPHINO FERNANDES, TAMIRES JUSTI ROCHA, THALES MENDES MILANESE, THIAGO APOLINARIO BILLIERI, TIAGO BITENCORT MARQUES, TULIO DAMINELLI BORGES, VANESSA CRISTINA CARPES DA SILVA, VANESSA FELISBERTO BILESIMO, WAGNER JOSÉ DENONI FREITAS, WELLINGTON ZOMER NUNES
Linguagem: POWERBUILDER, SQL
Campo de Aplicação: IF-10
Tipo de Programa: AT-02
Data da Criação: 01/01/1999
Regime de Guarda: Sigilo Até 29/05/2022
Procurador: DMARK REGISTROS DE MARCAS E PATENTES LTDA - CPF:03389474000165

Processo: 13332-4 **080**
Título: DOMÍNIO ATENDIMENTO VERSÃO 02
Titular: DOMÍNIO SISTEMAS LTDA. - CPF/CNPJ:02825945000178
Criador: ADRIANO DIAS, ADRIANO FRANCISCO, ALESSANDRA TEREZINHA DA SILVA, ALEXANDRE DE ALMEIDA, ALEXANDRE NIERO, ALEXANDRE ROBERTO LEMES MARTINS, ALINE CORREA RAMOS, ALISSON DE VILLA GERONIMO, ALISSON DOS SANTOS SILVA, ANDERSON FELISBERTO MANOEL, ANDERSON RICARDO DOS SANTOS RODRIGUES, ANDERSON SILVESTRI FERRO, ANTONIO JOSÉ VIEIRA JUNIOR, ANTONIO MARCOS DE OLIVEIRA, BRUNO BRISTOT LOLI, CAMILA MOTTA WOSNIESKI, CARLA EYNG, CESAR EDUARDO FRANCO ISE COLONETTI, CIRILO PINTER COLOMBO, CLEVERSON REINERT, DANGELO ROSSO ZANETTE, DANIEL DE MEDEIROS BOFF, DAVI GONÇALVES, DIEGO GOMES ANTONELI, DIEGO MACHADO MEDEIROS, DIEGO MARIANI DE MELO, DIEGO MARTINS DA ROCHA, EDGAR SOUZA DA CRUZ, EDIVALDO LUCIO, EVERSON NERI FRANCELINO, FAGNER LEANDRO DE SOUZA, FELIPE CORAL SASSO, FELIPE ORTMEYER HENRIQUE DA SILVA, FERNANDA D AGOSTIN, FERNANDO NAZARIO PIZZETTI, FLARIS BARRETO MARTINHAGO, GABRIEL GUADANHIM GENEROSO, GUILHERME FRANCISCO DE SOUZA, GUILHERME TEODORO DE OLIVEIRA, GUSTAVO GRIGGIO DE SOUZA, HEMERSON BEZ BIROLO, HENRIQUE COLOMBO GUINZAIN, HENRIQUE PIAZZA LUCIANO, HERLON HILBERT, HERON POTRIKUS CRESTANI, IURI SONEGO CARDOSO, JAISSON RODRIGUES DEMBOSKI, JEFERSON LUIZ BATISTI, JESSICA RONCONI DONDOSSOLA, JULIANA GUADANHIM GENEROSO, JULIANO MARQUES, LEONARDO BENEDET, LUANA GASPAR SOARES, LUCAS VITORINO GONÇALVES, MARCELO DEHON BATISTA DE PRA, MARCIO DAGOSTIM DE CASTRO, MARCONDES DE BORBA, MARIANA ANTONIO SARTORI, MARIANI COLONETTI, MARIANNA SANTOS SAGGIORATO, MARILIA TEIXEIRA PIRES, MARINA KURTZ SCHMIDT, MARITY EYNG NUERNBERG, MATHEUS MEDEIROS ANACLETO, MELISSA DA PAZ TEIXEIRA, MICHAEL CELSO BITENCOURT, PAULA CRISTINA VIEIRA RONSANI, PAULO HENRIQUE ELI, PAULO ROBERTO DABOIT MILANEZ, RAFAEL CECHINEL SILVESTRI, REGINALDO DAROLT, RENAN ROSSO DA SILVA, RICHARDSON PICININI CORREIA, ROBERTO MENDES GARCIA, ROBERTO VEFAGO CAROLLI, ROGERIO BRUM HERMANY, ROGERIO DAMACENO DE FARIAS, SAMUEL LODETTI GHELLERE, SIMONE PEREIRA DA CUNHA, SUELEN JUVENCIO DAMAZIO, TALINE FELTRIN DE SOUZA, TAMARA JOSEPHINO FERNANDES, TAMIRES JUSTI ROCHA, THALES MENDES MILANESE, THIAGO APOLINARIO BILLIERI, TIAGO BITENCORT MARQUES, TULIO DAMINELLI BORGES, VANESSA CRISTINA CARPES DA SILVA, VANESSA FELISBERTO BILESIMO, WAGNER JOSÉ DENONI FREITAS, WELLINGTON ZOMER NUNES
Linguagem: POWER BUILDER, SQL
Campo de Aplicação: IF-10
Tipo de Programa: AT-02
Data da Criação: 15/09/2009
Regime de Guarda: Sigilo Até 29/05/2022
Procurador: DMARK REGISTROS DE MARCAS E PATENTES LTDA - CPF:03389474000165

Processo: 13333-6 **080**
Título: CAPTA CLIENTE
Titular: NOVOCLIENTE TECNOLOGIA LTDA - CPF/CNPJ:14962497000133
Criador: GUILHERME LEMOS SANTOS
Linguagem: PHP
Campo de Aplicação: AD-01, AD-02, AD-03, AD-05, AD-10
Tipo de Programa: GI-01, GI-02, GI-04, GI-05, GI-07
Data da Criação: 14/06/2012
Regime de Guarda: Sigilo Até 29/05/2022
Procurador: RICARDO PREIS DE FREITAS VALLE CORREA - CPF:63149591015

Processo: 13450-3 **080**
Título: PLATAFORMA E COMMERCE
Titular: EDUARDO MALVEIRO PEREIRA LEITE - CPF/CNPJ:06542126864
Criador: EDUARDO MALVEIRO PEREIRA LEITE
Linguagem: PHP
Campo de Aplicação: IF-09, SV-03, TC-02
Tipo de Programa: GI-01, GI-02, GI-04, GI-07, SO-07
Data da Criação: 10/06/2010
Regime de Guarda: Sigilo Até 21/06/2022
Procurador: SUL AMÉRICA MARCAS E PATENTES LTDA. - CPF:60848983000142

Processo: 13451-5 **080**
Título: CLIENTE TR - 69
Titular: EDUARDO MALVEIRO PEREIRA LEITE - CPF/CNPJ:06542126864
Criador: EDUARDO MALVEIRO PEREIRA LEITE
Linguagem: PASCAL
Campo de Aplicação: TC-02
Tipo de Programa: CD-04, GI-01, SO-06, SO-08
Data da Criação: 13/08/2012
Regime de Guarda: Sigilo Até 21/06/2022
Procurador: SUL AMÉRICA MARCAS E PATENTES LTDA. - CPF:60848983000142

Processo: 13452-0 **080**
Título: IP TOUCH
Titular: EDUARDO MALVEIRO PEREIRA LEITE - CPF/CNPJ:06542126864
Criador: EDUARDO MALVEIRO PEREIRA LEITE
Linguagem: C, PHP, PYTHON
Campo de Aplicação: TC-02
Tipo de Programa: CD-01, CT-01, SO-07, TI-01, TI-04
Data da Criação: 10/06/2010
Regime de Guarda: Sigilo Até 21/06/2022
Procurador: SUL AMÉRICA MARCAS E PATENTES LTDA. - CPF:60848983000142

ANEXO C – Certificado de Registro de Computador

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA E COMÉRCIO EXTERIOR
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

CERTIFICADO DE REGISTRO
DE PROGRAMA DE COMPUTADOR

Processo: 13281-6

O INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL expede o presente Certificado de Registro de Programa de Computador, **válido por 50 anos** a partir de 1º de janeiro subsequente à data de criação indicada, em conformidade com o art. 3º da Lei Nº 9.609, de 19 de Fevereiro de 1998, e arts. 1º e 2º do Decreto 2.556 de 20 de Abril de 1998.

Título: **COPA-TRAD: CORPUS PARALELO DE TRADUÇÃO**

Data de Criação 01 de Junho de 2011

Titular: 83.899.526/0001-82 **UNIVERSIDADE FEDERAL DE SANTA CATARINA**

Criadores: 807.832.529-00 **LINCOLN PAULO FERNANDES**
039.255.359-77 **CARLOS EDUARDO DA SILVA**

Linguagens: **JAVASCRIPT, PHP**

Campo de Aplicação: **CO-03**

Tipos de Programa: **FA-01, GI-01, GI-08, UT-01**

Documentação Técnica em depósito **SOB SIGILO até 16/05/2022.**

A exclusividade de comercialização do programa de computador objeto deste Certificado não tem a abrangência relativa à exclusividade de fornecimento estatuída pelo art. 25, inciso I da Lei Nº 8.666, de 21 de Junho de 1993, para fins de inexigibilidade de licitação para compras pelo poder público.

Expedido em 26 de Março de 2013.

Rodrigo Moerbeck de Almeida Rego
Chefe da Divisão de Registro de Programas de Computador e Topografia de Circuitos Integrados

Breno Bello de Almeida Neves
Diretor de Contratos, Indicações Geográficas e Registros