UNIVERSIDADE DE LISBOA

FACULDADE DE LETRAS

Validação de Regras Linguísticas para Melhorar a Qualidade do Português Europeu e Português do Brasil num Contexto de Pós-edição

JOÃO PEDRO GASPAR GERARDO

Dissertação orientada pela Professora Doutora Helena Gorete Silva Moniz e Doutora Marina Sánchez Torrón, especialmente elaborada para a obtenção do grau de Mestre em Tradução

2022

Índice

Índice de Figuras

Lista de Abreviaturas

ALPAC: *Automatic Language Processing Advisory Committee*

ANN: *Artificial Neural Networks* (Redes Neuronais Artificiais)

APE: *Automatic Post-editing* (Pós-edição automática)

BLEU: *Bilingual Evaluation Understudy*

CIA: *Central Intelligence Agency*

FAQs: *Frequently asked questions* (perguntas frequentemente perguntadas)

GPU: *Graphic Processing Unit* (Unidade de Processamento Gráfico)

IBM: *International Business Machines Corporation*

MQM: *Multidimensional Quality Metrics* (Métricas de qualidade multidimensionais)

NL: Holandês

NLP: *Natural Language Processing* (Processamento de Linguagem Natural)

NMT: *Neural Machine Translation* (Tradução Automática Neuronal)

PE: Português europeu

PB: Português do Brasil

QE: *Quality Estimation* (Estimativa de Qualidade)

R&D: Research & Development (Investigação e Desenvolvimento)

SMT: *Statistical Machine Translation* (Tradução Automática Estatística)

SVO: Sujeito-Verbo-Objeto

TA: Tradução Automática

Resumo

O presente trabalho foi desenvolvido no âmbito da unidade curricular "Estágio Curricular", exercida no segundo ano letivo do Mestrado em Tradução Da Faculdade de Letras da Universidade de Lisboa (FLUL). Este estágio terá sido realizado na *Unbabel*, uma empresa que oferece soluções de tradução baseadas em inteligência artificial, para domínios de apoio ao cliente, tais como as FAQs, ou seja, aos artigos de *Perguntas Frequentes*, o *chat* (diálogos escritos) e os *tickets* (por *tickets* entenda-se *emails* mandados por clientes para a equipa de apoio ao cliente de uma empresa). O foco deste trabalho será demonstrar o impacto e a eficácia de regras linguísticas em duas variedades da língua portuguesa, o português europeu e o português do Brasil.

O objetivo deste projeto foi investigar o impacto e eficácia de destacar erros presentes num ambiente de pós-edição de traduções feitas por modelos de tradução automática. Para este fim, foram identificados erros comuns através da análise de conjuntos de dados, em contexto de apoio ao cliente, de inglês para português europeu e inglês para português do Brasil. Com base nesta análise, foram criadas regras linguísticas com o objetivo de prevenir erros e melhorar a qualidade da pós-edição, assim como torná-la mais eficiente. Após uma fase de deteção de erros, criaram-se regras linguísticas como solução para os erros encontrados.

Para este efeito, analisaram-se conjuntos de dados constituídos por mensagens de apoio ao cliente, sob a forma de *tickets*, com os pares de línguas inglês-português europeu e inglês-português do Brasil, a fim de desenvolver soluções para os erros encontrados. Visa-se, desta forma, melhorar a qualidade e a eficiência dos editores que operam a pós-edição dos textos com erros de edição que afetam a qualidade. Após a fase de deteção de erros e de desenvolvimento das respetivas regras linguísticas, iniciou-se a fase de validação. Recorrendo a conjuntos de dados diferentes, testou-se a eficácia das regras desenvolvidas. Os resultados foram documentados e analisados na perspetiva de os editores aceitarem e implementarem as soluções propostas, embora exista a possibilidade de os editores escolherem ignorar as sugestões propostas pelas regras linguísticas. Contudo, assumindo que os editores tenham optado por aceitar as sugestões, observou-se um aumento significativo nos índices de qualidade dos conjuntos. Ao comparar os conjuntos de dados antes e depois de aplicadas as regras, registou-se, em termos de média, um aumento de 5 pontos no conjunto de dados no par linguístico inglês-português europeu e de mais de 11 pontos no conjunto de dados no par linguístico inglês-português do Brasil com base na métrica *standard Multidimensional*

Quality Metric (MQM). Em casos específicos, identificou-se um aumento muito considerável de quase 44 pontos nas mensagens com os valores mínimos para o par inglês-português e de 53 pontos para o par inglês-português do Brasil.
Palavras-chave: **Tradução Automática, Pós-edição, Sistemas de Identificação de erros, avaliação de qualidade, MQM**

Abstract

This work was done during the course of the Master's internship, in the second year of the Master's in Translation Studies in the Faculty of Arts and Humanities of the University of Lisbon (FLUL). The internship was at Unbabel, a company that focuses on AI-based translation services for customer service domains such as FAQs, chat and tickets (emails).

The goal of this research is to investigate the impact and effectiveness of highlighting errors in the machine translation output in a post-editing environment. To that end, linguistic rules addressing common errors in European Portuguese and Brazilian Portuguese were created. Common errors in European Portuguese were identified by analyzing and annotating English to European Portuguese and English to Brazilian Portuguese datasets in the context of customer service. Based on this analysis, a set of linguistic rules was created, aimed at preventing errors and improving the overall quality and efficiency of the post-editors. An error detection phase and subsequent creation of linguistic rules to solve the detected errors was followed by a testing phase of said rules. The results were documented and analyzed with the assumption that the editors would accept and implement the proposed solutions, although it is possible the editors choose to ignore and skip the suggestions proposed by the linguistic rules. However, assuming the editors accept the suggestions, there was a significant increase in the overall quality of the test datasets. By comparing the datasets before and after applying the rules, an overall increase of 5 MQM points was observed in English to European Portuguese; for English to Brazilian Portuguese, such increase was of 11 MQM points. Maximum increases were of 44 MQM points in English to European Portuguese and of 53 points in English to Brazilian Portuguese[1].

Keywords: **Machine Translation, Postediting, Error Detection Systems, Quality evaluation, MQM**

[1] This work was developed within the "Curricular Internship" unit, during the second year of Translation Studies Master's in the Faculty of Arts and Humanities of the University of Lisbon (FLUL). This internship was at the company, Unbabel, specialized in translation solutions based on Artificial Intelligence for customer support domains, such as FAQs, chat, and tickets.

Agradecimentos

Dedicado aos meus pais, por todo o apoio e pelos sacrifícios que fizeram por mim.

Para as professoras, Anabela Gonçalves e Sara Mendes que sempre me apoiaram e guiaram pelo melhor caminho. Para a professora Helena Moniz por todos os seus conselhos, recomendações, palavras de sabedoria, tremenda paciência e por acreditar em mim. Para a minha supervisora, Marina Sánchez-Torrón, para quem não tenho palavras que cheguem para agradecer tudo o que fez por mim, durante e depois do estágio, pela ajuda, pelos conselhos, por todo o apoio e pela tua amizade.

Para toda a equipa de serviços linguísticos. Para toda a equipa de *LangOps* que me apoia desde o dia em que os conheci. É o meu maior privilégio trabalhar lado a lado com vocês. Desde conselhos profissionais a conselhos e amizades para a vida. Um especial para a Vera Almeida que viu algo em mim e me deu uma oportunidade única, e para a Cristina Bugheanu, uma pessoa incansável no que toca a ajudar a sua equipa e qualquer pessoa, e a melhor *manager* que alguma vez irei conhecer. Para todas as pessoas que conheci ao longo do estágio e, por gestos e ações, ou por palavras de incentivo, me ajudaram a seguir em frente. Por me terem tornado uma pessoa melhor em tantos aspetos diferentes. Um sincero

Obrigado.

1. Introdução

A era da informação, trazida pela globalização, pela criação e pelo desenvolvimento da Internet em 1989, ofereceu à sociedade imensas possibilidades e facilidades, mas também trouxe novos desafios. Embora, inicialmente, a língua inglesa tenha sido a mais disseminada na Internet, outras línguas, tais como o chinês e o espanhol, têm vindo a mostrar a sua importância e influência. Mais recentemente, a língua portuguesa tem sido alvo de destaque devido à vasta demografia que se verifica no Brasil e tornou-se, segundo um estudo da *Statista* em 2020, a sexta língua mais falada na Internet. Contudo, neste país, o português assume uma variante diferente do português europeu (PE), o português do Brasil (PB), levantando questões quanto ao resultado do estudo. Portugal conta com cerca de 11 milhões de falantes de português europeu, enquanto o Brasil tem mais de 210 milhões de habitantes que comunicam em português do Brasil. Olhando para esta enorme diferença demográfica, as empresas usam frequentemente a simples escolha lógica de apenas utilizarem a variante PB.

Desde agências de viagem a *websites* de jogos, a tradução portuguesa tem sido feita quase exclusivamente para o PB, talvez por ser desconhecida para muitos a existência das duas variantes ou da significância das suas diferenças. Contudo, há imensos desvios linguísticos que, por vezes, chegam a ser ofensivos para os falantes de PE e vice-versa.

Visto que o estágio foi efetuado no departamento de *Community*, ou seja, o departamento que lida com os editores/tradutores, os avaliadores e os linguistas, o tema da tese alinha-se com o foco do departamento: preservar os interesses da comunidade e melhorar as condições e ferramentas ao seu dispor. A investigação original proposta para esta tese tinha como objetivo criar e testar regras linguísticas que distinguissem as duas variedades portuguesas em questão, tendo em conta que os relatórios internos da empresa apontavam para editores de PE editarem textos de PB e vice versa. Desta forma, decidiu-se detetar todos os tipos de erros presentes nos conjuntos de dados disponibilizados nas duas variedades e procurar desenvolver soluções que apresentassem alternativas aos editores.

1.1 Motivação

Este projeto surgiu durante o estágio curricular, exercido no segundo ano letivo do Mestrado em Tradução da Faculdade de Letras da Universidade de Lisboa na empresa *Unbabel*, especializada em oferecer soluções de tradução baseadas em inteligência artificial para domínios de apoio ao cliente, tais como as FAQs, o *chat* e os *tickets* (por *tickets*, entenda-se *emails* mandados por clientes para a equipa de apoio ao cliente de uma empresa).

Houve uma oportunidade de abordar um dos problemas mais urgentes, através do sistema proprietário de identificação de erros, *Smartcheck*. Este tema surgiu devido ao facto dos erros entre o português europeu e o português do Brasil serem comuns e, adicionalmente, acrescentam-se à análise de dados da empresa que referiam a ocorrência de vários editores de uma destas variedades linguística editarem traduções da outra variedade. O foco deste trabalho será demonstrar o impacto e a eficácia de regras linguísticas em duas variedades da língua portuguesa, o português europeu (PE) e o português do Brasil (PB).

Embora os dados não tivessem corroborado este facto, houve ainda a oportunidade de fornecer aos editores mais ferramentas, para agilizar o processo de pós-edição, através do desenvolvimento de soluções testadas para problemas comuns e/ou urgentes encontrados nos conjuntos de dados disponibilizados. Desta forma, adotou-se a estratégia de criar e inserir regras na plataforma *Smartcheck* que seriam mostradas aos editores como sugestões de um corretor ortográfico.

1.2 Objetivos

Originalmente, o foco inicial da tese era detetar erros de variedade linguística entre o PE e o PB, sob a forma de sinais de contaminação do português europeu em textos escritos em português do Brasil e vice-versa. Posteriormente, procurou-se a deteção de erros em geral e possíveis soluções, a fim de aumentar os índices de qualidade do produto da empresa. Estes erros têm diversas origens, podendo surgir da tradução automática, do uso de dados da variedade errada durante o treino dos sistemas de tradução automática ou da pós-edição, realizada pelos editores a quem são distribuídos textos com índices de qualidade aquém dos desejados. O objetivo era desenvolver soluções, nomeadamente através de regras, para o sistema de identificação de erros, o *Smartcheck*. Estas regras detetam padrões e são criadas a partir de conceitos familiares na área de processamento de língua natural explorados posteriormente na seção 3.3, a "tokenização" das palavras, recorrendo a *PoS* (*Part-of-Speech*) *tagging*. Com este processo, é possível conferir aos modelos de tradução automática a capacidade de compreensão linguística aos através da separação das palavras nas suas respetivas categorias gramaticais, via *TurboParser* (Martins *et al.* 2014), que inclui informação morfológica, e através de operadores lógicos. Deste modo, é possível criar as regras específicas, para cada par linguístico e cliente, que atuam como avisos na interface dos editores. Através do *Smartcheck*, não só se torna possível apresentar soluções alternativas aos editores, como proporcionar melhorias na qualidade da sua experiência e produtividade, enquanto trabalham na plataforma. Assim sendo, esta tese vai assentar no estilo, com ênfase

no registo lexical/gramatical, e na falta de discriminação entre as variedades linguísticas PE *vs*. PB.

1.3 Estrutura da tese

Primeiramente, no ponto 2, será feita uma pequena apresentação da empresa e dos conceitos básicos necessários para entender a estrutura da mesma. Segundamente, no ponto 3 serão partilhados alguns conceitos básicos da Tradução Automática, o contexto de como a mesma surgiu e se desenvolveu. No mesmo capítulo, serão mencionadas algumas diferenças entre as variedades portuguesas e, de seguida, será exposto no ponto 4 a metodologia adotada assim como um projeto-piloto realizado antes da investigação principal. Posteriormente, no ponto 5, serão analisados os resultados, as principais questões e os principais erros, assim como as possíveis soluções a considerar, culminando nas conclusões presentes no ponto 6.

2. A entidade de acolhimento do estágio[2]

A *Unbabel* é uma *start-up* portuguesa, especializada na tradução e fornece serviços de Tradução Automática (TA) e de pós-edição humana com enorme velocidade e qualidade. Este sistema híbrido consiste na tradução de um documento feita por um sistema de tradução automatizado e customizado para cada cliente, respeitando todas as regras de proteção de dados. Posteriormente, é segmentado e distribuído para a comunidade de tradutores e editores para corrigir possíveis erros. Quando o documento estiver completo na sua íntegra, segue para o cliente, mas, também, para um anotador que dará *feedback,* desde o processo automático às alterações implementadas pelos editores. A empresa de *software* foi fundada em 2013 por Vasco Pedro, João Graça, Sofia Pessanha, Bruno Silva e Hugo Silva, e a ideia inicial surgiu num dia em que o grupo praticava surf, um passatempo que é encorajado e bem representado na empresa.

A sede da empresa está situada em São Francisco, na Califórnia, e tem escritórios em Lisboa, Nova Iorque e São Francisco. Emprega 170 funcionários oriundos de 27 países, representando 17 línguas faladas na mesma empresa. Já foi vencedora de vários prémios tais como: dois prémios de inovação TAUS (*Translation Automation User Society*) em 2015 e 2017, os prémios *WMT16* e *WMT19* (*Best Global Machine Translation Quality Estimation*), o prémio *National Innovation Award* em 2019 e o prémio 2019's AI 100 (*List of Most innovative Artificial Intelligence Startups for Disruptive Technology of the Year*).

O objetivo da empresa é obter uma comunicação universal, construindo "torres de conhecimento", visto que a mesma se inspirou no mito cristão *A Torre de Babel* (Génesis 11:1-9) que explicava a existência das várias línguas faladas no mundo. Este mito retrata as gerações posteriores à Arca de Noé que se juntavam e comunicavam numa só língua. Em conjunto, começaram a planear construir uma cidade com uma torre que alcançasse os céus e, assim, ficarem famosos. Deus, descontente, desceu à Terra, confundiu a língua que os homens falavam em várias e dispersou-os pela Terra. Segundo a enciclopédia *Britannica*, o mito poderá ter sido inspirado pela torre do templo *Marduk,* chamada *Bab-ilu (Gate of God)*. A semelhança entre as palavras Babel e Balal (balal sendo o verbo confundir) foi, provavelmente, a razão do trocadilho utilizado na Bíblia, visto que foi aí que Deus terá misturado todas as línguas do mundo.

---

[2] Este estágio curricular foi realizado antes da integração com a Lingo24. Por esta razão, a natureza da sua descrição é focada apenas na Unbabel.

A *Unbabel* pretende edificar "torres de conhecimento" e uma compreensão universal em todo o mundo. Com o objetivo de unir a humanidade através da compreensão entre as mais diferentes línguas, conta com uma comunidade de milhares de tradutores, editores, anotadores e linguistas que trabalham no total com 29 línguas. Tendo em conta que a comunicação é a base da civilização, da cultura e do comércio, num mundo digital, a comunicação tem de transcender fronteiras, fusos horários, mercados e línguas. Dada a natureza e o objetivo da empresa, os serviços destinam-se, principalmente, a grandes empresas e são, na sua essência, com diferentes exigências de qualidade e prazos, FAQs (perguntas frequentes), *emails* (*tickets* de apoio ao cliente) e mensagens por *chat*. As línguas com mais volume de trabalho são o alemão, o francês e o inglês, mas a empresa também destaca línguas como o hindi e vietnamita, que, no fim, culminam na oferta de 72 pares de línguas possíveis.

Dentro da empresa, existem vários departamentos que trabalham em vários aglomerados denominados *clusters* com diversos encargos e responsabilidades. Dentro de cada *cluster*, existem equipas de várias áreas e especialidades, desde engenheiros, a cientistas da linguagem ou a linguistas, que trabalham em conjunto para melhorar os produtos e serviços fornecidos. O estágio decorreu dentro do *Community Cluster*, mais especificamente, na equipa de *R&D* (*Research & Development*). No topo deste departamento está sempre o cliente, seguido da comunidade, que inclui tradutores/editores - que melhoram os textos em pós-edição, anotadores - que fornecem *feedback* para todas as fases, e linguistas que contribuem para a elaboração e atualização de guias linguísticos. Existe também a equipa de investigação e de processamento de linguagem natural (NLP- *Natural Language Processing*) e, por último, a plataforma que os editores e anotadores utilizam para desempenharem as suas tarefas.

Community tem como função construir e sustentar uma comunidade global de tradução que seja capaz de corresponder às exigências dos clientes da empresa. A sua missão é manter o equilíbrio entre a trindade velocidade-qualidade-custo, focando-se em ter sempre os melhores índices de qualidade (medidos através da métrica MQM- *Multidimensional Quality Metrics,* Lommel, 2015, que será descrita adiante) acompanhados de uma tradução rápida, tudo em conformidade com os custos. Por este motivo, a equipa de *Community R&D*, por um lado concentra os seus esforços em melhorar a plataforma onde os editores realizam as suas tarefas, ao implementar *quality of life changes* (funcionalidades que facilitam e/ou agilizam os processos) através da criação e/ou melhoria de ferramentas de apoio de que os editores e anotadores podem usufruir na plataforma. Por outro lado, foca-se em assegurar a

qualidade da comunidade, por exemplo, através da avaliação periódica dos editores. Esta missão exige um bom índice de qualidade de tradução para cada tarefa realizada. Recorrendo ao recurso externo, *Multidimensional Quality Metrics*, que mede a qualidade de uma tradução até 100 pontos percentuais, a empresa desenvolveu métricas compatíveis com este recurso. Se a tarefa contiver erros de tradução, o MQM calcula uma penalização consoante o tipo de erro e, por consequência, a gravidade do dito erro. A tipologia das MQM divide-se em três categorias: precisão, fluência e estilo, que são expandidas em outras subcategorias.

Deste modo, os sistemas de tradução beneficiam de dados que são alimentados durante as fases de treino, para o seu constante desenvolvimento, e os editores recebem *feedback* periódico. A qualidade desejada é dependente de múltiplos fatores. Entre estes está o género de conteúdo, visto que existem diferenças nos requisitos de *tickets*, mensagens por *chat* e FAQs, e os índices de qualidade estipulados pelo próprio cliente.

LangOps, *language operations*, é outro departamento relevante e é encarregado da responsabilidade de introduzir os novos clientes no sistema da *Unbabel*, assim como gerir e/ou criar recursos linguísticos, ou seja, memórias de tradução e glossários, e reforçar no sistema as regras impostas pelos clientes, bem como as suas preferências relativamente às traduções. O departamento assume um papel importante durante a relação entre *Unbabel*-cliente do ponto de vista de qualidade, estando em constante contato durante a fase de iniciação até ambas as entidades estarem satisfeitas com a qualidade das traduções. Desta forma, acaba por haver uma forte ligação com o departamento de *Community*.

Entenda-se por comunidade o conjunto de editores e tradutores que desempenham as tarefas de pós-edição, havendo ainda uma comunidade mais especializada composta por linguistas e tradutores com elevados níveis de competências, denominada *Community Pro* que desempenha, principalmente, as revisões das traduções finais e fornece *feedback* sobre a qualidade das traduções. Durante a fase de pré-produção, após ler o relatório do cliente, um dos especialistas da equipa de *LangOps* procura assegurar glossários e memórias de tradução com cobertura de dados e com qualidade, antes de os fornecer aos clientes para validação, implementando as alterações necessárias e reportando erros da tradução automática e/ou padrões de erros críticos.

Resumindo, nas figuras 1 e 2, retiradas de uma apresentação interna de Phil Brougham, *senior Product Manager*, tendo sido cedida a respetiva autorização para poder partilhar nesta tese, pode-se observar os dados da *Unbabel* e o efeito que têm nos seus clientes. Na figura 1, é destacado o valor de palavras que são traduzidas pela *Unbabel* mensalmente, ou seja, 150 milhões de palavras traduzidas mensalmente. Outro número a

indicar são os 100 mil tradutores que constituem a comunidade e componente humana e contribuem para o sucesso da empresa. A figura 2 exibe os efeitos nas empresas que recorrem aos seus serviços. A *Unbabel* contribui para uma redução significativa dos custos de clientes que pretendem expandir para mercados estrangeiros e o aumento da satisfação dos clientes.

Figura 1. *Unbabel stats*, imagem cedida e autorizada por Phil Brougham

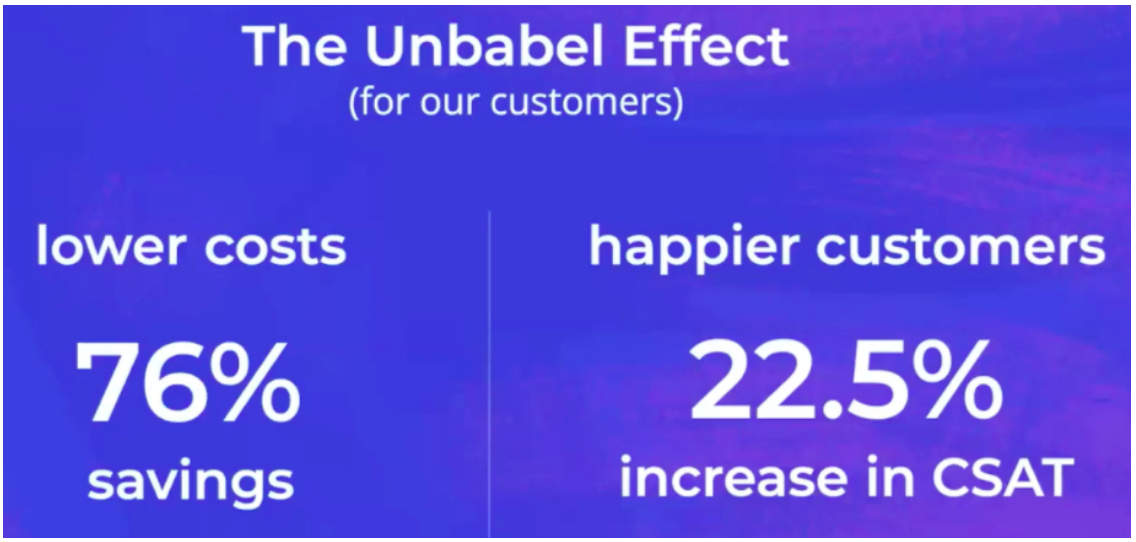

Figura 2. *Unbabel stats*, imagem apresentada internamente por Phil Brougham

2.1 *Pipeline* utilizada pela *Unbabel*

Ao longo da arquitetura deste sistema híbrido, no contexto da cooperação entre inteligência artificial e ser humano, um sistema de tradução automática irá traduzir inicialmente o produto na sua íntegra, enquanto a componente humana é integrada posteriormente, caso seja necessário, para manter a qualidade do produto final e eliminar potenciais erros não detetados pela componente artificial, de modo a melhorar a qualidade da

mesma. A arquitetura ilustrada na figura 3 demonstra o processo que é realizado para cada pedido do cliente relativamente a *tickets* (*emails* de apoio ao cliente).

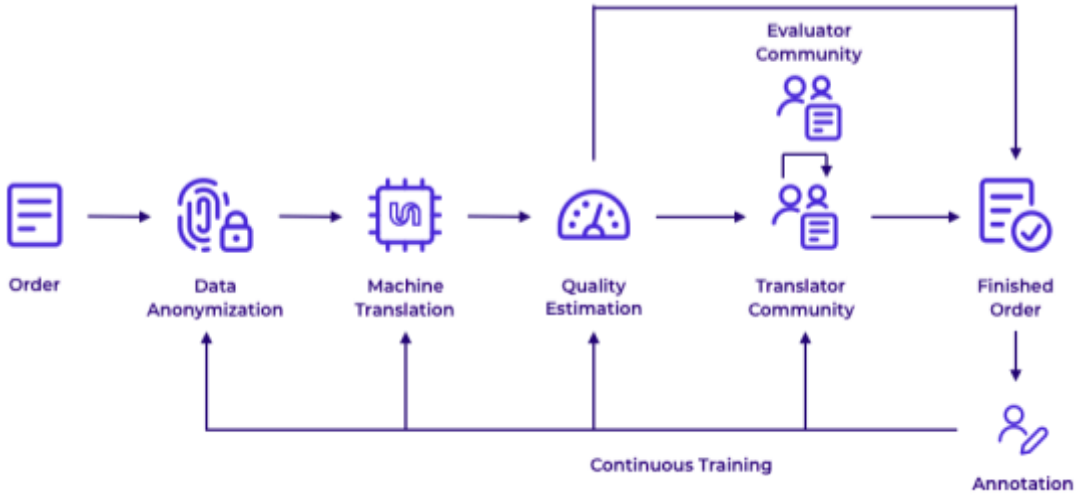

Figura 3: A arquitetura utilizada pela *Unbabel* para emails

Segundo um relatório interno, ilustrado na figura 3, num primeiro passo, o cliente submete os diversos textos para serem traduzidos, neste caso, *emails*, a traduzir que, na segunda etapa, terão todas as informações pessoais e sensíveis, anonimizadas de acordo com a Lei da Proteção de Dados Pessoais (lei n.º 58/2019, de 08 de agosto) e outras legislações relativamente à proteção de dados. Estas informações englobam endereços de *email*, nomes e apelidos, números de telemóvel e de crédito, entre outros. Para que o contexto não seja perdido devido à fase de anonimização, as informações são substituídas por *placeholders* (caracteres de preenchimento) como *NAME-0 ou EMAIL-0*. Após a anonimização, os textos prosseguem para a fase de tradução automática, esta componente de inteligência artificial é efetuada por um modelo criado e aprovado especificamente para cada cliente, visto que os mesmos são programados e treinados com dados dos clientes depois de serem revistos pelos anotadores. Este treino é constante e realizado regularmente com o objetivo de melhorar continuamente o sistema automático. Após cada sessão de treino, são comparados os valores de qualidade anteriores com os valores dos conjuntos de teste obtidos após o treino, a fim de determinar se houve melhoria nos índices de qualidade, calculada através de um sistema concebido internamente. De seguida, realiza-se a avaliação de qualidade que irá determinar se os pedidos dos clientes alcançaram o nível de qualidade desejado para serem entregues e concluídos, ou, caso não tenham índices de qualidade aceitáveis, se passarão para a fase de pós-edição executada pela comunidade de centenas de editores/tradutores e revista pelos avaliadores que transmitem *feedback* sobre os textos e editores.

Estas avaliações dos índices de qualidade e comparação dos resultados obtidos nos treinos dos modelos de tradução medem o nível de qualidade e eficácia das traduções, feitas pelos modelos de tradução automática ou por editores humanos. A qualidade destas traduções vai depender do tipo de conteúdo e das características associadas a cada um. Os *tickets* terão mais contexto ao longo da mensagem, assim como tempo necessário para que haja envolvimento humano no processo, tornando a expetativa de qualidade elevada. Por oposição, o conteúdo proveniente de canais de *chat* não passa por uma fase de edição humana após a tradução automática, devido à natureza quase instantânea da forma de comunicação. Os dados de chat são pré-treinados com dados editados deste domínio, de modo a assegurar a necessária qualidade. Porém, sendo um meio em que as mensagens são enviadas e respondidas em tempo real, as expectativas de qualidade não podem ser iguais às desejadas com os *tickets*. Por último, as FAQs, dispõem de uma quantidade de tempo suficiente para haver uma fase de edição sénior, além da fase de edição humana realizada também na arquitetura aplicada para os *tickets*. Explorando mais a componente de avaliação, as traduções têm um nível CUA, *Customer Utility Analysis*, atribuído para poder ser possível visualizar a qualidade das traduções. Os níveis designam-se por *Excelente, Bom, Moderado, Fraco*, cujos parâmetros variam consoante o tipo de conteúdo. O nível *Excelente* é atribuído a traduções que têm um nível de MQM entre 90 a 100 no caso de FAQs, 85 a 100 para *tickets*, 80 a 100 para *chat* e é um nível normalmente alcançado em traduções que tiveram uma fase de pós-edição humana. O nível *Bom* situa-se entre os valores 75 a 89 para FAQs, 70 a 84 para *tickets* e 60 a 79 para *chat*. Os níveis com uma tradução cuja mensagem pode ter alguma dificuldade a ser compreendida situam-se entre 60 a 74 para FAQs, 50 a 69 para *tickets* e 40 a 59 para *chat*. Por último, as traduções com erros que podem comprometer seriamente a mensagem são avaliadas como *Fracos* e situam-se entre menos de 0 a 59 para FAQs, menos de 0 a 49 para *tickets* e menos de 0 a 39 para *chat*.

As avaliações das traduções e dos treinos dos modelos são executadas, principalmente, pelo *COMET*, Rei *et al. 2020*, desenvolvido internamente na empresa. A métrica *Crosslingual Optimized Metric for Evaluation of Translation* foi criada com o propósito de prever automaticamente o que um humano acharia da qualidade de uma tradução, permitindo calcular o valor MQM das traduções e, desta forma, fazer a estimativa do índice de qualidade. *COMET,* é uma métrica automática baseada em redes neuronais, essencialmente, conjuntos de diferentes algoritmos, concebida em 2020 e considerada a ferramenta mais eficaz para calcular MQM na 5ª Conferência Mundial de Tradução Automática (WMT20). Quando o *Comet* avalia uma tradução automática, compara-a com o

texto original e com uma tradução de referência feita por um humano. Com esta metodologia, *COMET* "aprende" a prever as alterações que um humano preferiria e, desta forma, prevê um índice de qualidade. A esta métrica também se acrescentam anotações realizadas por humanos, de forma a avaliar o seu desempenho.

Uma das convicções usadas dentro da empresa – *you can't improve what you can't measure* – "não se pode melhorar o que não se pode avaliar", aplica-se a todas as fases do produto e departamentos internos, quanto mais informação e dados se tem sobre determinado sistema ou módulo, mais se pode progredir e melhorar. No que diz respeito a recolher dados sobre as traduções automáticas realizadas pelos modelos, a *COMET* vem fazer jus a esta convicção e necessidade. Tendo em conta que as traduções, e as línguas em si, são algo subjetivo e suscetível a mudança entre regiões e comunidades, foi constituída uma tipologia onde os vários tipos de erros têm uma severidade associada que influenciará a métrica MQM. Os erros menores incluem erros de pontuação, entre outros e não tendem a afetar a compreensão da mensagem, podendo, contudo, resultar num ligeiro sentido de estranheza a um nativo. Por sua vez, os erros graves afetam a compreensão, provocando uma maior sensação de estranheza. Porém, não tendem a impedir a leitura da mensagem e o seu sentido. Os erros que omitem informação essencial, podem causar consequências legais, financeiras ou outras aos clientes são designados erros críticos.

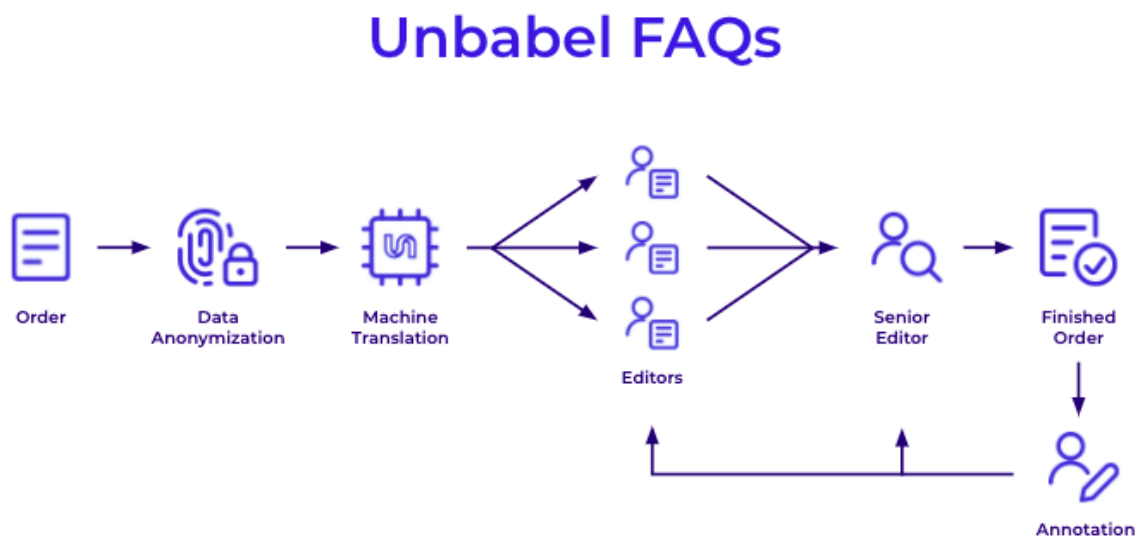

Figura 4: A arquitetura utilizada pela *Unbabel* para FAQs

No que diz respeito às FAQs, cada artigo submetido para tradução é fragmentado e disperso entre vários editores, culminando na revisão do artigo recomposto na sua íntegra por

um editor sénior. Porém, na fase final da arquitetura, o processo é semelhante à de *tickets*, havendo o processo de anotação no fim da tradução do artigo estar completa.

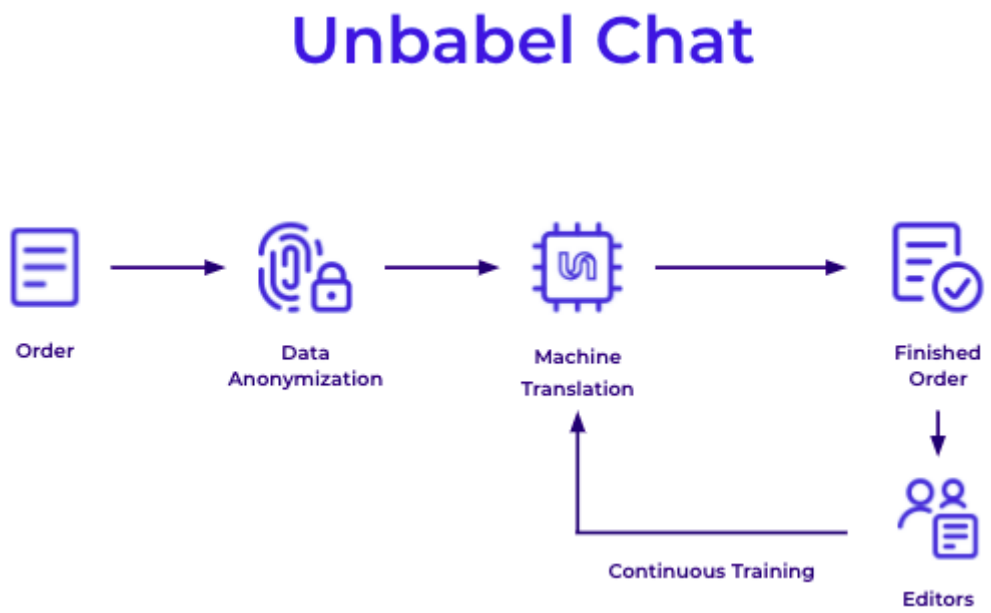

Figura 5: A arquitetura utilizada pela *Unbabel* para *chats*

No caso de *chat*, cujo conteúdo decorre num espaço de tempo demasiado curto, há uma necessidade de priorizar velocidade acima de qualidade, visto que, dado o tipo de conteúdo, não é possível recorrer à comunidade de tradutores. Deste modo, para compensar a perda do processo humano por velocidade, além dos treinos dos modelos de Tradução Automática, há um maior foco no treino de modelos de TA e na fase de anotação, que ocorre após a tradução ser entregue.

2.2 *Smartcheck*

Para que a comunidade de editores consiga fazer o seu melhor, é lhes concedido o acesso a recursos linguísticos que facilitam e agilizam o processo de pós-edição. Os editores podem usufruir de diversas ferramentas desde glossários e memórias de tradução a corretores ortográficos, e, além desses, podem fazer uso das sugestões provenientes do *Smartcheck*, um programa concebido para melhorar a consistência das traduções, assim como a aspectos gramaticais e ortográficos dos textos. Este não é um mero programa, mas um conjunto de ferramentas que consiste tanto em corretores ortográficos, para detetar erros ortográficos incluindo um da própria empresa, que identificam erros ortográficos, bem como em *hardcoded rules*, que lidam com erros puramente tipográficos. Por erros tipográficos, entendam-se duplo espaçamento entre palavras e parênteses não fechados, entre outros. Por

último, é ainda composto do sistema *Surf*, uma ferramenta desenvolvida dentro da empresa para a construção de regras específicas e de sugestões, customizáveis para cada cliente e língua. Este processo de criação de regras é realizado por linguistas que também supervisionam a sua eficácia ou os problemas que possam surgir com a sua criação, ou seja, falsos positivos ou inconsistência com guias de estilo. A execução das regras é facilitada com a *PoS tagging*, a "tokenização" de cada palavra em informação gramatical realizada pelo *TurboParser* (Martins *et al.* 2014), um analisador sintático que usa um sistema de dependências, no caso da *Unbabel* o sistema adotado é o *Universal Dependencies* (*UD*), e não de constituintes como o *LX-Parser* (Silva, João et al.), que assenta em relações entre palavras e permite elaborar regras que generalizam categorias gramaticais e características morfológicas. Seguindo o *website*, por dependências universais, entenda-se uma arquitetura elaborada para que exista uma referência geral a fim de obter anotações gramaticais nas mais diversas línguas feitas por vários linguistas. O sistema *UD* possibilita uma anotação gramatical das palavras e respetivas funções que têm numa frase, baseado em dependências (Marneffe *et al*., 2006; 2008 e 2014), em categorias gramaticais desenvolvidas por Petrov *et al.* (2012) e um sistema de categorias gramaticais para anotações, *Intersect*, desenvolvido por Zeman (2008). Estas categorias classificam classes de palavras e respetivas categorias morfológicas, seguindo critérios diferentes.

Explorando mais a *tokenização* e a segmentação de frases, as dependências universais assentam na sintaxe entre palavras, ou seja, nas relações de dependência e nas interações entre palavras. Neste processo é efetuada a divisão ao nível sintático e não ao nível fonético ou ortográfico. Por exemplo, a palavra "numa" segmenta-se como "uma" + "em". Nos casos de línguas com propriedades mais específicas, como, por exemplo, o chinês tradicional, o chinês simplificado e o japonês que não utilizam espaçamento entre palavras, é necessário um algoritmo mais especializado em segmentação ao nível da palavra. Após a segmentação, ocorre a identificação de cada entidade lexical reconhecida. As principais categorias gramaticais são:

Categorias Abertas	**Categorias Fechadas**
ADJ – Adjetivo	ADP – Preposição e *Postposition*
ADV – Advérbio	AUX – Auxiliar
INTJ – Interjeição	CCONJ – Conjunção Coordenativa
NOUN – Nome	DET – Determinante
PROP – Nome Próprio	NUM – Número
VERB – Verbo	PART – Partícula
	PRON – Pronome
	SCONJ – Conjunção Subordinativa

Tabela 1: categorias gramaticais em *Universal Dependencies*

UD para português não necessita de um algoritmo como para chinês ou japonês. Em termos de segmentação, as palavras são espaçadas e separadas por pontuação. A maioria dos corpora reunidos para português contém *tokens* constituídos por várias palavras, devendo-se ao facto de as contrações serem tão prevalentes na língua portuguesa. Entre estas palavras "multi-tokenizadas" destacam-se principalmente as contrações, como foi mencionado anteriormente, entre determinantes e preposições, assim como formas verbais cliticizadas em que se observa mesóclise, como por exemplo, "ensinar-lhe-ei", segmentado como "contar+lhe+ei". Também há que destacar as palavras hifenizadas que são consideradas como um *token* atualmente. Vale a pena mencionar que há exceções a esta segmentação, sendo este o caso dos acrónimos e das abreviaturas, tal como "EUA".

Segundo Costa (2016), um analisador sintático consiste num sistema que identifica as cadeias de dependência presentes em frases. Porém, como com todos os modelos automáticos de Processamento de Língua Natural, é necessário recorrer à análise estatística, de modo a lidar com as questões de ambiguidade, visto que a outra variante de modelos se baseia em regras que têm de ser inseridas manualmente. Deste modo, os modelos estatísticos demonstram mais potencial, apesar das suas limitações. A simples utilização dessas regras não garante que o analisador sintático encontrará a solução menos ambígua. A ambiguidade inerente às línguas naturais é um problema que pode ser tratado por meio da análise sintática probabilística. Segundo tal estratégia, atribuem-se pesos às regras, de modo a que o

analisador possa decidir qual a melhor regra a ser aplicada em um dado momento, lidando com a ambiguidade inerente de uma língua natural.

Adicionalmente, existe um subconjunto de regras que podem ser criadas e desenvolvidas para resolver, por exemplo, erros comuns entre editores. Estas regras linguísticas são criadas na *Surfboard* sendo chamadas regras *Surf* e são testadas num sistema "teste" (do original inglês *staging*), antes de serem utilizadas pelo *Smartcheck* em produção. O sistema *staging* consiste num espaço virtual onde as novas regras são testadas ao nível da sua eficácia e da sua cobertura, antes de serem aprovadas, rejeitadas ou modificadas, de modo a serem oficialmente ativadas e, assim, reconhecidas e aplicadas pelo *Smartcheck* em conformidade.

Embora sejam ambos provenientes do *Smartcheck*, existem algumas diferenças entre os dois exemplos que serão apresentados de seguida. A figura 6 demonstra um típico erro ortográfico, em que o editor só tem de selecionar o erro, escolher a opção com que quer substituir e clicar nessa opção para a implementar. Com as regras *Surf* é possível criar regras linguísticas que detetem erros gramaticais e/ou de estilo, ao invés da simples deteção de erros ortográficos. Adicionalmente, é possível adicionar uma categoria ao erro identificado e apresentar contexto, caso seja necessário, mencionando uma breve explicação, como por exemplo, "*Tu is too informal, please consider omitting it if possible*" ("Tu é demasiado informal, por favor considere omiti-lo se possível). Contudo, a figura 7 é ligeiramente diferente, sendo um erro captado por uma regra *Surf*. Continua a ser possível ver a descrição do erro e as sugestões, porém, para implementar a sugestão pretendida, é necessário inseri-la manualmente, aumentando a carga de trabalho manual dos editores em troca de melhor qualidade. Este ponto é relevante, visto que representa o maior ponto de fraqueza desta ferramenta. O *Smartcheck* exibe uma enorme dependência do editor, como todos os outros sistemas de deteção de erros, mas o ligeiro aumento do trabalho manual que as regras *Surf* exigem é uma desvantagem conhecida para a eficácia delas. Por definição, a ferramenta serve como um banco de sugestões, e, por essa razão, para que realmente afete a qualidade do produto, precisa que o editor em questão aceite e queira implementar as sugestões propostas, estando, portanto, vinculado às decisões do editor.

Figura 6: Exemplo de um erro detetado por um corretor ortográfico do *Smartcheck* sendo que é necessário selecionar a opção sugerida para a implementar

Figura 7: Exemplo de um erro detetado por uma regra *Surf* do *Smartcheck* em que é necessário inserir a sugestão manualmente

Em conjunto com as outras ferramentas de identificação de erros, através destas regras linguísticas, tanto em PE como em PB, podemos ajudar os editores destas variedades a melhorar a qualidade das suas pós-edições, passando pela deteção de erros que poderão escapar durante a revisão do editor. Da mesma forma, permite-nos ajudar a agilizar o processo de pós-edição destacando visivelmente erros com sugestões de correção específicas e de confiança.

3. Fundamento teórico

Esta seção dedica-se à explicação e à introdução da área da Tradução Automática (TA) e a um breve resumo da sua história, bem como à contextualização das principais arquiteturas utilizadas e dos conceitos essenciais recorrentes nesta tese. Serão mencionadas as posições de figuras de renome quanto à TA e aos períodos históricos que mais marcaram esta área de investigação. Também serão discutidas e exploradas as duas arquiteturas principais e mais desenvolvidas, a Tradução Automática Estatística e a Tradução Automática Neuronal, sendo esta considerada o estado da arte atual.

Pode-se definir a Tradução Automática simplesmente como "a tradução automática de um texto escrito numa língua humana para outra língua". Em *The Routledge Handbook of Translation and Philosophy* (Rawling & Wilson, 2019:428) há referência a uma outra definição de TA dada por Dorothy Kenny, "o meio que garante os mitos de significado e de compreensão universais, assim como transparência linguística" (tradução própria aproximada). Esta frase, além de fazer alusão aos valores da *Unbabel*, provoca uma reflexão relativamente à utopia que seria um mundo sem obstáculos à comunicação. Teríamos alcançado um nível tecnológico que nos permitiria falar com outra pessoa em línguas diferentes e desconhecidas e, em tempo real, ter a sua tradução? O mais próximo dessa utopia, atualmente, é a divulgação e expansão da língua inglesa como uma "língua de negócios", o que levanta a questão da hegemonia da língua inglesa relativamente a todas as outras. Até recentemente, grande parte da comunicação virtual era feita, principalmente, através da língua inglesa. Por outro lado, durante a globalização, o emprego da TA atingiu enormes proporções de utilização e vários autores, como Cronin (2013), defendem que este uso da TA contribuiu para a diversidade linguística nos meios virtuais (Rawling & Wilson, 2019:428). Deste modo, as empresas multinacionais viram esta tecnologia como uma chave essencial para a sua contínua expansão.

Todavia, há quem argumente, como é o caso de Dorothy Kenny (2019:438), que é necessário haver distinção entre tradutor humano e TA. Pode-se argumentar que é principalmente devido à questão de se uma tradução realizada automaticamente poderá ser considerada uma tradução fidedigna, visto que as traduções contemporâneas, como as realizadas para livros e textos científicos, são traduzidas por um tradutor humano. De facto, existe a distinção entre tradução humana, em que o agente principal é o humano que dispõe de ferramentas e de recursos, e a TA, em que o processo que embora seja executado exclusivamente pela máquina, tem de ter sempre interação humana, seja em pré ou pós-edição. Contudo, de igual forma, existem vários artigos, como *Aliens and Linguistics:*

Language Study and Science Fiction de Walter E. Meyers (1982) que descrevem os tradutores profissionais como "ciborgues" que já não conseguem trabalhar independentemente destas tecnologias, como, por exemplo, o *SDL Trados*, o mais popular e relevante no mercado, e o *Wordfast Anywhere*, uma das alternativas de acesso gratuito. Tal deve-se ao facto de estas tecnologias aumentarem a eficiência dos tradutores exponencialmente, especialmente devido às memórias de tradução, aos glossários e à TA.

Os recentes avanços na área de TA levantam algumas reflexões sobre a continuação da profissão de tradutor. Adicionalmente, há críticas vindas da comunidade de tradutores profissionais. Uma boa parte da comunidade em regime de *freelance* repudia o progresso desta área, devido ao receio de que a sua profissão seja colocada em risco. É um tópico sensível para muitos e fácil de se simpatizar. Nem todos se conseguem adaptar às novas necessidades da área da Tradução como, por exemplo, seguindo uma carreira como linguista tendo a função de ajudar os cientistas da linguagem a melhorar os sistemas de TA. Porém, os argumentos sobre a substituição da tradução humana pela TA não têm em consideração o facto de que é mais prevalente a complementação que a tradução humana oferece à TA, sob a forma de fornecimento de dados para o treino de sistemas ou sob a forma de prestação de serviços de pré e/ou pós-edição, como é o caso da arquitetura da *Unbabel*.

Relativamente a esses sistemas, em primeiro lugar, é fundamental que exista mais contextualização, para facilitar a compreensão dos sistemas descritos posteriormente. Na definição prévia de TA, foi mencionada a transferência de texto com o pressuposto que é escrito. Existem estudos que se encontram, atualmente, a desenvolver outras tecnologias que possam fazer uso de texto "não escrito", tal é o caso de tradução automática de fala e até de sinais gestuais, *e.g.*, o modelo de *DeepHands* (*Koller et al.*, 2015) e *Avatar-based Translation from European Portuguese to Portuguese Sign Language* (*Coheur et al.*, 2015).

3.1 Contextualização da Tradução Automática

De acordo com Dorothy Kenny (2019), a Tradução Automática começou a ganhar destaque após a Segunda Guerra Mundial. Provavelmente, relacionada com intenções bélicas e de recolha, ou partilha, de informação. Contudo, de acordo com Hutchins (2011), um dos historiadores mais importantes e relevantes da área, os primeiros sistemas de tradução automática só apareceram entre 1950 e 1960.

Em 1933, foram criadas duas patentes de sistemas de TA, uma do francês Georges Artsrouni, e uma de Petr Trojanski, na Rússia. Estes programas seriam, essencialmente, dicionários multilíngues que traduziriam palavras de uma língua de partida para a

correspondência do seu significado na língua de chegada. Curiosamente, esta patente russa viria a ser a arquitetura mais semelhante às atuais.

Warren Weaver, também considerado um dos grandes impulsionadores da TA, divulgou um memorando em 1949 que incentivou uma vaga de investigações na área da TA, não só nos Estados Unidos, mas a nível global. Incluiu objetivos e diversos métodos opostos à abordagem de criar um sistema que traduzisse palavra por palavra. Desta vaga, surgiu, em 1954, um sistema russo-inglês que despoletou entusiasmo e progressiva investigação. Esta demonstração de Georgetown com parceria da IBM, *International Business Machines Corporation*, uma empresa especializada em informática, serviu como uma apresentação aos meios de comunicação, alimentando progressivamente o entusiasmo desta nova área. De seguida, apareceram três principais focos de estudos de TA. No Reino Unido e nos Estados Unidos, financiados pela *CIA* e pelo exército no âmbito da situação geopolítica, enfatizava-se o desenvolvimento e o progresso de sistemas russo-inglês, enquanto na União Soviética eram explorados outros pares de línguas.

Em 1964 foi estabelecido o comité *ALPAC*, *Automatic Language Processing Advisory Committee,* que divulgou em 1966 o seu relatório. Este relatório atrasou o desenvolvimento desta área, devido aos resultados desencorajadores que apresentava, levando a um desinvestimento quase total na área durante uma década por parte dos centros de investigação estado-unidenses. Seria impossível discutir TA sem mencionar este relatório, que após tantos anos, continua a assombrar a área, sendo mencionado ocasionalmente a vinda de um "*second ALPAC report*".

John Hutchins (2004) documentou o surgimento desta nova área e tecnologia. Durante os primeiros anos de investigação, a TA viu períodos de grande entusiasmo e consequente aumento no volume de investigações, porém o oposto também se sucedeu. Críticas vindas de figuras mundialmente conhecidas faziam diminuir o entusiasmo pela nova área. As desvantagens e as óbvias fraquezas dos primeiros sistemas de TA eram suficientes para contrariar o entusiasmo. A comissão teve como abordagem determinar se a investigação e o desenvolvimento da TA conseguiria atingir o objetivo da área, a redução de custos e/ou o alcance de melhores resultados. O que nem sempre é do conhecimento geral é o facto de a comissão apenas ter tido em consideração as necessidades do Governo e do exército estado-unidense, ou seja, sistemas que não analisassem documentos de russo para inglês, ou que tivessem outros objetivos foram descartados. Existe um comentário no relatório da comissão sobre os principais problemas da Tradução (1966:16) que interessa explorar, para contextualizar o estado da Tradução em geral durante a era da Segunda Guerra Mundial:

There is no emergency in the field of translation. The problem is not to meet some nonexistent need through nonexistent machine translation. There are, however, several crucial problems of translation. These are quality, speed, and cost.

Por muito promissora que a TA fosse, eram identificados vários problemas com a trindade mais importante da TA - qualidade, velocidade e custo. A qualidade é o principal objetivo de todas as traduções, sejam elas feitas por humanos ou por máquinas. Contudo, na altura, os meios para medir a qualidade eram qualitativos e faltavam-lhes objetividade. Por esta razão, foram feitas experiências de avaliação para traduções feitas manualmente e por máquinas. A principal mensagem neste aspeto foi que a qualidade deve ser apropriada às exigências dos leitores, caso contrário, será um desperdício de recursos. Em termos de velocidade, havia imensos atrasos por parte dos sistemas de TA, sendo o período mais curto cerca de 15 dias para a tradução de 50 páginas, provando que havia muito para melhorar. Por último, relativamente ao custo, foi determinado que os Estados Unidos pagavam aos tradutores humanos valores compreendidos entre 9 a 66 dólares por cada 1000 palavras.

A conclusão do relatório foi que a tradução humana seria mais barata, visto que o custo da pós-edição apresentava uma fraqueza de TA neste aspeto. Nos casos em que não havia pós-edição, a comissão determinou que um leitor demoraria duas vezes mais para ler a tradução. A *ALPAC* também criticou a demonstração de Georgetown, apontando a necessidade de recorrer à pós-edição, tendo a mesma representado um gasto de tempo e custo superiores comparativamente aos gastos com uma tradução humana. O principal uso que a TA teria, aos olhos da *ALPAC*, seria, discutivelmente, como um agente auxiliar ao tradutor humano, sob a forma de ferramenta de tradução assistida por computador. Deste modo, o desenvolvimento da TA quase parou por completo nos Estados Unidos.

Porém, segundo Hutchins & Somers (1992:7), em 1976, foi criado o protótipo de um arquétipo novo, o sistema Météo (1982), pelo grupo científico TAUM, no Canadá. Devido às necessidades geopolíticas e sociais, havia uma procura de traduções para o par de línguas inglês-francês. Este sistema foi treinado para traduzir boletins meteorológicos para transmissões televisivas diárias e mostrou resultados bastante positivos. No continente europeu, a Comissão das Comunidades Europeias implementou um sistema com o mesmo par de línguas que o Météo, o Systran (1968). Posteriormente, foram adicionados novos pares de línguas, tendo esse novo sistema variantes criadas especificamente para organizações como a NATO. Hutchins & Somers (1992) mencionam o surgimento dos sistemas de TA comerciais como o maior desenvolvimento dos anos 80. Surgiram vários sistemas

desenvolvidos por empresas como a Mitsubishi, Toshiba e Siemens, que realçam as características da Tradução Automática - embora tenham índices de qualidade linguística fracos, são eficazes em termos de custo.

Antes de se falar nos diferentes tipos de sistemas de Tradução Automática, há que mencionar os modelos de sistemas baseados em regras linguísticas. Dorothy Kenny (2019) menciona como através destas regras, criadas manualmente, estes modelos podiam funcionar de diferentes formas. Entre essas formas observaram-se traduções "palavra a palavra" com base em dicionários, com a ajuda de regras relativamente à estrutura frásica, apenas respondendo aos problemas mais comuns e fáceis de resolver; ou através de sistemas *interlingua*, com o objetivo de ser independente da linguagem natural, embora se tenha determinado pouco prático e difícil de reproduzir em grande escala.

3.1.1 Tradução Automática Estatística

Vários autores, como Sánchez-Torrón (2017) Dorothy Kenny (2019), argumentam que a *Data-Driven Machine Translation*, começando pela arquitetura da Tradução Automática Estatística (SMT) surgiu, primeiramente, no memorando de Warren Weaver em 1949, que sugeriu a aplicação do trabalho criptográfico de Claude Shannon (1948) na TA. Visto que o memorando e as ideias de Weaver foram recebidos de diferentes formas, houve quem explorasse esta moção, como Kaplan (1950) e Reifler (1955), e houve quem discordasse com a mesma, como Bar-Hillel, contestando através do argumento do contexto ser válido apenas para "leitores inteligentes" e não para máquinas. Contudo, a inovação da tecnologia estatística só viria a ser alcançada nos anos 80-90 com os cinco modelos SMT de Brown (1990, 1993), cada um melhor do que o anterior, chamados "modelos IBM" 1 a 5. Pode-se argumentar que por detrás desta nova vaga de inovação pode ter sido devido à facilidade do acesso às ferramentas, tendo sido disponíveis nos finais da década de 90, e o progressivo desenvolvimento tecnológico, em poder computacional e capacidade de armazenar dados, assim como o crescente volume de dados produzidos graças à divulgação da Internet a nível mundial. Visto que os internautas queriam consumir conteúdos disponíveis noutras línguas além da sua, consequentemente, aumentou a procura por sistemas de TA. Por esse motivo, empresas internacionais, envolvidas na área da Tradução Automática, devido à sua forte ligação com o setor de apoio ao cliente, como a IBM, a empresa pioneira na área de investigação e seguida pela Google, Amazon e Facebook, começaram, sem exceção, a investir na investigação da área da TA. Por último, mas igualmente importante, os motivos deste investimento também se deram devido aos avanços em equipamento informático e

software de métricas de avaliação automática, *e.g. BLEU* (Papineni *et al*.) e *METEOR* (Lavie & Denkowskie, 2014), que incentivam a competitividade entre investigadores. Estas métricas fornecem um índice de qualidade, comparando o desempenho de TA com o de um humano. Quanto mais elevado for o índice, mais próximo está do desempenho de um humano.

Um sistema tradicional de SMT calcula modelos de probabilidade através de bitextos ou corpora paralelos. O texto monolíngue contém o modelo linguístico da língua de chegada e é, na sua essência, uma lista de palavras, ou sequência delas, com uma probabilidade associada às sequências de n-gramas. Este modelo linguístico fornece uma medida matemática da precisão de um determinado enunciado, e, desta forma, foca-se apenas na língua de chegada e não no processo de tradução. Este processo é do domínio de corpus bilíngue, ou seja, um modelo de tradução. Este modelo é constituído por tabelas em que palavras, ou frases, estão associadas a palavras, ou frases, de outra língua, em conjunto com uma probabilidade destes pares serem traduções mútuas. Se a tabela associa palavras da língua de partida a palavras da língua de chegada, é denominado *word-based* SMT, por exemplo os cinco modelos de Brown. Porém, se associar frases, denomina-se *phrase-based* SMT. De igual forma, denominam-se sistemas *syntax-based* SMT, os que associam também a informação sintática numa ou em ambas as línguas do par linguístico.

Deste modo, pode argumentar-se que a tradução em si decorre num processo de descodificação, em que é selecionada a opção com a maior probabilidade de ser uma tradução de qualidade. Posteriormente, foram desenvolvidos e acrescentados aos sistemas SMT descodificadores *log-linear* (Och e Ney, 2004), que providenciam mais modelos, tal é o caso de reordenamento de frases, que é extremamente útil para pares linguísticos com tipologias diferentes, e o caso da tipologia portuguesa "sujeito-verbo-objeto" *vs*. a irlandesa "verbo-sujeito-objeto".

As limitações da SMT *word-based* eram facilmente identificáveis. Sánchez-Torrón (2017:13) cita López (2008) que argumenta o facto de as palavras serem corretamente traduzidas, porém eram erradamente reorganizadas na frase da língua de chegada. Isto deve-se ao facto de o modelo linguístico considerar o contexto das palavras traduzidas, enquanto o modelo de tradução não considera o contexto, mas as probabilidades de ocorrência, demonstrando uma falha na compreensão do contexto local na língua de partida.

Por outro lado, os sistemas SMT *phrase-based*, em que as frases inteiras eram consideradas uma unidade de tradução, mostraram melhores resultados em comparação, isto porque, ao contrário dos SMT *word-based*, tanto o modelo linguístico, como o modelo de tradução identificam o contexto local das palavras, demonstrando melhores resultados no

contexto de expressões idiomáticas e de ambiguidades provenientes de palavras com múltiplos sentidos. Além disso, também possibilitaram a tradução de uma palavra para várias e vice-versa, como exemplificado em Sánchez-Torrón (2017:13), do inglês, *of course*, para o alemão *natürlich*. Contudo, este tipo de sistemas também exibiam falhas graves, tal como a incapacidade de lidar com relações de dependências distantes e com recursividades, e.g. "O João que adora assobiar tentou chamar a Ana que estava desatenta". Estas dificuldades poderiam, porém, ser resolvidas com sistemas *phrase-based* hierárquicos, que se podiam explicar na sua essência como frases traduzidas inseridas em frases traduzidas.

Além da Google, que utilizou esta arquitetura até 2015/2016 até passar para a arquitetura neuronal, um exemplo concreto de um sistema SMT encontra-se em Sánchez-Torrón (2017:16), em que se destaca o trabalho de Koehn (2005) que consistiu em criar 110 sistemas SMT com dados do Parlamento Europeu, e que obteve resultados que realçam as limitações desses mesmos sistemas: os sistemas que traduziram textos a partir do alemão tiveram melhores resultados do BLEU do que os que traduziram para o alemão. Curiosamente, o pior sistema SMT foi o que traduziu para finlandês, uma língua bastante complexa.

O terceiro tipo de sistema SMT, *syntax-based*, tinha como objetivo superar as dificuldades dos sistemas *phrase-based* descritas acima, focando-se na informação de estruturas árvores para a língua de partida, de chegada ou ambas. Uma vantagem direta deste sistema verificava-se no resultado final por este ser sintaticamente bem estruturado e reordenado, visto que era influenciado pelo contexto sintático. Contudo, identificavam-se outras dificuldades ao nível da ordem de palavras e a verbos frasais.

Embora os resultados estivessem a par com os SMT *phrase-based* e até demonstravam ser melhores em certos casos, segundo Koehn (2010a), este sistema rapidamente perdeu popularidade dentro da comunidade científica, devido ao surgimento dos sistemas de Tradução Automática Neuronal (NMT).

3.1.2 Tradução Automática Neuronal

A investigação do paradigma SMT, rapidamente, mudou de rumo para a arquitetura de Tradução Automática Neuronal (NMT), constituída por redes neuronais artificiais (ANN) e inspirada pelo funcionamento do cérebro humano, devido aos rápidos avanços informáticos. O mesmo foi feito pelas empresas mencionadas no ponto 3.1.1. Por exemplo, a Google, anunciou em setembro de 2016 o desenvolvimento do sistema *Google Neural Machine Translation* (Tradução Automática Neuronal da Google - tradução própria aproximada). A

NMT é considerada a causa que desencadeou a quarta revolução industrial, proporcionada pela Inteligência Artificial e pela importância dos dados, assim como a sua proteção.

Resumindo, um modelo neuronal consiste em várias redes artificiais, formadas por milhares de neurónios distribuídos em camadas que a informação percorre e são treinados através de um ajuste aos valores atribuídos a uma tradução. São essencialmente conjuntos de diferentes camadas de informação que processam um texto calculando, a cada camada, a opção com melhor qualidade.

O texto da língua de partida é inserido na camada superficial, a de entrada, e o texto da língua de chegada é extraído na camada de saída. Estas duas estão conectadas por um mecanismo de camadas escondidas, o que torna extremamente difícil os investigadores compreenderem a causa por detrás dos erros cometidos pela TA deste paradigma. Excluindo o neurónio inicial de entrada de informação, todos os neurónios recebem informação de outros neurónios da rede através de ligações que têm um peso que reflete cada valor atribuído pelos neurónios anteriores. Quando a soma destes valores atinge o limite de ativação de um neurónio, a sua informação é transmitida a outros neurónios. Desta forma, os sistemas NMT são mecanismos capazes de lidar com outros tipos de tradução, como reconhecimento de imagem e discurso, e.g. o modelo de *DeepHands* (*Koller et al.*, 2015) e *Avatar-based Translation from European Portuguese to Portuguese Sign Language* (*Coheur et al.*, 2015).

A vantagem das arquiteturas NMT sobre as SMT deve-se à maior facilidade em lidar com palavras para as quais o sistema não recebeu treino. Os sistemas NMT são capazes de explorar as semelhanças entre as palavras que conhecem e desconhecem, e, assim, são superiores aos modelos SMT, baseados em n-gramas, no que diz respeito a prever as sequências de palavras desconhecidas ao sistema, mas que contêm funções gramaticais semelhantes a outras que o sistema conhece. Tal como os sistemas SMT, os NMT requerem dados de entrada e de saída para serem treinados e desenvolvidos, bem como corpus, a fim de realizarem as traduções. Relativamente à realização da tarefa, o sistema NMT divide o conteúdo consoante a dimensão da frase do texto de partida e esta é primeiramente codificada por uma rede neuronal e posteriormente descodificada por outra rede neuronal, palavra a palavra. É importante mencionar que estes codificadores e descodificadores são treinados em conjunto para maximizar o índice de qualidade.

Pode ser argumentado que uma desvantagem deste processo é o tamanho da frase do texto de partida poder causar uma descida no índice de qualidade. Quão maior for uma frase, maior a possibilidade de o modelo confundir ou simplesmente errar algo ao longo da tradução. Com o objetivo de ultrapassar este obstáculo, acrescentou-se ao processo

codificador-descodificador um mecanismo de alinhamento (Bahdanau, Cho, & Bengio, 2015; Luong, Pham, & Manning, 2015) que causa a divisão de uma longa frase num conjunto de vetores, em vez de incorporar toda a informação num único vetor. Estas diferenças no processo de tradução são o que distinguem os sistemas SMT dos sistemas NMT, e, consequentemente, as vantagens e desvantagens de cada um. A fim de obter índices de qualidade aceitáveis, é necessário estar constantemente a treinar os sistemas. Com esta perspetiva, para cada treino de um sistema SMT *phrase-based*, é inevitável treinar os pesos dos valores atribuídos de várias componentes (as mesmas treinadas separadamente) o corpus linguístico, nomeadamente o de tradução, e os modelos de reordenamento de frases. Por contraste, os sistemas NMT são treinados num único modelo, em que o processo se baseia na codificação e na descodificação de palavras. Contudo, as desvantagens dos sistemas NMT assentam na enorme dificuldade em compreender a origem dos erros cometidos pelo sistema, sendo menos transparente relativamente aos sistemas SMT, e devido ao facto de os mesmos não conseguirem assimilar grandes quantidades de glossários e vocabulários, o que provoca uma restrição relativamente ao número de ocorrências de determinada palavra. Adicionalmente, os sistemas NMT têm uma desvantagem em comum com os sistemas SMT: ambos têm dificuldades com línguas complexas, por exemplo, o finlandês, e outras que contenham palavras compostas compridas, como é o caso da língua alemã. É importante destacar o facto de que os recursos necessários para treinar um sistema NMT são dispendiosos e demorados, sendo que a uma unidade de processamento gráfico (*GPU-graphic processing unit*) topo-de-gama demora pode demorar semanas a treinar um sistema NMT, comparativamente aos poucos dias que se demora a treinar um sistema SMT, dependendo da quantidade de dados utilizados.

3.2 As variedades do português europeu e português do Brasil

Nesta seção serão mencionadas algumas das principais diferenças entre as variedades brasileira e portuguesa, como a questão das formas de tratamento, que diferem não só entre os países, como dentro do território do Brasil e português. Como acontece em qualquer estudo de uma determinada língua, ou variedades, é necessário fazer um enquadramento histórico e social. Dentro de um país, existem variedades linguísticas. Relativamente a Portugal, as diferenças regionais não têm as mesmas dimensões e são, geralmente, questões de pronúncia de certas palavras. Por outro lado, devido ao enorme território geográfico, o Brasil possui uma quantidade de diferenças gramaticais, sociais e culturais que dividem, de

certa forma, o país. Desta forma, é complexo encontrar uma versão da variedade brasileira padrão que se possa comparar com a versão portuguesa padrão.

3.2.1 Breve resenha histórica sobre as diferenças entre português europeu e português do Brasil

Embora haja várias diferenças relevantes entre as variedades, a ferramenta *Smartcheck* não possui atualmente a capacidade para poder ajudar os editores a distingui-las melhor. Por esta razão, nesta seção, haverá ênfase nas formas de tratamento, na posição e utilização dos pronomes clíticos, determinantes e possessivos, visto que são as áreas onde o impacto da ferramenta é significativo.

Antes de contrastar as diferenças entre as duas variedades, é de mencionar a situação linguística atual do Brasil. Pode ser argumentado que existem duas formas bastante distintas de falar português do Brasil. Sem poder ignorar a situação socioeconómica, o grau de instrução da população pode variar substancialmente, resultando em duas variantes de PB. O PB que é falado nas ruas em contexto social ou entre amigos e desconhecidos, em que a distinção entre formas de tratamento representa diferentes níveis de formalidade. E a variante acentuadamente formal que Duarte e Serra (2015) consideram ser a "gramática da escola" ensinada no Brasil. Esta é considerada por alguns falantes como mais "estrangeira", e, possivelmente mais complexa do que a gramática portuguesa, mencionando um distanciamento entre o que é falado contemporaneamente e o que é escrito em gramáticas tradicionais, baseadas ainda no modelo europeu. Por esta razão, existem vários constrangimentos sobre a escrita e a língua na variedade brasileira, tendo em conta que uma língua padrão se entende como a variedade escrita encontrada em jornais, revistas, trabalhos académicos, ou seja, por alguém que esteja em constante contato com a escrita da língua e a faça circular. Esta "gramática formal" distancia-se significativamente da língua realmente falada pela maioria dos falantes, de tal forma que os jornais se veem obrigados a criar manuais de redação para os seus contribuidores, a fim de poderem orientar a sua escrita. Duarte e Serra (2015) mencionam igualmente o facto de esta gramática ser demasiado tradicional e arcaica, sendo fortemente baseada num modelo gramatical demasiado concentrado na escrita europeia e vista como sendo normativa. Uma gramática deve ser, na sua essência, uma gramática descritiva, uma representação de como os seus falantes realizam a sua língua. Edição após edição, as inconsistências entre conceitos e a ausência de atualização de exemplos continuam a acentuar essa diferença entre a gramática prescritiva e tradicional, adotada pelas elites brasileiras, e a gramática descritiva, efetivamente usada pela

maior parte da população, dificultando a aprendizagem das gerações mais novas. Estas afirmações são partilhadas pelos falantes nativos dentro da *Unbabel* com vasta experiência linguística que também contribuíram para este projeto.

As comparações entre as variedades PE e PB são sustentadas em Martins e Kato (2016a) e são apresentadas algumas das divergências entre as duas variedades. Além de reconhecerem questões de vocabulário e expressões idiomáticas, é estabelecido o marco que define as evoluções separadas que as variedades atravessaram ao longo da sua gramática histórica. Com base no trabalho de Tarallo (1993), o século XIX é apontado como o começo das diferenças entre estas variedades ao nível da escrita. Em Lopes e Cavalcante (2011), é evidenciada a entrada do novo pronome, *você*, no sistema pronominal do Brasil. Por volta dos anos 30, começou-se a usar os dois pronomes em peças de teatro. Porém, tal foi feito acidentalmente, visto que os autores tinham a intenção de recorrer a ambas as formas, mesmo tendo como objetivo manter a distinção entre relações simétricas e assimétricas. Ao longo da evolução do sistema linguístico brasileiro, a conclusão é a total mistura entre os pronomes, incluindo a mistura entre a 2ª pessoa sob a forma do pronome *tu,* e a 3ª pessoa, recorrendo ao pronome *você*.

Um exemplo histórico destas mudanças é exemplificado em Nascimento *et al.* (2018) em que as autoras dão destaque às diferenças dos sistemas de formas de tratamento entre as variedades PE e PB. Começando com a acentuada presença no sistema linguístico de *Vossa Mercê* que, em pleno século XIV, tinha uma tonalidade de reverência. Posteriormente, a forma de tratamento ao se dirigir ao rei tomou a forma de *Vós*. Dada tal substituição, a forma *Vossa Mercê* adquire um estatuto de respeito inferior, mas começa a ser utilizada de forma mais livre entre as camadas restantes da sociedade. Ao longo dos séculos, outras formas surgiram como *Vossa Majestade, Vossa Alteza* e outras variantes mais desviadas da expressão *Vossa Mercê* do século XIV começaram a manifestar-se. Entre estas, encontram-se as formas de tratamento *vosmecê, mecêa, vosse,* e a forma que sobreviveu no nosso sistema linguístico, *você*. Esta forma nominal em específico terá o seu surgimento e utilização entre as variedades PE e PB mais aprofundadas posteriormente.

Ao longo da investigação deste projeto, foram analisados aspetos no que diz respeito à divergência entre variedades, sendo expostas mais adiante, incluindo as que não puderam ser exploradas através da ferramenta *Smartcheck*. Entre estes aspectos, serão mencionados com mais importância as formas de tratamento e a combinação, ou omissão, do artigo definido quando este precede um pronome possessivo. Embora não tenham sido contemplados neste estudo, os aspetos linguísticos que não puderam ser considerados para a utilização da

ferramenta, mas que serão brevemente mencionados são erros de estrutura frásica ou de reordenamento, assim como os diversos casos de uso da forma *se*. Do mesmo modo, embora haja diferenças no uso dos clíticos, tal não é evidente nos erros dos editores dentro dos conjuntos de dados. A escolha dos aspectos apontados prende-se com com os diversos exemplos analisados durante a fase de investigação, os quais mostram, como se irá descrever no capítulo 5, níveis de inconsistência vários dos editores quanto à utilização ou omissão do artigo definido quando este precede um pronome possessivo. Estas questões relembram as afirmações em Duarte e Serra (2015): o facto de a variedade do PB em regime formal parecer outra língua aos olhos de um falante nativo comum.

3.2.2 Sistema de formas de tratamento

Tal como descrito por Nascimento (2020): Tradicionalmente, e num sentido estrito, entende-se por formas de tratamento o conjunto das formas (palavras e expressões) que o falante usa para se dirigir ao ouvinte (ou ouvintes) num ato de comunicação linguística, ou seja, para referir o(s) destinatário(s) ou recetor(es) do seu discurso. São exemplos dessas formas os pronomes de 2.ª pessoa tu, vós e você e expressões nominais como o senhor ou Vossa Excelência, que codifica as atitudes que os falantes têm para com os destinatários do seu discurso, em particular no que respeita às relações sociais e/ou familiares que com eles mantêm.

O PE possui um sistema de formas de tratamento bastante complexo comparativamente a outras línguas europeias, visto que tem diversas opções consoante as situações de comunicação, orais ou escritas, ou ainda fatores de natureza social (Cintra, 1972:7). No terceiro volume da Gramática do Português da Fundação Calouste Gulbenkian (2020:2701), Nascimento expõe as características do sistema português como sendo um sistema de formas de tratamento tripartido. Em primeiro lugar, formas de tratamento pronominais, considerando este um sistema fechado e equilibrado devido à predominância de poucas formas nominais estabelecidas na língua portuguesa: *eu, tu, você.* Segundo*,* formas de tratamento nominais, significativamente mais diverso e rico, o que leva a ser um sistema mais "instável" do que o pronominal. Constituído por nomes de parentesco desde *pai, mãe, tio, avó, etc.*, mas também expressões como *o senhor, a senhora*, que são designadas em Nascimento (2020:2720) como formas nominais de convivência de carácter geral; a utilização de nomes das profissões como *o professor, o senhor doutor, senhor engenheiro, etc.*, e a combinação de nomes precedidos de artigos definidos como *a Ana*. Por último, o sistema de formas de tratamento verbais, tendo em conta que sendo a língua portuguesa uma

língua de sujeito nulo, é bastante comum a não realização do sujeito em orações finitas, restando recorrer à forma verbal para preencher essa informação na frase através de pessoa e número: (*tu*) *compreendes, (você) pode repetir?*.

Relativamente à seleção da forma de tratamento, é dependente de vários fatores. Primeiramente, depende das preferências pessoais do falante, determinadas por tradições de grupos sociais, familiares, regionais e por desenvolvimento pessoal. Segundamente, a relação de intimidade ou familiaridade com o ouvinte e as relações socioculturais da região, assim como o contexto em que a ação discursiva ocorre decidem o nível de formalidade. Outros fatores essenciais são a hierarquia social e profissional das entidades envolvidas no discurso, falantes e ouvintes, como um ato discursivo entre um aluno e o seu professor. Nascimento *et al.* (2018) descreve estas várias formas de tratamento do PE como acompanhadas de informação sobre os intervenientes no discurso ao nível do grau de relacionamento, por outras palavras, a dêixis social.

Como em todas as línguas *vivas*, existe uma constante evolução, adaptabilidade e inovação linguística ao longo do tempo. Algumas formas de tratamento vão caindo em desuso como a tradição ou maneirismo de tratar o pai e a mãe pela forma nominal *você* ou conjugação verbal na terceira pessoa em caso da não realização do sujeito. Porém, surgem outras formas como *puto* e *mano*, mas, de igual modo, torna-se cada vez mais recorrente em caso de utilização dos pronomes possessivos *meu* e *minha* como novas formas nominais por parte dos falantes mais jovens da língua, os agentes mais prováveis a inovarem a língua. Outro exemplo da constante evolução da língua e o nível de variação que deriva dela é o facto de, atualmente, as camadas mais jovens começarem a preferir a utilização de estrangeirismos como *dude* e *bro*, devido à atual dominância da língua inglesa nas diversas redes sociais. Mesmo que não tenha sido um estrangeirismo, o caso do pronome pessoal *stor* começar a ser utilizado como a forma de tratamento preferencial por parte dos jovens quando se dirigiam a um professor, especialmente no ensino secundário. Segundo o *site Infopédia*, a palavra tem origem na amálgama de *se(nhor) + (dou)tor*. Contudo, destacando que, embora seja um amálgama de formas de tratamento formais, a forma resultante é considerada informal/coloquial e parte da gíria académica que os professores tendem a não aprovar.

Relativamente às formas de tratamento pronominais, Nascimento *et al.* (2020) expõe o facto de os elementos combinados de formas pronominais de pessoa, número e género que referem os participantes de um ato discursivo formarem uma série pronominal. Cada um destes elementos serve um propósito informativo na oração. Consoante a sua função sintática, como sujeito ou como um dos diversos géneros de complementos, adquirem uma

nomenclatura específica em relação com a função que desempenham. Para o elemento que realiza a função sintática de sujeito, designa-se uma forma nominativa, para as formas pronominais que desempenham as funções de complemento direto e indireto, são identificadas como acusativas e dativas, respetivamente. Para a forma que realiza o complemento de preposição, existe a designação de oblíqua e genitiva, esta última perante uma forma que realiza a função de complemento do nome.

As diferentes formas de tratamento serão agora expandidas mais detalhadamente devido aos contrastes culturais relativamente ao processo discursivo entre as variedades PE e PB. Observando o contraste com que as formas nominais *você* e outras semelhantes como *vosmecê, mecêa* e *vosse* entraram nos sistemas linguísticos de PE e PB. Nascimento (2018:247) delimita o século XIX como o ponto de partida para ambas as variedades, porém tem uma facilidade a estabelecer-se no sistema linguístico do português do Brasil relativamente ao sistema europeu, que já tinha um pronome anteriormente bastante estabelecido na língua, *tu*, que manteve a sua popularidade comparativamente a esta nova forma que se apresenta como uma alternativa completamente diferente. Do outro lado do mundo, a entrada do pronome *você* no sistema linguístico do Brasil já competia com o pronome *tu*, trazido pelos colonizadores e entre os anos 30 e 50, *você* já era uma forma de tratamento comum na maior parte do território, negando à forma nominal *tu* a mesma popularidade que possui em Portugal.

3.2.2.1. Formas de tratamento no sistema PE

Nascimento (2018) identifica, atualmente em Portugal, várias estratégias que podem ser utilizadas ao iniciar um ato discursivo. Sendo a primeira a utilização de um dos pronomes de segunda pessoa do singular anteriormente mencionados, *tu* e *você*, posteriormente selecionando a alternativa que melhor se insere no contexto e nos fatores socioculturais do momento, como já foram mencionados anteriormente;

A segunda estratégia traduz-se na não realização do sujeito na oração ou durante o discurso, aproveitando o facto do português ser, de facto, uma língua de sujeito nulo e recorrendo aos elementos restantes, como forma a forma verbal, para indicar as informações necessárias;

A terceira alternativa aproveita um aspeto identificado na variedade PB, a utilização do pronome de 2ª pessoa do singular em junção com um verbo na terceira pessoa ou recorrendo a uma das expressões mencionadas previamente, *i.e. você/o senhor deseja mais alguma coisa?*;

A última estratégia mencionada por Nascimento (2018) remete para a possibilidade de aproveitar o pronome *vós*, embora tenha caído em desuso e provoque mais estranheza do que o pronome *você*.

Em teoria, a seleção da forma de tratamento na 2ª pessoa do singular mais apropriada parece depender de outros fatores externos à própria língua. Não mencionando o papel nominativo que a forma *tu* partilha com a forma de tratamento *você*, a forma *tu* pode, igualmente, adquirir o estatuto de complemento direto ou indireto através do pronome acusativo, *te*, por exemplo, *liguei-te ontem à noite*; *convidei-te para ir beber um café*. Do ponto de vista social, já foi mencionado que o pronome *você* pode chegar a ofender o recetor da mensagem. Tal verifica-se especialmente quando este integra uma faixa etária mais velha e parece ter mais incidência quando o falante é mais jovem. Estes, ao interagirem com familiares, amigos, ou recém conhecidos e até desconhecidos, costumam recorrer à forma de tratamento *tu,* se o recetor não tiver uma diferença de idade significativa, ou se o recetor tiver um estatuto social superior, como uma relação entre aluno e o seu professor em que a regra geral é a utilização da terceira estratégia mencionada por Nascimento (2018). Recorrer a uma expressão como pronome nominal, neste caso a profissão como forma pronominal ou aplicar a segunda estratégia e preferir não realizar o sujeito: *(professora), pode repetir?*. Tal não se verificava em gerações anteriores que se dirigiam aos seus pais e a entidades com estatuto social superior através da forma *você*. Esta "tradição" tem vindo a cair desuso, embora ainda esteja presente atualmente.

Por outro lado, o pronome *você*, estabelecido anteriormente como a variante evoluída da forma de tratamento nominal *vossa mercê*, que tinha como recetoras as entidades num pináculo do estatuto social, como a alta nobreza e realeza, herdou o facto de ser a correspondência na 3ª pessoa. É algo que no contexto da língua portuguesa se revela paradoxal, visto que é uma forma de tratamento que "compete" diretamente com a forma nominal de 2ª pessoa, *tu*. Tornando-se uma forma pronominal de 2ª pessoa ao nível semântico e uma forma pronominal de 3ª pessoa ao nível gramatical. Outra situação a realçar é a série pronominal que o pronome *você* ancora. Sendo a forma nominativa o *você*, ela não adquire outras funções sintáticas. Estas são identificadas pelas formas *o*, *a* e *lhe* para os pronomes acusativos e dativos, respetivamente: "Se você quiser saber o que realmente aconteceu, convido-o para almoçar e conto-lhe tudo"

Como já foi mencionado, esta forma de tratamento pode ser considerada uma forma respeitosa, mas também uma forma ofensiva. Para evitar a segunda possibilidade de interpretação e, de igual modo, a utilização da forma nominal *tu*, começou a ser preferida a

não realização do sujeito e a realização de expressões como alternativas. Porém, há outra parte da população, menos instruída, e jovem, que recorre à utilização da forma de tratamento *você*. Seja devido à falta de instrução, devido ao contato com a cultura brasileira (através de redes sociais, por exemplo) ou por outros motivos. Esta tendência de generalizar a utilização desta forma pode, de facto, contribuir para uma normalização da forma de tratamento, perdendo a possível interpretação ofensiva, e ser considerada apenas como uma alternativa mais formal ao pronome *tu*.

Como foi referido anteriormente, têm surgido alternativas à utilização das formas de tratamento *tu* e *você*. Entre estas realça-se o emprego das formas nominais como *o senhor, a senhora, a dona,* entre outras, em que, normalmente, são antecedidos por um artigo definido que atua em concordância da forma empregue. As adoções destas formas de tratamento possuem outras variantes com um tom de formalidade elevado, utilizado para situações de prestação de serviço, por exemplo: *os meninos, as meninas*. Igualmente, podem conter um pronome possessivo inserido entre o artigo definido e a forma nominal *a minha senhora*. O facto deste artigo definido ser realizado em conjunto com o nome próprio do recetor providencia um certo nível de formalidade e até de distanciamento. Estas formas nominais de convivência de carácter geral têm a sua origem, possivelmente, na população detentora de menor estatuto social. É importante mencionar que nem todas estas formas de tratamento de convivência geral são utilizadas em relações de menor para maior estatuto social. As alternativas *os meus amigos, as minhas amigas* são algumas das formas de tratamento de carácter geral que representam um tom de familiaridade. Contudo, um elemento a destacar é a polaridade entre as formas plural e singular. As formas *o amigo, a amiga*, são representadas e interpretadas como uma relação entre estranhos ou semelhantes ao nível do caráter linguístico das formas *(o) colega(s), a(s) colega(s)*. Uma forma nominal que representa a amizade entre indivíduos também é utilizada em relações mais distantes do ponto de vista da intimidade. Outra justificação para recorrer a esta forma de tratamento com a interpretação de intimidade distante pode ser, precisamente, para fabricar uma falsa sensação de proximidade entre estranhos e, assim, facilitar o processo de comunicação. Igualmente exposto, foi a alternativa de se dirigir ao ouvinte pela profissão e esta também pode ser realizada com a presença de um artigo definido a preceder à forma nominal como, por exemplo, *O doutor, pode receitar alguma coisa para as dores?*. Acima destas formas de tratamento, existem as de maior formalidade, exemplificadas em Nascimento (2020:2725), que ilustram uma entidade com claras diferenças em termos de estatuto social e podendo até designar a sua profissão:

Meritíssimo (senhor) Juiz, *Excelentíssimo (senhor) Reitor*. Estas formas também podem incluir a integração da forma nominal *senhor,* dependendo do orador.

Ao longo desta seção foi, principalmente, discutida a utilização de *tu vs. você*, contudo existe ainda uma outra opção que só é realmente utilizada por crentes na religião cristã. A decrescente utilização da forma de tratamento *vós*, encontrada maioritariamente no Norte do país ou em discurso religioso, sendo o recetor Deus, no caso da religião cristã, mas caiu completamente em desuso em todas as outras ocorrências de discurso. Sendo que, em Portugal, a principal religião é o Cristianismo, existem algumas formas de tratamento especiais e geralmente específicas a entidades de grande importância entre a comunidade cristã. Nascimento (2020:2726) destaca formas nominais na 3ª pessoa, como *Sua Excelência, Sua Santidade, Sua Eminência*. Estas formas de tratamento ou menções a Deus são a única instância em português europeu em que é grafado o pronome possessivo.

Num outro espetro do sistema linguístico PE, observamos formas de tratamento depreciativas com o intuito de insultar o recetor através da equivalência ou comparação do recetor animais como *camelo, burro,* entre outros, e a decadência das formas *gajo* e *tipo*.

3.2.2.2. Formas de tratamento no sistema PB

Relativamente ao PB, tal como defendido por Nascimento (2020), Podemos afirmar que, no PB, o uso normal das duas formas [tu e você] não sugere distribuição complementar, ou seja, não há distinção de grau de cortesia ou de maior ou menor familiaridade nas localidades em que ambas as formas se encontram em variação. Apenas ocasionalmente encontramos relatos de omissão do pronome (presumivelmente um você nulo, mas possivelmente um o senhor/a senhora nulos) quando o falante, em situação assimétrica, não sabe como se dirigir a um interlocutor mais velho ou pouco conhecido. Tudo indica, porém, que entre os mais jovens já não se encontra essa estratégia, tal é a frequência de um pronome expresso de segunda pessoa, particularmente em contexto inicial.

Foi partilhado em Nascimento (2018:255), relativamente às formas de tratamento no português do Brasil que, no caso PB, a utilização de *tu* e *você* não são distribuídos igualmente no território brasileiro. Destaca-se o trabalho de Scherre *et al.*(2015) que consistiu na recolha de dados sobre formas de tratamento, concretamente este paradigma de *tu vs. você*, no Brasil. Umas das conclusões foi a surpreendente preferência de uso geral do pronome *tu*. Contudo, mesmo com a utilização do pronome *tu*, determinaram-se diversos conjuntos de preferências como a "presença de *tu* com concordância, em graus variados, motivada pelo contexto de mais formalidade ou pelo aumento da escolarização, especialmente onde *tu* é reconhecido

como de uso mais natural na comunidade local". Abaixo estão representadas as conclusões dos autores da recolha e culminam num total de seis subsistemas diferentes.

(i) O uso exclusivo de *você* e as formas reduzidas de *cê* e *ocê*, em Minas Gerais. Este subsistema é observado em várias regiões que pertenciam à capitania de São Paulo, criada em 1709;

(ii) Preferência do pronome *tu* com a taxa de preferência de 60% e uma taxa com concordância inferior a 10% (i.e. *tu não vai viajar para Portugal?*) e pouca utilização da forma pronominal *você*. Encontrado no norte (Amazonas) e sul (Rio Grande do Sul);

(iii) Preferência do pronome *tu* com a taxa de 60% e a taxa de concordância entre 40 e 60%, e o pronome *você* com pouca relevância. Utilizado em Pará e Santa Catarina, norte e sul respetivamente;

(iv) Igual distribuição na taxa de preferência dos pronomes pessoais *tu* e *você*, em que a utilização de *tu* é demonstra preferência ao pronome, mas com uma taxa de concordância inferior a 10%. É falado nas regiões nordeste (Maranhão) e sul (Santa Catarina);

(v) Igual distribuição nas taxas de utilização, sem predomínio de uma forma sob a outra e com uma percentagem ao nível da concordância a rondar os 10 a 40%. Mais observado nas regiões do norte e nordeste.

(vi) Variação entre os pronomes *você* e *tu*, havendo predominância da forma *você* sem concordância. Um subsistema detetado de norte a sul como em Distrito Federal, Rio de Janeiro, São Paulo, Minas Gerais, Paraná e Maranhão.

Estas conclusões provam que não existe um sistema apenas com uma das formas de tratamento. Normalmente, a opinião geral é a ideia do uso frequente do pronome *você*. Não só se verificou que nenhum dos pronomes substitui completamente o outro, como se identificaram zonas com uso quase definitivo de *tu,* como vai ser explorado adiante, tal é explicado pela supremacia da língua falada sob a "língua das gramáticas".

Já foi mencionado ao pormenor a polaridade entre os pronomes pessoais *tu vs. você*, contudo é importante realçar que a principal diferença entre os sistemas de formas de tratamento é o facto de os falantes brasileiros não se dirigirem ao ouvinte mencionando o seu nome e, o que provavelmente poderá ser mais surpreendente, o facto de não usarem os títulos de profissão como *professor* ou *doutor*. Como alternativa, um aluno dirigir-se-ia ao seu professor recorrendo à forma nominal *senhor*, como uma forma de cortesia, um símbolo de respeito e diferença entre os estatutos sociais dos intervenientes no discurso. Tendo o símbolo de respeito e posição na sociedade, os falantes brasileiros ao interagirem com personalidades de renome, pessoas desconhecidas, ou perante alguém que esteja num patamar

socioeconómico superior, recorrem às formas *o senhor, a senhora*. Ao nível familiar, as gerações mais velhas mantêm a utilização das formas *o senhor, a senhora*, como claro sinal de respeito perante a hierarquia familiar. Contudo, esta mentalidade muda nas regiões urbanas, havendo um relaxamento desta mentalidade de exercer sempre o máximo respeito no sentido em que os jovens tratam os seus familiares por um dos dois pronomes, *tu* e *você*.

O sistema dos pronomes pessoais na variedade brasileira é representado na Tabela 2 e comparado com o sistema PE em Nascimento (2020:2736-2739), sendo destacadas diferenças entre as duas variedades. Entre as observações feitas, notou-se apenas um aspeto em comum entre os sistemas. O facto de ambas as variedades recorrerem às mesmas formas ao mencionarem a 1ª pessoa do singular. Por outro lado, ocorreu a extinção das formas *vós* e variantes da mesma como *convosco*, normalmente recorrentes na 2ª pessoa do plural, bem como as formas *si* e a variante *consigo*, tendo o seu desuso sido derivado pela preferência à forma *vocês*. Ainda no tópico de extinção de formas, as formas *nós/nos* utilizadas para enunciar a 1ª pessoa do plural estão, de igual modo, a cair em desuso. Sendo estas apenas utilizadas nas comunidades rurais e a serem substituídas por *a gente* pela população com um estatuto social mais elevado.

	Formas Tónicas			Formas Átonas (clíticas)		
	Suj (nominativo	CD (acusativo)	CP (oblíquo)	CD (acusativo)	CI (dativo)	Passivo/ Nominativo
1sg	eu	eu	mim, comigo	me	me	
1pl	nós a gente	nós, a gente	nós, connosco a gente	nos	nos	
2sg	tu você	tu você	ti, contigo, você, ~~si, consigo~~		te lhe	
2pl	~~vós~~ vocês	vocês	~~vós, convosco~~ vocês	~~vos~~ os, as, se	~~vos~~ lhes	
3sg	ele, ela	ele, ela	ele, ela ~~si, consigo~~	o, a, se	lhe	se
3pl	eles, elas	eles, elas	eles, elas	os, as, se	lhes	

Tabela 2: Comparação de sistemas PE e PB por Nascimento (2020: 2736), quadro extraído de *Gramática do Português vol. III*

As formas pronominais nominativas registaram uma redução no paradigma flexional verbal, algo que em PE enfrenta mais resistência devido à maior restrição gramatical. O melhor exemplo desta situação é o mesmo paradigma presente no sistema PE, a preferência de utilização do *tu vs. você*. Já foi mencionado anteriormente a divergência de usos entre o pronome de 2ª pessoa do singular mais popularmente utilizado, contudo, também há uma competição semelhante relativamente à 2ª pessoa do plural. Como se pode observar no quadro, há uma tendência para substituir o pronome *nós* pela expressão nominal *a gente*. A forma nominal *nós* é mantida em orações cujo pronome seja precedido ou seguido de um quantificador numeral (*dois de nós/ nós dois*). Esta substituição vai contribuir para a redução da necessidade de flexionar o verbo da oração que contenha *a gente*. A redução flexional dentro da variedade PB é fortemente representada nestas disputas entre a representação ou omissão de concordância relativamente às 2ª e 3ª formas do plural.

É relevante mencionar brevemente que parece estar a surgir um erro quando se discute as diferenças entre as variedades portuguesas. É um facto que as orações adverbiais gerundivas estão fortemente presentes no discurso brasileiro. Contudo, acontece o mesmo em PE. Em ambas as variedades, estas orações podem ter um sujeito expresso ou não realizado, visto que a língua portuguesa permite tal opção. A única diferença entre as variedades neste aspeto encontra-se na posição do sujeito relativamente ao verbo. Na escrita portuguesa, a posição do sujeito é constantemente pós-verbal, enquanto na escrita brasileira o mesmo não se verifica, sendo o sujeito, normalmente, encontrado em posição pré-verbal.

Como já descrito, especialmente na variedade PB, as diferenças dentro da mesma variedade são a forma escrita *vs.* a forma oral. Estas baseiam-se na escolarização inspirada na gramática lusitana que tende a recuperar algo que na variedade PE poderá até já estar em completo desuso. Esta tendência claramente conservadora acaba por resultar num entrave à evolução normal de uma língua, cuja consequência culmina na existência de duas gramáticas brasileiras diferentes, a falada "moderna e real" e a escrita "conservadora e aportuguesada". Retomando a questão de qual gramática utilizar em artigos jornalísticos, segundo Nascimento (2020:2750), há uma preferência a sujeitos pronominais expressos na 3ª pessoa em jornais brasileiros, cerca de 56% nas entrevistas transcritas e 51% nos textos jornalísticos. No outro lado do oceano Atlântico, na variedade PE apenas 22% das entrevistas transcritas e 7% em textos jornalísticos contêm o sujeito de 3ª pessoa expresso. Outra característica a destacar neste género de texto na variedade PB é o facto da evolução da língua parecer estar a incluir a retoma do sujeito, quer ele já esteja mencionado na mesma oração numa anterior. Contrastando com o PE, que apresenta clara preferência na não realização do sujeito, exceto

quando o pronome expresso e o seu antecedente realizam funções gramaticais diferentes (*o Salvador viu o cão a saltar bem alto, mas ele nunca irá voar*), ou circunstâncias altamente específicas, como um antecedente ser o sujeito em posição pós-verbal numa oração participial (*Embrulhadas as prendas, elas já podem ser distribuídas*).

Relativamente à taxa de utilização de cada pronome, em entrevistas, os pronomes *você* e (*a*) *gente* representam a maioria dos dados. Já em textos jornalísticos, há uma grande taxa de utilização da 1ª forma do plural não realizado, seguido do clítico *se* e 3ª pessoa do plural, *eles, elas*. Outras formas começaram a ser detetadas neste género, ao nível de textos mais informais, sendo observadas 13 ocorrências de *a gente* e 7 de *você*. Por comparação, é seguro afirmar que a variedade PE prefere o clítico *se* com uma taxa de utilização de 69% e o sujeito não realizado de 1ª e 3ª pessoa do plural com 27 e 4%, respetivamente. A forma *se*, por si só, representa uma diferença importante entre estas variedades. No PB, este pronome só é presente em contextos sob a exclusiva função de preposição. Já em PE, tem também a possibilidade de realizar a função de sujeito, estando, desta forma, presente em orações não preposicionais.

3.2.4 Determinante e possessivo

Durante a investigação e a fase inicial de criação de regras através da análise de um conjunto de dados de traduções de inglês para PB, de cariz formal, verificou-se uma enorme inconsistência na ocorrência de determinantes com possessivos. Embora esteja correta a utilização do artigo definido antes do possessivo, a realidade presente nos conjuntos de dados utilizados durante este projeto é que é a sua utilização, ou omissão, foi puramente opcional.

Embora não haja nenhuma regra assente sobre esta questão em gramáticas, é, de facto, uma questão que remonta, novamente, à questão da "gramática falada e a gramática culta" mencionada anteriormente. Segundo os falantes nativos consultados durante a investigação e fase de planeamento deste projeto, é considerado prática comum omitir o artigo definido quando este precede um pronome possessivo. Foi mencionado que, embora não se considere errada a presença, ou omissão, do artigo, é uma combinação utilizada em cenários de ênfase ou quando o falante pertence a um estrato social elevado e conversador. Esta temática será retomada posteriormente no contexto de anotação e pós-edição.

3.3 Ferramentas de identificação de erros

Algumas destas ferramentas são universalmente conhecidas e utilizadas, como o corretor ortográfico do *Microsoft Word* e o *Google Translate*. Contudo, têm vindo a surgir

outras aplicações que apoiam a escrita, dos quais a *Language Tool*, um projeto disponível desde 2003, e mais recentemente, em 2012, o *Grammarly*.

A ferramenta Language Tool é capaz de suportar mais de 30 línguas e baseia-se em padrões de erros e regras linguísticas, para ajudar o escritor a melhorar os seus textos e facilitar a sua compreensão.

Relativamente ao *Grammarly,* a aplicação só fornece apoio para a língua inglesa, mas reconhece as diferenças entre as variedades americanas, britânicas, canadianas e australianas, e é um serviço que afirma poder identificar 250 tipos de erros diferentes, desde erros ortográficos, gramaticais e de estilo. Além do reconhecimento das variedades inglesas, também oferece indicações de vocabulário mais correto para o contexto através da deteção da tonalidade do texto. De modo a diminuir a ausência de linguagem corporal e modulação vocal em mensagens escritas e permitir um maior grau de atenção à forma como o escritor estrutura a sua mensagem e como ela será possivelmente entendida do ponto de vista dos leitores.

Utilizando estes programas de identificação de erros disponíveis gratuitamente, os editores, enquanto trabalham na plataforma da *Unbabel*, têm acesso também ao corretor desenvolvido pela *Unbabel*, o *Smartcheck*. O *Language Tool* é semelhante ao *Smartcheck* na medida em que ambos são corretores multilíngues baseados em regras linguísticas. Estas ferramentas fornecem assistência preciosa durante o processo de pós-edição.

3.4 Pós-edição e TA

Segundo Comparin & Mendes (2017: 1), a área da Tradução Automática tem sido considerada bastante relevante nas últimas décadas. Com a crescente globalização, esta área científica veio a ter um papel principal nos mercados. Por um lado, tendo em conta a variação na qualidade, tornou-se aparente a necessidade de juntar sistemas de tradução automática à pós-edição, a fim de obter tradução de alta qualidade. Por outro lado, a pós-edição, sendo um processo realizado por um humano, torna-se, naturalmente, numa fase mais demorada e cara.

Visto que a TA se tornou uma das áreas de investigação mais importantes, subproduto do surgimento da área da Inteligência Artificial, como referido anteriormente, empresas internacionais como a Google, a Amazon, e o Facebook necessitam de um sistema de tradução automática no âmbito dos seus modelos empresariais, fornecer grandes volumes de traduções em relativamente pouco tempo. Por esta razão, dependem da qualidade das suas traduções. Contudo, embora tenham sido alcançados progressos espantosos na área da tradução automática, como a noção e desenvolvimento de sistemas complexos como NMT, os

resultados, do ponto de vista da qualidade, são inconsistentes, seja consoante o modelo usado, o facto de ser SMT ou NMT, ou o par de línguas envolvido. Por esta razão, é necessário haver um processo adicional que atenue esta inconsistência. Atualmente, o processo de pós-edição é realizado através de agentes humanos e, contrariamente à fácil acessibilidade e eficiência das ferramentas de tradução assistida por computador, a inversão dos papéis, em que o humano é o agente secundário no processo da tradução, acrescenta mais custos e resulta num processo mais demorado, a fim de obter índices de alta qualidade.

Hutchins (1997:8) menciona a investigação de Yehoshua Bar-Hillel (1951:1-2), matemático e linguista isrealita pioneiro na área da computação da TA e linguística formal que compreendia as limitações da TA e tornou óbvio o facto que a TA sem qualquer intervenção humana só é possível através da redução da qualidade. Ficou estabelecido pelos cientistas e investigadores que a pós-edição é um passo vital e importante para a TA:

It seems obvious that fully automatic MT, i.e. one without human intervention between putting the foreign text into the reading organ of the mechanical translator and reading off its output, ins achievable only at the price of inaccuracy. (...) It appears that a post-editor is indispensável for elimination of semantical ambiguities.

Com o crescente desenvolvimento e progressivo sucesso da TA, principalmente dos sistemas de TA mais recentes e com resultados encorajadores dos sistemas de *Data-driven Machine Translation*, ou seja, a SMT e, principalmente, a NMT para ser mais concreto, seria de esperar que a pós-edição teria de acompanhar esse crescimento. Consiste num processo de correção dos resultados finais das TA, realizado por humanos, mais concretamente, tradutores profissionais. No fim de cada texto produzido por um sistema de TA, se não obter os requisitos de qualidade necessários, terá de ser submetido para pós-edição. Esta tarefa visa corrigir os erros que foram detetados e também fornece feedback para melhorar o sistema. Como já foi referido, do ponto de vista da qualidade das traduções, a pós-edição feita por humanos tem sido bastante benéfica e até crucial para a TA. O processo que, embora seja executado exclusivamente pela máquina, tem de ter sempre interação humana, seja sob a forma de fornecimento de dados para o treino dos sistemas ou através de pré e/ou pós-edição. Com estes dois aspetos em consideração, a avaliação da qualidade serve para controlo de qualidade e identificação de erros, e, igualmente, contribui com *feedback* no que diz respeito a melhorias a implementar nos sistemas. No contexto da *Unbabel*, a pós-edição é feita em regime de *crowdsource*, através de uma comunidade de centenas de editores, que trabalham com diversos segmentos de texto diferentes, distribuídos aleatoriamente.

Outro aspeto da pós-edição traduz-se no facto de que combate as fraquezas da TA e permite, no caso da *Unbabel*, uma melhor compreensão dos problemas que a TA teve com um determinado texto, que possibilita compreender melhor os "mistérios" dos sistemas neuronais usados atualmente. Adicionalmente, contribui com outro género de *feedback* para a melhoria dos sistemas utilizados através do fornecimento de mais dados para alimentar os modelos, como por exemplo, expressões idiomáticas, registos muitos informais, a chamada *slang language*. Este *feedback* constitui informação adicional que os sistemas poderão não ter recebido durante o seu treino.

Complementarmente, além de avaliar a eficácia do sistema, existe, de igual forma, a necessidade de prever a dificuldade ou o esforço necessário durante a fase de pós-edição. Mediante esta previsão, os investigadores e criadores dos sistemas podem melhorar o desempenho dos modelos com respetivas adaptações em pré-edição de futuros textos. O resultado desta avaliação denomina-se Estimativa de Qualidade (QE). Suplementarmente, esta avaliação do esforço contribui para determinar até que ponto compensa recorrer ao processo de pós-edição, ao invés de traduzir de raiz. Na maioria dos casos, a pós-edição da TA pode ser mais eficaz do que a tradução de raiz, assumindo que têm o nível de qualidade desejado pelos investigadores. O problema surge na inconsistência das TA. Os níveis de inconsistência variam, possivelmente, devido ao vocabulário e à sua complexidade, consoante o tipo de texto (*FAQ*, *tickets*, *chat*) e a qualidade do texto de partida, mas igualmente dos dados inseridos para treinar o sistema. Estes aspetos são todos possíveis áreas de estudo e investigação dentro da qualidade de tradução.

Contudo, o potencial demonstrado foi aplicado de forma diferente. Em vez de ser a máquina o agente principal e o tradutor humano o assistente, o humano foi o agente principal, sob a forma de ferramentas de tradução assistida por computador (*CAT Tools*), que incluem memórias de tradução, glossários, entre outros. Atualmente, pode-se argumentar que esta é a melhor combinação entre TA e tradutor humano e é um tema ainda muito explorado. Contudo, o processo de pós-edição humano não elimina todos os erros, sendo essencial a existência e desenvolvimento de sistemas de identificação de erros, neste caso o *Smartcheck*, sob a forma de regras *Surf* como iremos observar no estudo-piloto presente no capítulo 4.

Para mais referências, existem, do mesmo modo, investigações relativamente à pós-edição automática, tendo sido discutida a *Automated Post-Editing and Quality Estimation (APE-QUEST)*, Depraetere *et al.* (2020) no "17th Machine Translation Summit". Este modelo consiste em treinar um sistema com dados da língua de chegada, ou seja, um sistema monolíngue, que no fundo pode ser considerado bilíngue no argumento em que é um

sistema de TA capaz de traduzir “mau português” para “bom português” que consiga produzir textos que sejam bem aceites na língua de chegada de forma coerente e com o objetivo de eliminar a sensação de “estranheza”. Está a ser desenvolvido no domínio muito específico de resultados desportivos e a obter resultados promissores.

4. Metodologia

Neste capítulo serão clarificados os procedimentos e estratégias adotadas ao longo do projeto. Incluindo a descrição do estudo-piloto que serviu de base para uma melhor compreensão e preparação para o desenvolvimento de regras no *Smartcheck*.

Explorando as capacidades da ferramenta desenvolvida internamente na *Unbabel*, o *Smartcheck*, verifica-se que é semelhante à ferramenta *Language Tool* na medida em que ambos são corretores multilingues baseados em regras linguísticas, fornecendo assistência preciosa durante o processo de pós-edição. O *Smartcheck* é uma ferramenta que engloba diversos corretores ortográficos e outras funcionalidades como regras tipográficas e regras linguísticas, *Surf*, produzidas e mantidas na plataforma *Surfboard*. O foco desta investigação passa pela criação destas regras linguísticas *Surf*, que permitem a criação de regras específicas por língua ou até por cliente, e estas regras irão realçar determinadas palavras ou expressões, como um corretor ortográfico na interface dos editores. Estes, por sua vez, enquanto trabalham na plataforma da *Unbabel*, têm acesso não só aos programas de identificação de erros disponíveis, como *Grammarly* e/ou *Language Tool*, mas também ao *Smartcheck*.

Ao longo deste capítulo iremos analisar o impacto das caraterísticas que distinguem as duas variedades. O PE exibe clássicos sinais de uma língua de sujeito nulo, enquanto o PB, com o seu enorme potencial de variação, principalmente ao nível da língua falada, demonstra sinais mistos. Mencionando brevemente outro levantamento de diferenças ao nível do vocabulário e/ou culturais descobertas ao longo deste projeto e algumas referidas posteriormente: *senha vs. palavra-passe* e *guardar vs. salvar*, *mascarar vs. fantasiar*, *rapariga vs. moça*.

Uma enorme vantagem da ferramenta *Smartcheck* prova ser a capacidade de poder detetar erros em aberturas de *emails*. Adicionado ao facto de que as traduções podem ser melhoradas neste campo, foi tomada uma especial atenção às formas de tratamento em ambas as variedades.

Nascimento (2020:2735) menciona como a escrita-padrão, aprendida através do processo de escolarização, inclui fortes marcas da gramática portuguesa, levando a uma forte distinção entre a gramática da fala e a da escrita. Embora o PE tenha, de facto, um sistema linguístico algo restrito e com pouca taxa de variação linguística quando comparado com o Brasil, o mesmo não se verifica no sistema linguístico do Brasil. Tal poderá dever-se à proximidade geográfica de outra língua gramaticalmente fluída, o inglês, e, desta forma, até os sistemas de identificação de erros têm dificuldades em detetar todas as possibilidades de

ocorrências de erros, por exemplo, relativamente ao género das entidades referidas por *you*. Falantes nativos destacaram este facto em que a língua brasileira não tem uma "linha" que delimita o que é gramatical e o que é agramatical tão bem definida como a variedade europeia.

Antes de começar a investigação, foi proposta uma oportunidade de analisar um problema que estava a mostrar-se frequente no sistema de TA com o par de línguas inglês-holandês (EN-NL). Este problema tinha sido reportado por um cliente que necessitava de ser respondido o mais depressa possível e consistia no uso incorreto da forma verbal *moet/moeten* (v. dever), uma vez que era considerada agressiva para os leitores, visto que as mensagens eram, essencialmente, de aconselhamento. Desta forma, foi criada uma regra *Surf*, presente na figura 8, que pretendia reduzir a taxa de ocorrência da citação *moet/moeten* e das suas realizações em dados de mensagens pós-editadas. O objetivo da análise foi centrado em confirmar se tal redução se verificou num conjunto de dados constituído por 1050 *tickets* e o impacto da regra proposta e aplicada no *Smartcheck*. A experiência permitiu testar frequências em casos reais, e obter experiência, em primeira mão da plataforma *Surfboard*, onde são criadas as regras a serem aplicadas pelo *Smartcheck*.

```
rule STYLE_NL {
    category: modal,
    description: 'If a translation of must please make sure it does not sound rude',
    severity: major
}
# modal
=> {
   any_word(/[Mm]oet/, /[Mm]oeten/)
}
```

Figura 8: Regra *Surf* sugerida para implementar no *Smartcheck* em produção para sinalizar a palavra moet/moeten em NL

Num primeiro passo, foram comparados os textos produzidos pelo sistema TA e os pós-editados, e, posteriormente, foi documentada a frequência das seguintes palavras e combinações de palavras presentes em ambos. Houve a necessidade de ter em conta os afixos *u*, *jullie* e *zou*, pois *u* corresponde à segunda pessoa do plural formal, você, enquanto que *jullie* e *zou* são marcas de pluralidade de sujeito.

Moet, moet u, u moet, moeten, moeten u, u moeten, moeten jullie, zou moeten.

Contrariamente ao estudo-piloto, o projeto final não pôde ser realizado de forma automatizada, visto que os problemas não tinham sido identificados, exigindo trabalho manual intensivo. Porém, à luz da metodologia do estudo-piloto EN_NL, foram elaborados e disponibilizados dois conjuntos de dados de traduções automáticas e pós-editadas de correio eletrónico, ou seja, *tickets* de suporte com os pares de línguas inglês-português europeu (283 *tickets* no total) e inglês-português do Brasil (301 *tickets* no total), e, tendo em conta a relevância dos dados e o tema em questão, bem como as restrições de tempo, foi decidido trabalhar apenas com os dados de cariz formal. Posteriormente, estes conjuntos foram divididos em dois: um de desenvolvimento constituídos por 138 mensagens em PE e 114 mensagens em PB, utilizado para detetar os erros, cometidos pelos editores, mais comuns e urgentes que pudessem ser resolvidos pelo sistema de identificação de erros; outro para testar as regras criadas, composto por 115 mensagens PE e 187 PB.

Seguindo os guias desenvolvidos pela *Unbabel*, os erros são atribuídos a uma de três categorias possíveis: estilo (do inglês *style*), fluência (do inglês f*luency*) e precisão (do inglês *accuracy*), que agrupam diversas subcategorias, como se pode observar na tabela 4.

Segundamente, foram criadas 25 regras *Surf* para o português europeu e 20 para o português do Brasil, com a sua distribuição exemplificada na tabela 3, tendo em consideração os guias linguísticos da empresa, das gramáticas e de falantes nativos com vasta experiência na área da linguística. Numa terceira e última fase, foi usado o segundo conjunto de dados para testar as regras e foi calculado o seu índice de qualidade, antes e depois de as regras serem aplicadas, o que, hipoteticamente, melhoraria o índice de qualidade, assumindo que os editores aceitam as sugestões do sistema.

	PE	**PB**
Estilo	12	9
Precisão	4	5
Fluência	9	6

Tabela 3: Distribuição das regras *Surf* criadas no âmbito da investigação

Estilo	Fluência	Precisão
Adição	Codificação de caracteres	Registo Gramático
Omissão	Gramática	Registo Lexical
Erro de Tradução	Tipografia	Variedade Linguística Errada
Erro de Tradução	Inconsistência	
Halucinação da TA	Inteligibilidade	

Tabela 4: Tipologia de erros implementada na *Unbabel*

O índice de qualidade é afetado por cada erro com diferentes níveis de impacto e consoante o número de palavras numa determinada mensagem. Segundo o guia de anotação criado pela empresa, erros de pontuação, por exemplo, são considerados menores, por outras palavras, não causam perda de significado nem confusão ao leitor. Enquanto os erros graves, como erros de variedade linguística, são mais relevantes, por poderem levar o leitor a uma leitura não completa ou alterarem partes do conteúdo da mensagem, resultando num desconto mais significativo e, por último, os erros críticos, por exemplo o caso detetado no conjunto de dados de português do Brasil, "Querida (nome)," em que a recetora da mensagem pode sentir-se ofendida, resulta num erro com a máxima severidade, um erro crítico. Outras consequências deste tipo de erro, sob a forma de alteração do significado da língua de partida na língua de chegada ou alucinações da TA, são os possíveis danos da reputação da empresa ou a representação errada do funcionamento de determinado produto ou serviço.

A tarefa inicial foi a de identificar, manualmente, todos os erros de variedade linguística, ou seja, sinais de contaminação de português do Brasil em textos de português europeu e vice-versa, seja na TA, por possível uso de dados errados durante o treino ou durante a pós-edição, pois tinha havido relatórios internos de editores de uma destas variedades a inserir sinais de contaminação durante a pós-edição, que resultaria num aumento significativo deste tipo de erro. Posteriormente, verificou-se que tal não era o caso e não foram identificados volumes suficientes para justificar as conjeturas iniciais. Por esta razão, tomou-se a decisão de explorar todos os tipos de erros que fossem identificados, por exemplo, falhas morfológicas ou agramaticalidades. Tal foi feito através de dados recolhidos do mesmo cliente durante o mesmo espaço de tempo. Com o objetivo de também criar regras para o português do Brasil em mente, foram consultados regularmente dois falantes nativos com vasta experiência linguística.

5. Resultados

Nesta seção vão ser apresentados os resultados das várias tarefas realizadas ao longo do estágio. Numa primeira parte, serão descritos os impactos que a regra linguística para mensagens de inglês para holandês teve. Na segunda e terceira observação será comentado os efeitos que as regras linguísticas criadas para PE e PB tiveram no conjunto de dados.

Numa primeira análise superficial, observam-se na tabela 5 os resultados dessa fase de ambas as variedades em termos de qualidade. As taxas de erros por cada mil palavras e pode ser argumentado que a variedade PB tem o dobro dos erros da variedade europeia, à exceção dos erros de gravidade crítica. Em 96 mensagens de PE foram detetados 190 erros e 328 no PB em 168 mensagens. Estes resultados podem ser ilustrados de outra forma, por exemplo, em cada mil palavras, encontram-se 24,13 erros no caso PE e 41,09 no caso PB. Perante estes factos podemos assumir que a qualidade, relativamente ao cliente ao qual os dados foram extraídos, é superior na variedade PE do que na PB.

	PE	PB
% de mensagens com erro(s)	66,2%	89,8%
Erros por mil palavras	12,23	20,55
Erros críticos por mil palavras	0	0,17
Erros graves por mil palavras	10,79	16,56
Erros menores por mil palavras	1,44	3,81

Tabela 5: Representação dos erros por mil palavras

Em termos de comparação de qualidade segundos os parâmetros seguidos pela *Unbabel*, os resultados exibidos na tabela 6 demonstram que o índice de qualidade dos textos de inglês para PE subiu mais de 5 pontos e as de inglês para PB mais de 11 pontos, valores assaz promissores. É importante mencionar que os valores mínimos não são 100 nas colunas que correspondem aos índices de qualidade com as regras aplicadas porque foi decidido incluir alguns erros que o *Smartcheck* não consegue resolver, por agora. Estes erros são, por vezes, igualmente descartados e, assim, pretende-se dar visibilidade ao impacto que estes erros têm nos textos. Estes erros são, por exemplo, de localização, posição errada do símbolo

monetário ou pontuação, sendo de destacar o especial impacto que tiveram numa mensagem PB em que baixou até 93,22, onde o valor mínimo expectável é cerca de 98. Por outro lado, é possível observar nos valores mínimos dos conjuntos antes das regras serem aplicadas que o impacto destas regras consegue equivaler a melhorias que chegam a 45 e 60 pontos, valores assaz promissores em todos os parâmetros.

	PE (sem regras)	PE (com regras)	PB (sem regras)	PB (com regras)
Média de qualidade	94,23	99,97	88,65	99,94
Mediana de qualidade	95,95	100	90,57	100
Desvio-padrão	6,93	0,2	9,35	0,52
Valor máximo	100	100	100	100
Valor mínimo	54,55	98,15	40	93,22

Tabela 6: Resultados em valores de MQM antes e depois da aplicação das regras

5.1 Análise do Estudo-piloto

Primeiramente, há que observar o impacto que a regra EN_NL teve em *email* recolhidos de um só cliente no mesmo espaço de tempo, e serviu para adquirir experiência para a análise dos conjuntos PE e PB.

A análise revelou que as instruções da regra foram eficazes a diferentes níveis. Foram observadas descidas diferentes nas taxas de ocorrência das palavras *moet* e *moeten*, com 232 em TA para 140 (8,7%) em textos pós-editados e 194 para 182 (1,15%) respetivamente. Relativamente às combinações, a utilização de *moet u* desceu de 108 ocorrências em TA para 62 (4,38%) em texto pós-editado, e da mesma forma a combinação *u moet* teve uma redução de ocorrência de 32 em TA para 22 em texto pós-editado (0,9%). Contudo, as combinações *moeten jullie*, *jullie moeten*, *moeten u* e *u moeten* não justificaram a implementação de um aviso devido a falta de relevância nos dados. Por último, *zou moeten*, verificou-se uma descida pouco significativa, de 30 em TA para 28 em pós-editado (0,2%). Não foi possível contabilizar cada palavra ou combinação de palavras presentes nos dados. O sistema conseguiu apenas verificar a presença ou omissão das mesmas. Apesar destes obstáculos, a análise verificou que todas as combinações de palavras detetadas obtiveram descidas na taxa

de utilização após a implementação da regra na plataforma, sendo substituídas por sinónimos ou simplesmente omitidas com diferentes graus de sucesso.

5.2 Análise e Resultados do conjunto de dados inglês-português europeu

Antes de descrever os resultados e a posterior análise dos mesmos, é de referir que ao longo da fase em que a eficácia das regras criadas foi testada, foram encontrados novos erros no conjunto de dados utilizado para testar a validação das regras *Surf* criadas. Porém, estes erros não foram tidos em consideração para a criação de regras posteriores. Foram identificados e registados para um possível trabalho futuro.

Na figura 10 está representada a amostra onde as regras foram testadas e pode-se observar que em 115 mensagens, foram encontrados pelo menos um erro. Por sua vez, a figura 11 revela a categorização de cada erro identificado. Embora a quantidade de mensagens com erros possa parecer grave, é necessário ter em consideração a severidade de cada erro que foi discutido previamente e agir em conformidade. Relativamente à distribuição dos erros pelas categorias, os erros de estilo são os que constituem a maior percentagem. Foram detetados 152 erros de estilo, mas somente 27 erros de precisão e 11 de fluência. Em termos de severidade, podemos presumir que a maior parte dos erros de estilo foram classificados como graves, visto que engloba erros identificados como registo de língua errados. Relativamente aos erros de precisão, também é constituído por erros principalmente de severidade grave, englobando erros como palavras não traduzidas ou uso de falsos amigos. Por último, os erros de fluência são principalmente de menor severidade, sendo na sua maioria pontuação e capitalização indevida.

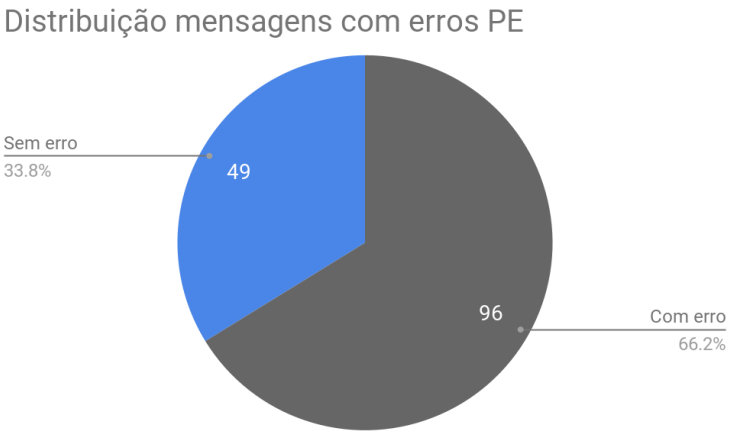

Figura 10: Representação de mensagens de português europeu com pelo menos um erro

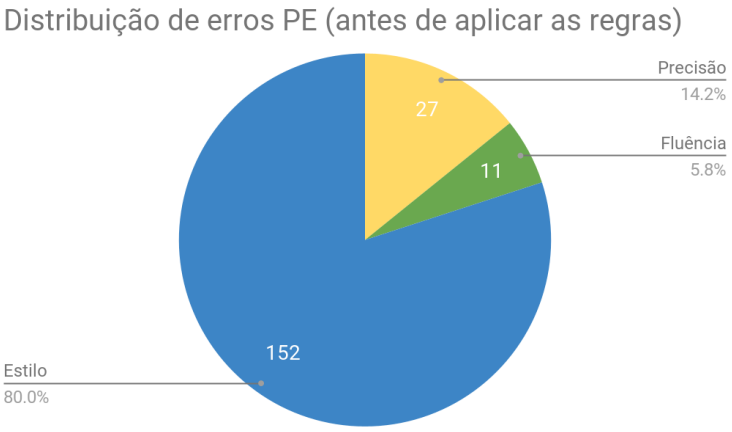

Figura 11: Distribuição de erros no conjunto de dados PE

A figura 12, por sua vez, demonstra a variação de erros de português europeu e a esmagadora maioria dos erros de estilo serem erros identificados como registos de língua inapropriados. Entre estes erros, foi identificado um demasiado prevalente, fácil de corrigir e evitar, sob a forma de "Olá", encontrado 132 vezes e categorizado como grave.

Foi verificado que a única regra presente na plataforma relativamente a este erro, era a sinalização para acrescentar uma vírgula após "Olá" ou "Oi". Por outras palavras, verificou-se que algumas das regras previamente criadas não distinguem entre registos formais e informais. Esta regra pode ser bastante útil para manter a consistência no caráter formal presente, e exigido, nas mensagens. Além de tentar evitar a utilização de "Olá", é vital impedir o uso da palavra "Oi", visto que esta pode ser considerada extremamente informal e passar um significado de desrespeito para com a pessoa a quem a mensagem seja dirigida. No entanto, os erros de registo lexical não foram identificados apenas por "Olá/Oi". Também houve registo de ocorrências da abreviatura "foto" que pode contribuir para uma tonalidade mais informal num contexto de comunicação tipicamente formal. Considerando que foi assumido que as regras seriam criadas para serem aplicadas em *emails* de cariz formal, foi desenvolvida uma regra que sugerisse a palavra na sua íntegra, "fotografia".

Com muito menos prevalência, o segundo e terceiro tipos de erros identificaram-se sob a forma de 13 problemas de entidades mencionadas e 11 ocorrências de palavras ou expressões não traduzidas. Sendo o principal objetivo desta ferramenta o de contribuir para uma melhor qualidade dos textos, acrescenta-se de igual forma, o auxílio a programas como os reconhecedores de entidades (*Named Entity Recognizers*). Um exemplo comum de erro de entidade mencionada e de expressão não traduzida recolhido do conjunto de dados é o caso

FCA, sigla de *Financial Conduct Authority*, uma instituição reguladora do Reino Unido. Para resolver este problema, foi criada uma regra que sinalizasse tanto a sigla como a expressão a fim de ambas serem evitadas e substituídas por "Autoridade de Conduta Financeira do Reino Unido". Relativamente a erros de palavras ou expressões não traduzidas, o exemplo mais recorrente foi a deteção da palavra "app". Pode-se argumentar que esta palavra pode não ser um erro sendo usada em português sobre certas circunstâncias, como conversa com ou entre jovens ou que trabalhem com tecnologia, por exemplo. Porém havendo uma tradução exata na língua portuguesa, e podendo dar uma tonalidade mais casual e informal à mensagem, como a palavra "foto" mencionada previamente, decidiu-se criar uma regra *Surf* que sugerisse a forma portuguesa "aplicação".

De igual modo registaram-se 10 ocorrências de erros de variedade linguística, como "retorno de chamada" e "contato". Mas, como já foi referido, não são o tipo de erro mais prevalente nem, discutivelmente, o mais urgente, pelo menos, tendo em conta a amostra desta investigação. A solução passa por sinalizar a expressão "retorno de chamada" e sugerir "chamada de retorno". Relativamente à palavra "contato"[3], foi determinado que não seria prático criar um regra linguística para cada erro ortográfico, especialmente um que não afeta a leitura da palavra de forma significativa. Adicionalmente, é bastante provável que a solução para este erro seja dada por um outro corretor ortográfico que não o *Smartcheck*.

Relativamente aos erros de menor gravidade, há a mencionar, ao nível da pontuação, a presença de 7 erros por omissão de vírgula nas aberturas e despedidas das mensagens, 3 erros devido à capitalização de dias da semana, meses, línguas, bem como 3 ocorrências de falsos amigos, devido à tradução de "preferred" para "preferido" e "preferida", ao invés de "preferível". Considerando que estes erros pertencem à categoria de falsos amigos, podem passar despercebidos pelos corretores ortográficos e prejudicar a compreensão da mensagem. Desta forma, criou-se a regra para identificar ambas as formas "preferido" e "preferida" e sugerir a alternativa correta. Os 3 problemas relativamente à hifenização estão relacionados com o mesmo género de erro: a omissão do hífen da palavra composta por justaposição "pós-inscrição". Durante o processo de criação da regra linguística que confrontasse esta situação e fornecesse uma alternativa de adicionar um hífen entre "pós" e "inscrição", observou-se que o alcance da regra em questão podia ser expandida. A regra resultante após desta teoria culminou na sua ativação perante a deteção das palavras "pós", "pré" e "pró", sugerindo o acrescento de um hífen após a palavra identificada.

[3] Nota que pode ser discutido que a palavra "contato" também podia ser categorizada como um erro de simples ortografia, considerando o Acordo ortográfico de 1990.

Por último, foram registados 3 problemas relativamente a unidades monetárias. Estas ocorrências decorreram com a posição agramatical do símbolo "€" e "£" e devido à inconsistência da sua posição ao longo do conjunto de dados que também envolve a omissão de um espaço entre o símbolo e o número mencionado. As circunstâncias deste tipo de erros podem causar problemas graves ao nível da compreensão da mensagem e, naturalmente, sendo o tópico de natureza monetária, a qualidade e fidelidade da tradução é vital. Contudo, como se pode observar, neste conjunto de dados, a presença de erros menores não tiveram um volume significativo de ocorrências.

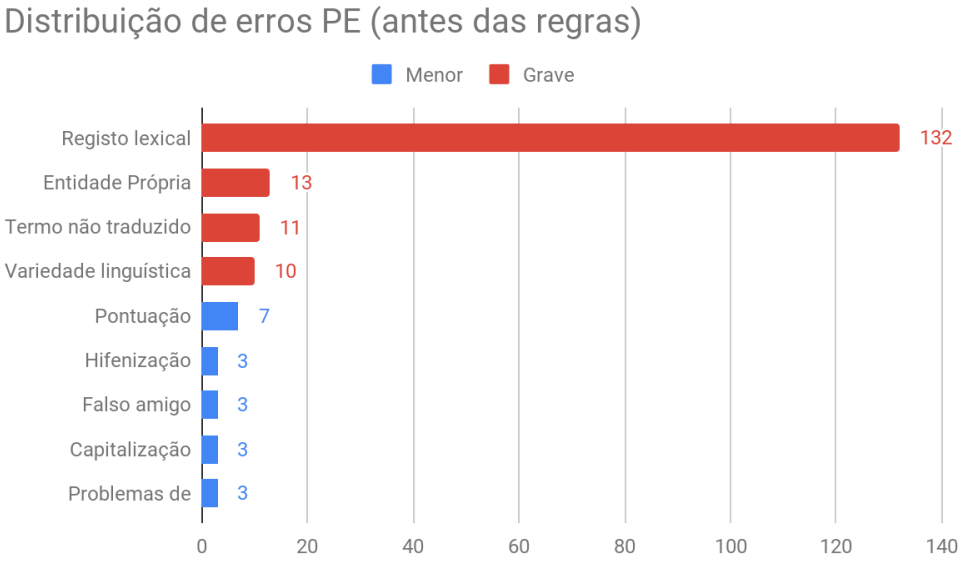

Figura 12: Distribuição de erros PE[4]

5.3 Análise e Resultados do conjunto de dados inglês-português do Brasil

No conjunto de dados de português do Brasil, foi aplicado o mesmo método que ao conjunto PE. Os novos erros descobertos no conjunto de dados utilizado para testar a regras criadas não foram tidos em consideração para este estudo, mas identificados para trabalho futuro.

Observando a figura 13, em comparação com a figura 10, é registado um aumento significativo de mensagens que continham pelo menos um erro. Ainda que se considere a ligeira diferença do tamanho das amostras, não justifica o desvio entre as duas figuras. Comparando de igual forma os resultados de ambas as variedades relativamente à

[4] Os gráficos das figuras 12 e 15 isolaram a palavra "Pontuação". O erro em questão é "Problemas de unidades monetárias".

distribuição dos erros identificados, na figura 14 pode-se verificar alguma semelhança à figura 11. É de destacar o facto que erros de estilo são bastante prevalentes em ambas as variedades, contudo a segunda categoria de erro mais identificada no conjunto de dados PB foi a fluência ao invés de precisão. Esta mudança entre as variedades poderá ser justificada por um aumento significativo na quantidade de erros de pontuação, normalmente em falta, e pela deteção de erros de concordância, não identificados no conjunto de dados PE.

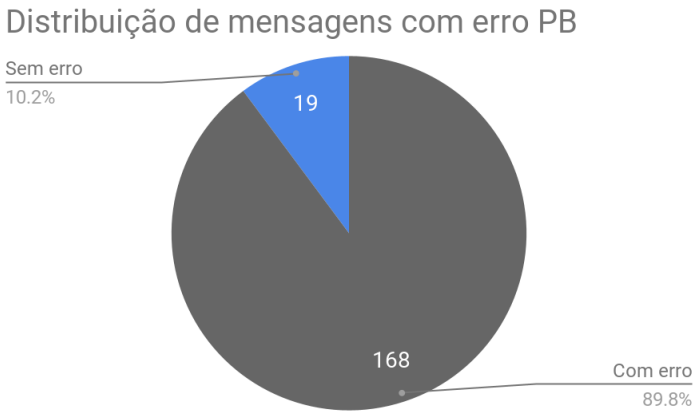

Figura 13: Representação de mensagens de português do Brasil com pelo menos um erro

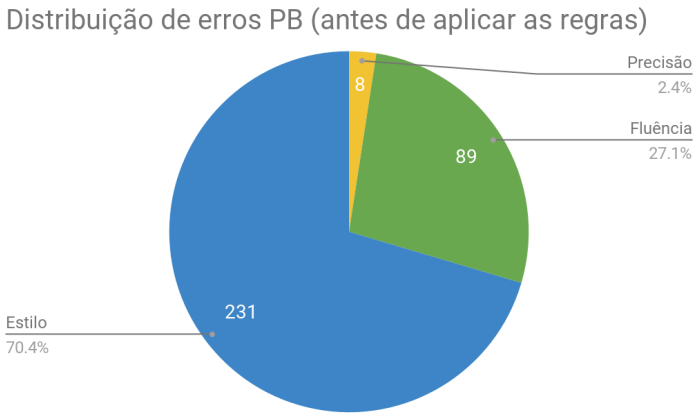

Figura 14: Distribuição de erros no conjunto de dados PB

Relativamente às respostas criadas para confrontar estes erros exibidos na figura 15, o principal tipo de erro, variedade linguística errada, poderá ter sido, até certo ponto, provocado pela investigação. Tal aconteceu durante a fase inicial de anotação manual e criação de regras no conjunto de dados para desenvolver as regras, foi detetada uma variação significativa entre

a utilização, e omissão, de artigos definidos quando estes precedem pronomes possessivos. Por esta razão, decidiu-se elaborar uma regra *Surf* para tentar diminuir a taxa de inconsistência relativamente a esta questão. Porém, não é o único tipo de ocorrência de erros de variedade linguística errada, visto que foram identificadas diversas vezes a utilização do verbo "guardar", que não é utilizado pela variedade brasileira, sendo normalmente preferido o verbo "salvar". Do mesmo modo, mas não tão relevante, criou-se uma regra linguística que destacasse a expressão "sinto (muito)", visto que, segundo os falantes nativos consultados, embora não seja errado como o caso do verbo "guardar", a expressão "lamento (muito)" é considerada a mais preferível pela população em geral. Ainda dentro da tipologia de erros de estilo, detetou um erro derivado da sigla AAE, Área Econômica Europeia, que causou alguma confusão perante os falantes brasileiros nativos que foram consultados. Acharam a expressão estranha e após ser investigada a relevância da sigla nas formas de comunicação brasileiras, como a Folha de São Paulo, por exemplo, chegou-se à conclusão de sugerir a alternativa de EEE, Espaço Econômico Europeu.

Avançando para outro tipo de erro, este também com mais de 100 ocorrências, e em comum com a amostra PE, os erros de registo lexical são prevalentes, partilhando, desta forma, a mesma regra que a variedade de português europeu. Foi discutido previamente a presença das 3 ocorrências de erros críticos no capítulo 4 devido ao uso da forma "Querida [nome]", numa mensagem de cariz formal. Considerando o cariz formal que mensagens trocadas por *email* normalmente exigem, a utilização deste pronome deve ser evitado por poder causar um sentimento de ofensa ao recetor. A solução para este dilema passou pela criação de uma regra *Surf* que sinalizasse o pronome "Querida" e sugerir na mesma a expressão "Caro" seguido do nome próprio da pessoa ou por "cliente".

Um caso a comentar e único ao conjunto de dados PB são os erros de formas verbais, neste caso, erros de concordância. Foram identificadas quase 50 ocorrências de casos em que o nome da entidade era do género feminino, mas era precedido com um artigo masculino. Foram detetados em situações semelhantes, onde um nome de origem britânica de uma empresa pode ter levado os editores a erro, a não ser que os editores tivessem pesquisado por um *website* da empresa em questão.

Comentando os erros menores, verifica-se uma presença muito mais significativa de erros de pontuação, identificados em aberturas e despedidas das mensagens e, de igual modo erros de localização, relativamente a unidades monetárias. Além de as aberturas e despedidas das mensagens, as diversas ocorrências de erros de pontuação derivam também de valores monetários que não foram bem localizados (por exemplo 1.000 vs 1,000) e de mensagens de

abertura fora do cariz formal como “Olá Sílvia! Obrigado,” que, neste exemplo, seriam identificados dois erros, devido à posição da vírgula e na utilização do ponto de exclamação devido ao tipo de comunicação e formalidade. De igual modo, há um aumento dos problemas envolvendo símbolos monetários e a justificação para destacar os erros de localização, mais concretamente, de unidades monetárias. Seja sobre a sua posição no que diz respeito ao valor a que são associados, ou da localização indevida dos próprios valores. Nos parágrafos seguintes é exibido o exemplo 7 de um *email* PB com vários erros de localização, e existe, complementarmente, a questão da capacidade de compreensão que os recetores da mensagem têm e o quão prejudicada esta pode ser, neste caso, apenas com erros menores. Este exemplo ilustra como erros classificados com uma menor severidade conseguem prejudicar severamente a compreensão de uma mensagem.

(7) Cartão de débito sem contato - 4,95 £ de substituição ou €/ 19,95 lei Entrega de cartão adicional - 4,95 £ ou €/ 19,95 lei Recarga de o Cartão de débito (EEA e não EEA) GRÁTIS Transferências bancárias recebidas GRÁTIS Saques de dinheiro em Caixa Eletrônico em todo o mundo - dentro de o Caixa Eletrônico , Recarga de dinheiro e vale mensal Sofort de 900 £ / €/ 4,500 lei grátis - 2 % de taxa depois .

Aqui estão os detalhes para sua conta Joint Contas recorrentes - €£0/mês - Cartão de débito sem contato : £4,95/4,95€ por pessoa - Suporte de retirada em Caixa Eletrônico de £200/€200 gratuito por mês com 2 % de cobrança por transação após a taxa ser usada - Recarregamentos de dinheiro em o Reino Unido têm uma taxa de : 3,5 % (3£ min) em o PayPoint 2 % (2£ min) em os Correios - Reembolso de £2000/€2000 gratuito de cartão de moeda corrente por mês com 2 % de cobrança por transação após a taxa ser usada - Taxa de 2 % (mínimo de £2/€2 por) para transferências bancárias internacionais de moeda Por favor , note que a conta Premium não pode ser convertida em uma conta bancária , pois são diferentes Avise -me se tiver alguma dúvida.

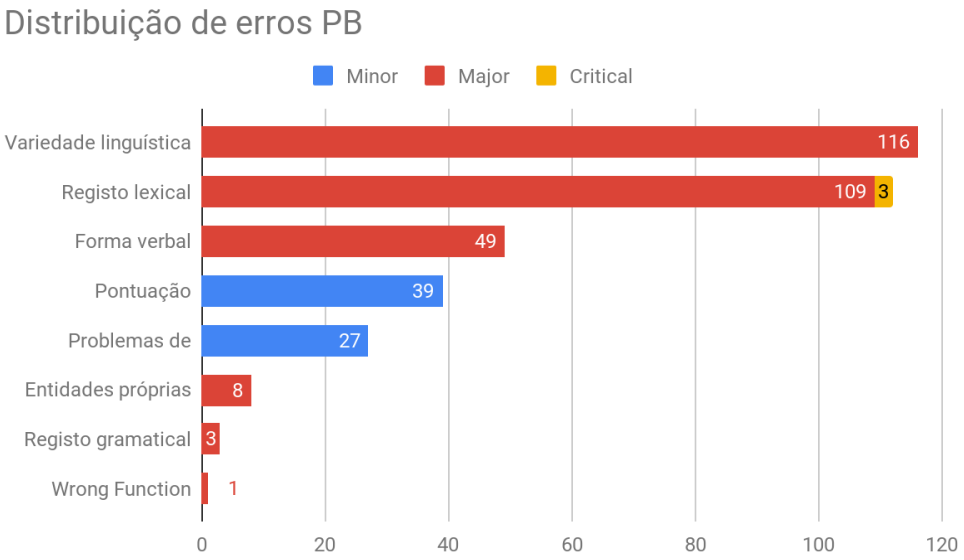

Figura 15: Distribuição de erros PB

6. Conclusão

Este trabalho indica a eficácia destes sistemas de sugestão de regras num contexto de pós-edição. O aumento da qualidade resultante da aplicação deste trabalho é relevante, podendo mesmo resultar em aumentos de 40-50 pontos de MQM em casos de trabalhos com valores mínimos, o que demonstra que análises periódicas podem originar aumentos significativos na qualidade do produto final. Este processo pode ser replicado em todos os pares linguísticos, para todos os clientes que recorram à *Unbabel* para traduzir *emails* ou FAQs. Criaram-se várias regras linguísticas para PE e PB, notando que estes resultados assentam no facto de os editores terem de considerar as sugestões propostas e aceitá-las. Contudo, estas regras, além de ajudarem os editores, também contribuem para a regularização do tom de formalidade das mensagens, como o caso da regra que sinaliza "Olá/Oi" e podem auxiliar outros sistemas, como programas de reconhecimento de entidades, por exemplo, siglas (*i.e*. sugerir EEE, Espaço Econômico Europeu, em vez de AEE, Área Econômica Europeia). Outras regras criadas com base em diferenças ao nível do vocabulário e/ou culturais foram: senha vs. palavra-passe e guardar vs. salvar, mascarar vs. fantasiar, rapariga vs. moça.

Por outro lado, foram excluídas deste projecto tentativas de resolver problemas de localização e de estrutura frásica, embora haja potencial para explorar estas temáticas no futuro. O exemplo 7 na seção 5.3 serviu como um reconhecimento da necessidade de as resolver.

Algumas destas regras podem ser consideradas subjetivas por natureza. Como realizado neste projeto, análises periódicas a conjuntos de dados específicos a um par de línguas podem resultar na criação, e consequente validação, de regras linguísticas a serem aplicadas por sistemas de identificação de erros, a fim de melhorar diretamente e em pouco tempo os índices de qualidade.

Sendo que este estudo consistia na análise de 139 mensagens em PE e 114 mensagens em PB com o objetivo de identificar os erros mais comuns e urgentes cometidos pelos editores e que pudessem ser resolvidos pelo *Smartcheck*, sob a forma de regras linguísticas, *Surf*. Após esta fase, as regras resultantes foram testadas em conjuntos de dados diferentes dos iniciais compostos por 115 mensagens PE e 187 PB. Em termos de resultados, como referida na seção anterior, verificaram-se aumentos de 5 pontos no conjunto de dados PE e de 11 pontos no PB. Estes aumentos comprovam a eficácia das regras *Surf* e o sucesso desta experiência, visto que após as regras terem sido aplicadas, a média de qualidade das

mensagens fica acima de 99 pontos, distinto dos valores iniciais de 94 e 88 para PE e PB, respetivamente.

Alguns exemplos de erros descritos nas seções anteriores demonstram o contributo das ferramentas de identificação de erros. Considerando as melhorias descritas, constata-se que as regras da ferramenta *Smartcheck*, com acesso às regras *Surf*, são relevantes para detetar erros que, embora possam causar dificuldade na compreensão da tradução, podem escapar aos editores que estejam desatentos ou que desconheçam por completo o erro.

Foi observado e comentado previamente as diferenças entre os resultados dos conjuntos no que diz respeito à quantidade de erros detetados e a sua tipologia. Esta variação pode ser devido ao quão fluida a variedade brasileira realmente é e ao facto de a variedade estar geograficamente próxima de outra língua fluida, o inglês, como já foi referido anteriormente.

Com um percurso académico focado mais em linguística do que em código informático, além de conhecimentos básicos, houve um desafio inicial de aprender uma linguagem computacional, especialmente uma que foi criada somente para a ferramenta *Smartcheck*. Como trabalho futuro, há vários campos a explorar, embora dependentes da tecnologia disponível. Estando atualmente a trabalhar na *Unbabel*, é frequente a necessidade de criar uma regra linguística *Surf*. Por um lado, a equipa que gere a comunidade de editores e tradutores, foca-se em criar regras que tenham alcance ao nível da língua. Por outro lado, numa vertente mais focada em apoiar os clientes, há diversas regras que são criadas regularmente com um alcance mais restrito, que correspondam a alternativas de traduções que os clientes possam preferir.

7. Bibliografia

Bar-hillel, Y., "The Present State of Research on Mechanical Translation". *American Documentation*, vol. 2, pt, 4, 1951, pp.229-237. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.5570&rep=rep1&type=pdf

Brito, A., "Gramática: História, Teorias, Aplicações", Universidade do Porto, 2010.

Comparin, Lucia, and Sara Mendes. 2017. "Error detection and error correction for improving quality in machine translation and human post-editing". Proceedings of the 18th International Conference on Intelligent Text Processing and Computational Linguistics– CICLing 2017

Costa, Luís, Diana Santos & Nuno Cardoso, *Perspectivas sobre a Linguateca*.Linguateca. 2008. ISBN: 978-989-20-1445-6. https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS

Costa, P., "Um analisador sintático neural multilíngue baseado em transições". Universidade Federal de São Carlos. 2016. https://repositorio.ufscar.br/bitstream/handle/ufscar/9065/DissPBC.pdf?sequence=1&isAllowed=y

Costa, T., "Um estudo diacônico das variadas realizações do objeto direto anafórico na imprensa baiana dos séculos XIX e XX", Universidade Estadual de Campinas, 2012

Cunha, C & Lindley Cintra, "Nova gramática do português contemporâneo". Lexikon Editora Digital, 2017

do Nascimento, Maria Fernanda, Eugênia Duarte, & Amália Mendes. "Sobre formas de tratamento no Português Europeu e Brasileiro." *Revista Diadorim* [Online], 20 (2018): 245-262. Web. 6 Set. 2021

Duarte, M., e C. Ribeiro Serra. 2015. "Gramática(s), Ensino de Português e 'Adequação Linguística'". *Matraga 22*, 36: 31-55. http://dx.doi.org/10.12957/matraga.2015.17046

Lommel, A. 2015. "Multidimensional Quality Metrics (MQM) Definition". QT21 - Quality Translation 21. 30 de dezembro de 2015. http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

Heidi Depraetere, Joachim Van den Bogaert, Sara Szoc, Tom Vanallemeersch: APE-QUEST: an MT Quality Gate. EAMT 2020: 473-474, https://www.aclweb.org/anthology/W19-6717.pdf

Hutchins, J.. "FIRST STEPS IN MECHANICAL TRANSLATION." (1997).

Hutchins, J.. "MT News International, no. 14", June 1996, pp. 9-12 http://www.hutchinsweb.me.uk/MTNI-14-1996.pdf

Hutchins, J & Somers, H. "An Introduction to Machine Translation", 1992, Academic Press

Kato, M. A. 2005. *Gramática Do Letrado*. In Marques, M. A. et al. (orgs.). *Ciências Da Linguagem: Trinta Anos de Investigação e Ensino*. Braga: CEHUM.

Kato, Mary A. & Ana Maria Martins 2016. European Portuguese and Brazilian Portuguese: an overview on word order. In: Leo Wetzels, Sergio Menuzzi & João Costa (eds.), The Handbook of Portuguese Linguistics. Hoboken, NJ: Wiley-Blackwell. 15-40.

Koehn, P, "Neural Machine Translation", Cambridge University Press, 2020

Martins *et al.*, 2014. http://www.cs.cmu.edu/~ark/TurboParser/

National Research Council. "Language and Machines: Computers in Translation and Linguistics". Washington, DC: The National Academies Press. 1966 https://doi.org/10.17226/9547

Quinn, William A. *Style*, vol. 16, no. 1, 1982, pp. 80–82. *JSTOR*, www.jstor.org/stable/42945405. Accessed 3 Jun. 2021.

Rawling, Piers, and Philip Wilson, eds. *The Routledge Handbook of Translation and Philosophy*. London: Routledge, 2019.

Rei, Ricardo *et al.*, "COMET: A Neural Framework for MT Evaluation", 2020 https://arxiv.org/pdf/2009.09025.pdf

Silva, João, António Branco, Sérgio Castro & Ruben Reis. Out-of-the-Box Robust Parsing of Portuguese. In *Proceedings of the 9th International Conference on the Computational Processing of Portuguese (PROPOR'10)*, 2010, pp. 75–85.

Tarallo, F., & Kato, M. Filling syntactic boundaries in spoken Brazilian Portuguese. *Language Variation and Change, 5*(1), 91-112. 1993. doi:10.1017/S0954394500001423

Testa, I, "Quality in human post-editing of machine-translated texts: error annotation and linguistic specifications for tackling register errors", Faculdade de Letras da Universidade de Lisboa, 2018. http://hdl.handle.net/10451/36289

Torrón, M, in *Produtivity in Post-Editing and in Neural Interactive Translation Prediction: a Study of English-to-Spanish Professional Translators*, 2017

Whitlam, J, "Modern Brazilian Portuguese Grammar: A Practical Guide", Routledge, 2011, ISBN 0-203-84392

8. Webgrafia

https://www.britannica.com/topic/Tower-of-Babel *consultado dia 5 de abril de 2021*

https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/ *consultado dia 16 de julho de 2021*

http://www.qt21.eu/mqm-definition/definition-2015-12-30.html *consultado dia 10 de maio de 2021*

http://www.cs.cmu.edu/~ark/TurboParser/ *consultado dia 10 de maio de 2021*

https://repositorio.ufscar.br/bitstream/handle/ufscar/9065/DissPBC.pdf?sequence=1&isAllowed=y *consultado dia 10 de maio de 2021*

https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS *consultado dia 22 de maio 2021*

http://home.olemiss.edu/~djr/pages/writer/articles/html/cyborg.html *consultado dia 8 de maio de 2021*

https://www.jstor.org/stable/42945405?read-now=1&refreqid=excelsior%3A1564839322bce844b57e540cdf101c3e&seq=1#page_scan_tab_contents *consultado dia 8 de maio 2021*

https://resources.unbabel.com/i/1315162-translation-quality-at-unbabel/0 *consultado dia 18 de junho de 2021*

https://help.unbabel.com/hc/en-us/articles/4408078076439-Quality-reporting-Monitor-the-quality-results-for-your-translations *consultado dia 26 de abril de 2021*

https://www.pgdlisboa.pt/leis/lei_mostra_articulado.php?nid=3118&tabela=leis&nversao= *consultado dia 18 Junho de 2021*

https://support.microsoft.com/en-us/office/check-spelling-and-grammar-in-office-5cdeced7-d81d-47de-9096-efd0ee909227#:~:text=To%20start%20a%20check%20of,your%20document%2C%20just%20press%20F7 *consultado dia 8 Julho de 2021*

https://support.google.com/toolbar/answer/146786?hl=en *consultado dia 8 Julho de 2021*

https://languagetool.org/ *consultado dia 8 Julho de 2021*

https://www.grammarly.com/ *consultado dia 8 Julho de 2021*