Impacto da qualidade de textos de partida criados por utilizadores e agentes e a propagação de erros em sistemas de Tradução Automática

Madalena Gonçalves[1,2], *Marianna Buchicchio*[2,3], *Helena Moniz*[1,4]

[1]Universidade de Lisboa, Faculdade de Letras, Lisboa, Portugal
[2]Unbabel, Lisboa, Portugal
[3]Universidade de Lisboa, CLUL, Lisboa, Portugal
[4]INESC-ID, Lisboa, Portugal

Abstract

This paper proposes a typology concerning errors and linguistic structures found in the source text that have an impact on Machine Translation (MT). The main objectives of this project were firstly, to make a comparison between error typologies and analyze them according to their suitability; analyze annotated data and build a data-driven typology while adapting the previous existing typologies; make a distinction between the errors produced by users and agents in the online Customer Support domain; test the proposed typology with three case studies; methodize patterns in the errors found and verify their impact in MT systems; finally, create a typology ready for production for its particular field. At first, it was made a comparison between different typologies, whether they consider a bilingual or monolingual level (e.g. Unbabel Error Typology, MQM Typology (Lommel *et al.*, 2014b) and SCATE MT Error Taxonomy (Tezcan *et al.*, 2017). This comparison allowed us to verify the differences and similarities between them and, also, which issue types have been previously used. In order to build a data-driven typology, both sides of Customer Support were analyzed — user and agent — as they present different writing structures and are influenced by different factors. The results of that analysis were assessed through the annotation process with a bilingual error typology and were calculated with one of the most highly used manual evaluation metrics in translation quality evaluation — *Multidimensional Quality Metrics* (MQM), proposed in the QTLaunchPad project (2014), funded by the European Union. Through this analysis, it was then possible to build a data-driven typology — Source Typology. In order to aid future annotators of this typology, we provided guidelines concerning the annotation process and elaborate on the new additions of the typology. In the interest of confirming the reliability of this typology, three case studies were conducted in an internal pilot, with a total of 26,855 words, 2802 errors and 239 linguistic structures (represented in the 'Neutral' severity — associated with conversational markers, segmentation, emoticons, etc., characteristics of oral speech) annotated, with different purposes and taking into account several language pairs. In these studies, we verified the effectiveness of the new additions, as well as the transfer of source text errors to the target text. Besides that, it was also analyzed whether the linguistic structures annotated with the 'Neutral' severity had in fact any impact on the MT systems. This testing allowed us to confirm the effectiveness and reliability of the Source Typology, including what needs improvement.

Keywords: Source text, Error annotation, Machine Translation, Customer Support.

Palavras-chave: Texto de partida, Anotação de erros, Tradução Automática, Apoio ao Cliente.

1. Introdução

A Tradução Automática (TA) tem crescido exponencialmente nos últimos anos, contudo o seu foco manteve-se essencialmente no texto de chegada (TC), pois através dele é possível avaliar o desempenho dos seus sistemas. A análise da qualidade do texto de partida (TP) só despertou recentemente interesse na TA.

O estudo desenvolvido neste artigo foi realizado na empresa Unbabel, uma empresa que junta TA com pós-edição humana, em 2021. Uma vez que o domínio da Unbabel é a assistência ao cliente *online*, em que é realizada a tradução de *emails* e conversas de *chat*, foram considerados ambos os interlocutores deste domínio: os agentes e os utilizadores.

Por um lado, temos os agentes, cuja maioria não é falante nativa da língua inglesa, a língua geralmente utilizada em *call centers*, e têm de enfrentar ambientes profissionais com condições precárias e repetitivas. Eles também têm de seguir guiões e terminologia próprios da empresa onde trabalham, enquanto cumprem curtos períodos de resposta, o que resulta em interações mais controladas.

Por outro lado, temos os utilizadores, que são os clientes no mundo inteiro. Estes são, na sua grande maioria, nativos da língua em que comunicam e têm uma maior liberdade na sua comunicação no que diz respeito à expressão de emoções. Sendo o conteúdo analisado neste estudo *chat*, a linguagem de *chat* é bastante específica. Uma vez que esta linguagem pertence a um meio tão dinâmico como a Internet, isso faz com que a linguagem de *chat* seja influenciada por diferentes fatores e, assim, esteja em constante mutação. Além disso, esta linguagem também é uma mistura do discurso oral com o discurso escrito, resultando numa linguagem única com propriedades híbridas. Estas conversas diferem de acordo com a disposição emocional dos interlocutores, o seu nível de escolaridade, o dispositivo utilizado, etc. Todos estes fatores afetam a escrita e leitura das mensagens, e, por fim, a qualidade, não só do TP, mas também do TC. O objetivo deste estudo é desenvolver uma metodologia para medir somente a qualidade do TP e compreender o seu impacto no TC. Para isso, recorreu-se a outras metodologias de avaliação de textos produzidos por falantes não nativos da segunda língua do seu país e a tipologias a nível bilingue, tendo em consideração especificamente o quadro da *Multidimensional Quality Metrics* (MQM) (Lommel *et al.*, 2014b).

2. Estado da arte em tipologias de erros

Uma vez que na TA o foco tem sido maioritariamente na qualidade do TC, o número de tipologias que consideram somente erros do TP ainda é bastante reduzido. Por essa razão, foi decidido fazer uma comparação entre várias tipologias. Neste artigo iremos apenas focar-nos em três: *MQM Typology*[1] (Lommel *et al.*, 2014b), Tipologia de Erros da Unbabel e *SCATE MT Error Taxonomy* (Tezcan *et al.*, 2017).

A tipologia da MQM (Lommel *et al.*, 2014b) é uma tipologia bilingue que tem em conta erros de tradução, contudo esta tipologia também considera que alguns erros podem ocorrer em textos monolingues. MQM é uma métrica de avaliação manual de qualidade reconhecida na área de TA por apresentar flexibilidade aos seus utilizadores e por ser aplicável a todas as línguas, pares de línguas, diferentes conteúdos e contextos (Lommel *et al.*, 2014b). Na sua listagem de classes de erros, cada erro está representado numa tabela com a sua denominação, definição e aplicação (quer esta seja somente a nível bilingue ou a nível bilingue e monolingue). Um exemplo disso será apresentado na Figura 1 com a classe de erro *Agreement*, que tanto se aplica a nível bilingue como monolingue.

[1] https://www.qt21.eu/mqm-definition/definition-2015-12-30.html

Agreement

ID	agreement
Definition	Two or more words do not agree with respect to case, number, person, or other grammatical features
MQM Core?	no
Automatable?	yes
Parent	word-form
Children	*none*
Applies to	source and target
Example(s)	• A text reads "They was expecting a report."

Figura 1: Lista de classe de erros na tipologia da MQM: Agreement.
Exemplo extraído de: https://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html#agreement

A Tipologia de Erros da Unbabel[2] é a tipologia atualmente utilizada na Unbabel e foi produzida em conformidade com a tipologia da MQM, embora especificamente elaborada para o contexto de apoio ao cliente. É uma tipologia bilingue que apenas considera erros de tradução, ou seja, foca-se na relação entre o TP e o TC. Esta tipologia é bastante extensa pela sua especificidade em classes de erros gramaticais, tais como 'Preposição Omitida'. Ainda que estas classes específicas ajudem posteriormente na análise de erros, a sua extensão dificulta a sua aprendizagem, o que pode tornar o processo de anotação mais lento e diminuir a concordância entre os anotadores (tema abordado na secção 6). Foi essencial incluí-la, por ter em conta o mesmo domínio, neste caso o do apoio ao cliente, que foi considerado na produção da nova tipologia.

A *SCATE MT Error Taxonomy* (Tezcan *et al.*, 2017) é uma tipologia que considera erros bilingues e monolingues de forma separada. Na Figura 2, temos a representação desta tipologia, em que os erros encontrados no lado esquerdo são erros considerados a um nível bilingue e no lado direito são os erros considerados a um nível monolingue. Algo que se verificou nesta tipologia foi que certos erros apenas considerados a nível monolingue também ocorrem a nível bilingue, como, por exemplo, erros de ortografia (que incluem pontuação e grafar a maiúscula) e de ordem de palavras.

[2] Esta tipologia é propriedade privada da Unbabel e por essa razão não será apresentada neste artigo.

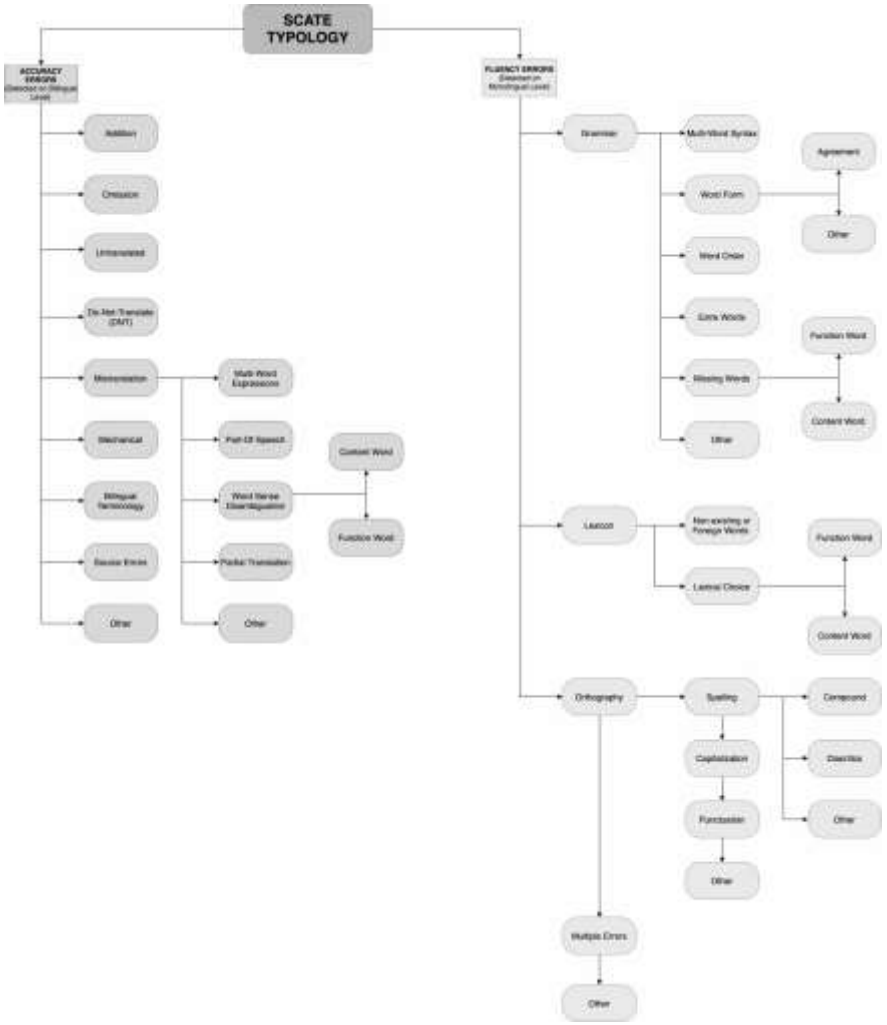

Figura 2: SCATE MT Error Taxonomy (Tezcan et al., 2017)

Deste modo, os sistemas de TA neuronais não produzem tantos erros de fluência, mas sim de precisão semântica, embora muito mais robustos, não são complementamente fiáveis Em conclusão, embora não haja uma tipologia que se centre exclusivamente em erros em textos monolingues, existem já tipologias que consideram certos erros para além da relação com o TC. As tipologias da MQM (Lommel *et al.*, 2014b) e *SCATE MT Error Taxonomy* (Tezcan *et al.*, 2017) consideram erros encontrados tanto no TP como no TC. Outro aspecto que se verificou foi a especificidade de algumas tipologias, como a de MQM (Lommel *et al.*, 2014b) e a Tipologia de Erros da Unbabel, que também podem apresentar dificuldades aos anotadores e, assim, diminuir a sua concordância. Por estas razões, os objetivos para a construção de uma nova tipologia são encontrar erros específicos do TP e diminuir a extensão destes de forma a que haja maior concordância entre os anotadores.

3. Construção de uma tipologia monolingue com base em dados

Desde o início, o objetivo foi criar uma metodologia com base em dados reais. Com isso em mente, primeiramente houve um esforço de anotação de erros com a Tipologia de Erros da Unbabel, que é uma tipologia bilingue inspirada na abordagem original do quadro da MQM. Refira-se que valores de 95 MQM em 100 são considerados níveis de tradução profissional (Sanchez-Torron & Koehn, 2016). Os dados anotados provieram dos agentes, em inglês, e dos utilizadores, em português do Brasil. Foram selecionadas estas línguas por duas razões: o inglês é a língua mais utilizada no domínio de apoio ao cliente e a principal língua de partida (LP) traduzida na Unbabel; uma análise prévia na testagem da Tipologia de Erros da Unbabel demonstrou que o português do Brasil tinha uma maior disparidade entre os dados.

Com estes resultados foi possível verificar quais os erros que também ocorriam a um nível monolingue e se havia erros específicos do TP que precisavam de ser adicionados.

	Agente	Utilizador
Conversas anotadas	170	179
Palavras anotadas	31,440	12,862
MQM	80.32	27.29

Tabela 1: Resultados de anotação do TP com a tipologia bilingue da Unbabel

Embora o número de conversas entre o agente e o utilizador esteja próximo, o número de palavras produzidas pelo agente é consideravelmente mais elevado, devido a guiões de instruções que os agentes têm de seguir na sua comunicação. Em seguida, iremos ilustrar os resultados das anotações dos agentes e dos utilizadores.

3.1 Agente

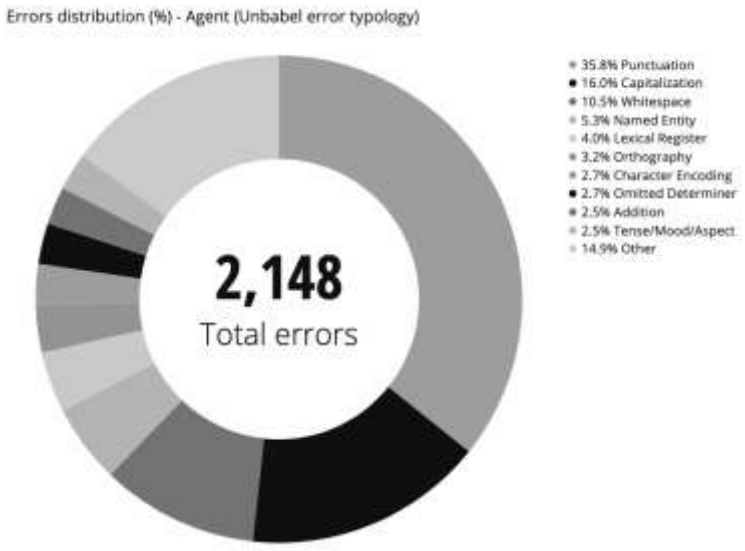

Figura 3: Resultados de anotação dos diálogos apenas do agente

Erros relacionados com ortografia, omissão de determinantes, adição de material linguístico e problemas quanto ao tempo, modo e aspeto verbal comprovam a dificuldade apresentada pelos agentes não nativos da língua inglesa. Rozovskaya & Roth (2010) afirmam que autores não-nativos de uma língua cometem uma variedade de erros de gramática e de escolha lexical. Lee & Seneff (2009) acrescentam que grande parte dos

erros gramaticais cometidos por não-nativos da língua em que comunicam são influência da sua língua materna. E quanto maior for a diferença entre a língua materna e a língua não materna, maior será a incidência de efeitos negativos de transferência de língua (Sirbu, 2015). Para além do seu conhecimento da língua inglesa, outros fatores também influenciam a escrita, tais como a sua escolaridade, o seu contexto cultural, o seu nível de proficiência na língua não materna e a sua motivação para a escrita (Ferris & Hedgcock, 2005 em Kraichoke, 2017).

3.2 Utilizador

Com conteúdo produzido por utilizadores, o contexto de produção dos *chats* muda por completo, como também alguns dos erros encontrados. Devido aos requisitos de manter o diálogo de forma fluída e rápida e à comunicação emocional na assistência ao cliente, os utilizadores tendem a ser menos cuidadosos com pontuação e ortografia.

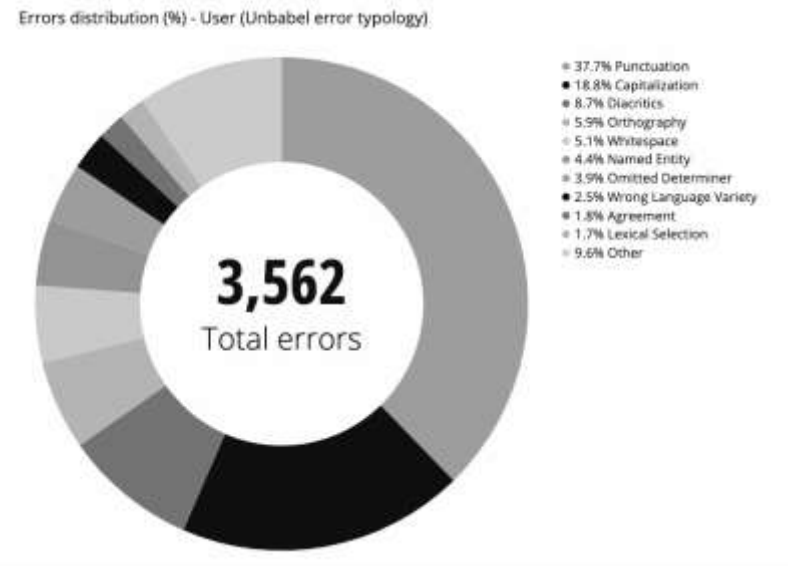

Figura 4: Resultados de anotação dos diálogos apenas dos utilizadores dos serviços

Como é possível verificar, o número de erros encontrados nos textos dos utilizadores é superior ao dos agentes. Os erros encontrados não diferem muito comparados com os do agente, contando-se como erros mais comuns a ortografia e a omissão de determinantes. Pois tal como Höhn *et al.* (2016) afirmam, ser falante nativo de uma língua poderá não significar ter níveis elevados de proficiência linguística. Para além destes erros, outros erros também se salientaram na anotação do utilizador, tais como erros de pontuação, grafar a maiúscula, acentuação e espaços em branco. Esses erros eram de certa forma expectáveis, uma vez que a comunicação do utilizador é mais espontânea e a linguagem de *chat*, sendo mais instantânea, faz com que estes erros pouco graves ocorram com maior frequência. A escrita do utilizador varia de acordo com as suas caraterísticas pessoais (Nars *et al.*, 2016). O modo como alguém se exprime em situações de stress é único e, no contexto de apoio ao cliente, imprevisível, afetando justamente a escrita, a sua tradução e a interação que ocorre entre os interlocutores.

4. Tipologia do Texto de Partida

Com a análise prévia baseada em dados, foi então possível criar uma nova tipologia, que denominamos Tipologia do Texto de Partida. Esta tipologia consiste em quatro categorias principais: *Accuracy*, *Fluency*, *Style* e *Design*. Tal como a Tipologia de Erros da Unbabel, esta tipologia também se inspirou na tipologia da MQM (*Lommel et al*., 2014b), por isso, os nomes das categorias e as suas definições correspondem tanto quanto possível às estipuladas pela MQM, tendo em conta o contexto monolingue do TP e o domínio de apoio ao cliente. Uma exceção relevante é a categoria de *Accuracy*. Esta categoria foi especificamente concebida para o contexto de tradução, para quando o TC não corresponde com precisão ao TP. Sendo esta tipologia exclusivamente concentrada em erros no TP, a sua definição na tipologia apresentada refere-se ao mapeamento ou reconstituição entre uma fonte produzida por um agente ou utilizador em tempo real e a fonte pretendida. É utilizada quando o significado semântico ou a conceptualização de uma ideia é comprometida.

Também se verificou que várias estruturas linguísticas no TP tinham grande impacto no TC, contudo estas estruturas não eram necessariamente erros no TP. Estas estruturas linguísticas são uma marca da linguagem de *chat*, em que as caraterísticas do discurso escrito e oral se encontram. As estruturas linguísticas referidas nesta tipologia foram as seguintes: *Incomplete Sentence*, *Code Switching*, *Emoticon*, *Conversational Marker*, *Idiomatic*, *Profanity*, *Abbreviation* e *Segmentation*. Em seguida serão explicadas as suas definições ilustradas com exemplos e se, previamente, estas estruturas eram consideradas noutras tipologias ou estudos.

Incomplete Sentence é utilizada quando uma frase é interrompida, sendo impossível inferir o seu significado. É geralmente o abandono de uma ideia e uma atualização da ideia pretendida. Esta classe também é considerada na tipologia de MQM (Lommel *et al.*, 2014b), embora com um nome diferente ('*Completeness*'). *Incomplete Sentence* é uma estrutura linguística recorrente nos agentes e nos utilizadores. Devido às restrições de tempo, os agentes têm de apresentar soluções o mais rápido possível, o que por vezes resulta em enviar frases incompletas. Para os utilizadores, isto resulta duma disfluência habitual em conversas *online* em que o discurso se interrompe e uma nova ideia surge. No exemplo (1), apresentamos uma frase incompleta de um agente em inglês.

(1) «Please try to sign in to the link and try to check if a "Continue" or "Next" button **[Ø] [Ø].**»

Code Switching ocorre quando outra língua, para além da LP, é utilizada. Esta classe também foi encontrada num estudo realizado por Hammarberg & Grigonytė (2014). Este fenómeno é bastante recorrente tanto nos agentes, como nos utilizadores. Esta classe tem grande impacto, pois os sistemas de TA não estão preparados para terem uma língua diferente do par de línguas identificado inicialmente na conversa, o que resulta em erros críticos no TC. No exemplo (2) apresentamos uma frase escrita em inglês por um agente a comunicar com um utilizador francês.

(2) «**Merci**. I will now forward this case.»

Emoticon é utilizada para o uso de *emoticons* no TP. O uso de *emoticons* e *emojis* tem-se tornado cada vez mais frequente na linguagem *online* porque a sua utilização é uma forma de eliminar ambiguidade e adicionar contexto emocional à mensagem transmitida (Otemuyiwa, 2017). Esta classe é uma nova adição à tipologia devido, por vezes, ao seu impacto no TC, podendo resultar na alteração do seu significado. No exemplo (3) apresentamos uma frase em inglês produzida por um agente que terminou a frase com *emojis* e a sua alteração crítica no TC em português presente no exemplo (4).

(3) «Hope you're having a great day!  »

(4) «Espero que tenha um bom dia!  »

Conversational Marker é utilizada quando marcadores discursivos são utilizados, pertencendo a diferentes classes gramaticais, tais como conjunções, interjeições, advérbios e orações lexicalizadas (Schiffrin, 1987, citado em Cabarrão *et al.*, 2018). Estes tornaram-se comuns às conversas *online* e são únicos e caraterísticos de cada língua. Como afirmam Cabarrão *et al.* (2018), os marcadores discursivos apresentam dificuldades na tradução devido à sua natureza idiomática. Dada a sua particularidade, a sua tradução nem sempre é a mais correta na LC. Esta classe foi uma nova adição à tipologia devido ao seu uso frequente na linguagem de *chat*. No exemplo (5) apresentamos um marcador discursivo utilizado no português do Brasil.

(5) «**Bom**, agora sim ficou bem claro»

Idiomatic é para o uso de expressões idiomáticas do TP. Uma expressão idiomática é uma expressão fixa numa determinada língua e esta é, por vezes, recuperada através da memória dos falantes. Devido à sua particularidade cultural e linguística, essas expressões apresentam dificuldades aos sistemas de TA. Esta classe foi uma nova adição à tipologia devido à sua utilização por ambos interlocutores. No exemplo (6) apresentamos uma frase inglesa comum utilizada por agentes.

(6) «Hello, **Mary's here**.»

Profanity é para quando um vernáculo obsceno é utilizado. Mesmo que este esteja escrito corretamente, pode também gerar problemas na tradução. Esta classe também é considerada na tipologia da MQM (Lommel *et al.*, 2014b), contudo com um nome diferente ('*Offensive*'). A experiência de apoio ao cliente pode ser stressante para os dois interlocutores, quer seja pelos momentos de espera por uma solução a um problema, ou pela falta de empatia implicada na linguagem de *chat*. Para libertar as suas frustrações, vários utilizadores recorrem ao uso de vernáculos obscenos na interação com os agentes. No exemplo (7) apresentamos o TP em alemão com o vernáculo obsceno destacado a negrito e a sua tradução em inglês para «tímido» presente no exemplo (8).

(7) «Rückerstattung für ein **scheiss** game aus euren Haus...lächerlich geld zu verlangen.»
(8) «Refund for a **shy** game from your home... ridiculous money to demand.»

Abbreviation é a categoria para o uso de abreviaturas. O crescimento da tecnologia e o surgimento da Internet no século XIX permitiram a criação de novos fóruns e, por consequência, de novos conceitos e termos. Tal como Mattiello (2013) afirma, termos abreviados tornaram-se populares pelo seu uso crescente em SMS e pela informalidade. As abreviaturas nem sempre são captadas pelos sistemas de TA, por isso, o seu uso tem impacto no TC. Devido ao seu uso frequente na linguagem de *chat*, esta classe foi uma adição à tipologia. No exemplo (9) apresentamos uma frase produzida por um utilizador português com três abreviaturas («ctt», «q» e «vc») em que uma delas não foi captada pelo sistema e ficou igual no TC inglês, como se pode verificar no exemplo (10).

(9) «Muito obrigado pela atenção. Entrarei em **ctt** com o email **q vc** forneceu.»
(10) «Thank you very much for your attention. I will enter **ctt** with the email **you** provided.»

Segmentation é utilizada para texto segmentado que leva a erros de tradução, como por exemplo ter uma frase dividida a meio. Esta classe também se encontra numa tipologia proposta por Miyata & Fujita (2021) com o nome '*Sentence Splitting*'. Este fenómeno pode ser realizado tanto pelos agentes como pelos utilizadores. Através das constrições de tempo do lado dos agentes e a rapidez da linguagem de *chat* do lado dos utilizadores, pode ocorrer um problema ao digitar e enviar o início de uma frase numa "bolha" de *chat* e noutra enviar o resto da frase. Uma vez que os sistemas de TA não têm modo de saber que a frase foi segmentada, cada segmento é

traduzido individualmente sem o seu contexto geral. Nos exemplos (11) e (12) apresentamos uma frase em inglês que ficou dividida em duas partes.

(11) «If you wish to know about your refund»
(12) «Go to our online site.»

De forma a auxiliar os anotadores durante o processo de anotação, forneceram-se Diretrizes de Anotação. Nessas diretrizes todas as categorias são definidas e exemplificadas; apresenta-se uma secção de esclarecimento de dúvidas sobre estas mesmas e duas árvores de decisão. Estas árvores de decisão foram realizadas com um processo de eliminação em mente, com respostas de "sim" ou "não". Uma árvore de decisão centra-se nas classes de erros (Figura 5), enquanto a outra árvore centra-se nas severidades utilizadas (Figura 6). Árvores de decisão são importantes na construção de uma tipologia, pois são relevantes para a consistência dos anotadores no *Inter-Annotator Agreement* (IAA), prática que permite avaliar a fiabilidade do processo de anotação e, assim, verificar as suas vantagens e o que necessita de ser revisto. Esta prática será apresentada e desenvolvida na secção 6.

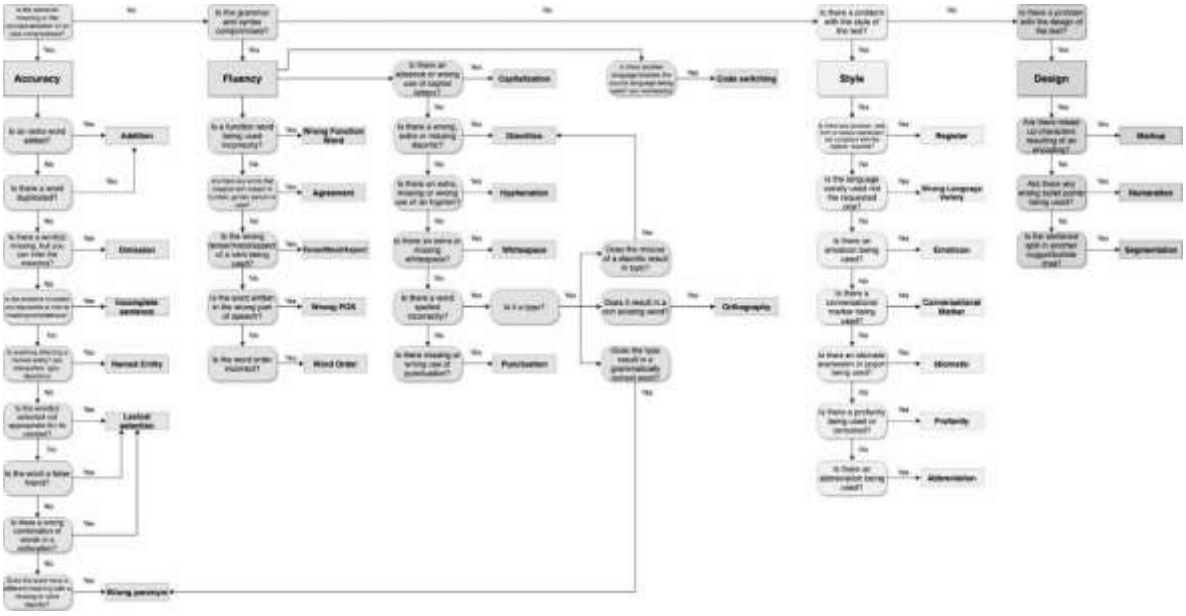

Figura 5: Árvore de decisão quanto às classes de erros

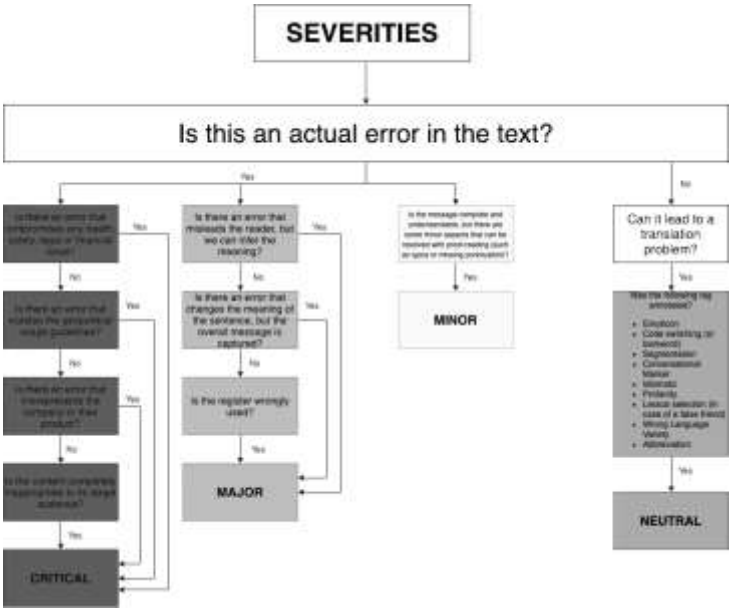

Figura 6: Árvore de decisão quanto às severidades

5. Resultados: piloto

Decidiu-se testar a tipologia e as suas respectivas diretrizes com diferentes dados e pares de línguas. Assim, foi decidido realizar três estudos com propósitos diferentes, de forma a verificar a eficácia e fiabilidade da tipologia proposta: anotação do utilizador (PT-BR_EN), anotação do agente (EN_FR) e piloto interno multilíngue.

5.1 Anotação do utilizador (PT-BR_EN)

O primeiro estudo focou-se nos textos produzidos pelo utilizador, em que a LP foi português do Brasil (PT-BR) e a língua de chegada (LC) inglês (EN). O conteúdo, selecionado aleatoriamente, provinha de diferentes clientes. O TC foi testado em três diferentes sistemas de TA comerciais disponíveis *online*, de forma a verificar diferentes traduções. É importante destacar que o TP foi anotado com a Tipologia do Texto de Partida, enquanto o TC de todos os sistemas de TA foi anotado com a Tipologia de Erros da Unbabel, sendo esta a mais adequada na avaliação de textos traduzidos no domínio de apoio ao cliente. Após o processo de anotação, os valores de MQM foram calculados. O valor de MQM do TP foi de 72.26, o que é longe do ideal, uma vez que são nativos da LP. Os valores de MQM do TC comparados com o do TP foram bastante baixos. Foi decidido então realizar uma comparação entre o TP e o TC com o valor de MQM mais baixo, de maneira a compreender a razão pela descida de qualidade. Desse modo, foi feita uma comparação entre o TP e o Sistema TA 1. As severidades utilizadas no processo de anotação foram: 'neutros' para as estruturas linguísticas no TP que podem impacto no TC; 'pouco graves' para erros que podem ser resolvidos através do processo de revisão; 'graves' para erros que levam a informação incorreta, mudança de significado ou o uso do registo incorreto; 'críticos' para erros que deturpam a imagem do cliente ou do seu produto, conteúdo impróprio para o seu público-alvo ou informação que poderá ter implicações financeiras, legais, de segurança ou de saúde.

	Nº de palavras	Neutros	Pouco graves	Graves	Críticos	MQM
Texto de Partida	2909	27	497	62	0	72.26

Sistema TA 1	2874	N/A	52	175	85	29.05
Sistema TA 2	3003	N/A	86	50	36	38.96
Sistema TA 3	2998	N/A	66	98	49	51.37

Tabela 2: Resultados do estudo "Anotação do utilizador (PT-BR_EN)"

Tal como é possível verificar na Tabela 2, existe uma discrepância evidente entre o valor de MQM do TP e o valor de MQM do Sistema TA 1. Uma das razões para esta diferença é a maior frequência de erros graves e críticos encontrados no TC, enquanto o número desses erros no TP é bastante reduzido ou até mesmo nulo (no caso dos erros críticos no TP). De forma a entender melhor a relação entre o TP e o TC, foi realizado um alinhamento entre as suas anotações. O seu resultado pode ser encontrado na Tabela 3.

	Mesmos erros encontrados em ambos TP e TC	Erros no TP que originaram erros diferentes no TC	Estruturas linguísticas neutras do TP que tiveram um impacto no TC
Número total	34	29	9

Tabela 3: Alinhamento entre TP e TC na Anotação do utilizador

Ao alinhar os dados anotados de cada texto foi possível verificar diferentes géneros de transferência de erros no TC por parte do TP. A transferência mais frequente foi a transferência dos mesmos erros do TP para o TC. Em seguida, também se denotaram erros encontrados no TP que originaram erros de diferente natureza no TC. Por fim, em 27 apenas nove estruturas linguísticas anotadas com a severidade Neutra tiveram de facto impacto no TC. Devido à especificidade destas estruturas linguísticas, dois exemplos serão apresentados para se compreender então o impacto destas e ao mesmo tempo realçar a importância da sua anotação.

	TP	Anotação TP	Classe de erros	Severidade TP	TC	Anotação TC	Classe de erros	Severidade TC
Ex. 1	Abs	Abs/	Abreviatura / Pontuação	Neutro/ Pouco grave	Abs	Abs	Inteligível	Crítico
Ex. 2	Responda ao que eu pergunto sff	/sff	Pontuação/ Abreviatura	Pouco grave/ Neutro	Answer what I ask sff	/sff	Preposição Omitida/ Inteligível	Grave/ Crítico

Tabela 4: Exemplos de abreviaturas no TP com impacto no TC

Na Tabela 4, podemos verificar duas estruturas linguísticas em português ("Abs" e "sff"), neste caso abreviaturas, que tiveram impacto no TC. Estas abreviaturas não foram captadas pelo sistema de TA, pelo que foram mantidas exatamente como se encontram no TP. Sendo que estas abreviaturas não têm qualquer significado na LC, estas estruturas foram anotadas como erros críticos.

5.2 Anotação do agente (EN_FR)

O segundo estudo focou-se no agente, em que a LP era inglês (EN) e a LC era francês (FR). O conteúdo provinha apenas de um cliente, devido aos guiões específicos que os agentes têm de seguir de acordo com a

empresa que representam. Neste estudo aplicou-se o mesmo processo de anotação, embora apenas com um TC, e o resultado foi o oposto do primeiro, em que o valor de MQM do TC foi mais elevado que o do TP. O motivo pelo qual este resultado ocorreu foi porque o sistema de TA utilizado era robusto aos problemas do TP.

	Nº de palavras	Neutros	Pouco graves	Graves	Críticos	MQM
Texto de Partida	9848	17	409	341	0	78.53
Texto de Chegada	10,707	N/A	211	226	2	87.41

Tabela 5: Resultados do estudo "Anotação do agente (EN_FR)"

Como é possível verificar na Tabela 5, embora tenham sido anotados dois erros críticos no TC, a frequência de erros pouco graves e graves anotados no TP é bastante elevada, se comparada com os erros anotados no TC. Decidiu-se, de novo, realizar um alinhamento entre o TP e o TC, para se aferir o impacto da qualidade do TP no TC.

	Mesmos erros encontrados em ambos TP e TC	Erros no TP que originaram erros diferentes no TC	Estruturas linguísticas neutras do TP que tiveram impacto no TC
Número total	59	40	0

Tabela 6: Alinhamento entre TP e TC na anotação do agente

Embora o TP tenha, de facto, tido impacto no TC, em que a transferência de erros foi bastante elevada, as 17 estruturas linguísticas anotadas com a severidade Neutra não tiveram nenhum impacto no TC.

5.3 Piloto interno multilíngue

O terceiro, e último, estudo focou-se num piloto interno multilíngue em que a tipologia e as diretrizes foram testadas internamente na Unbabel. Foram reunidos voluntários para que houvesse *feedback* diferente. Este estudo também foi uma oportunidade para verificar se a tipologia se aplicava a diferentes línguas, para além do inglês e do português, e como funcionaria com utilizadores de diferentes contextos culturais. As línguas anotadas foram neerlandês (NL), polaco (PL), romeno (RO), português do Brasil (PT-BR), italiano (IT), espanhol (ES), alemão (DE) e inglês (EN). Os resultados de cada língua encontram-se na Tabela 7, com os seus respectivos códigos de linguagem.

	Nº de palavras	Neutros	Pouco graves	Graves	Críticos	MQM
NL	2884	27	103	22	5	90.88
PL	1519	9	125	91	0	61.82
RO	536	2	106	14	2	63.43
PT-BR	1838	36	185	31	0	81.5
IT	977	13	67	3	4	87.51
ES	1560	13	186	145	0	85.71
DE	1942	34	186	145	0	53.09
EN (agente)	2842	51	193	43	0	85.46

Tabela 7: Resultados do estudo "piloto interno multilíngue"

Os valores de MQM variam bastante de acordo com o par de línguas anotado, com neerlandês com o maior valor de MQM (90.88) e alemão com o valor mais baixo (53.09).

Novamente, este estudo focou-se essencialmente no impacto das estruturas linguísticas anotadas no TP com a severidade Neutra na qualidade do TC. Por essa razão foram verificados o número de segmentos anotados, o número total de erros anotados e o número de estruturas linguísticas anotadas como neutras.

	Segmentos no TP	Número total de erros	Estruturas neutras
DE	236	365	33
EN	294	287	52
PT_BR	212	252	35
PL	140	225	9
ES	127	180	13
NL	357	157	27
RO	61	124	2
IT	159	87	24

Tabela 8: Análise de estruturas neutras no Piloto interno multilíngue

Em praticamente todas as línguas, exceto o italiano e o neerlandês, o número de erros é muito mais elevado que o número de segmentos, o que significa que poderá ocorrer mais do que um erro por frase.

6. Inter-Annotator Agreement

O acordo entre os anotadores (do inglês, *Inter-Annotator Agreement* - IAA), é uma prática comum após o processo de anotação utilizada para diferentes objetivos, tais como comprovar e melhorar diretrizes de anotação, identificar ambiguidades ou dificuldades ou ter acesso a diferentes interpretações (Artstein, 2017). O IAA permite a avaliação da fiabilidade de todo o processo de anotação e, assim, verificar as suas vantagens atuais e o que necessita de ser melhorado. Como Lommel *et al.* (2014a) afirmam, uma das maneiras mais eficazes para verificar a fiabilidade de um método de avaliação é a sua aplicação consistente por parte dos seus anotadores.

No presente artigo, o IAA vai ser utilizado nos dados anotados no piloto interno multilíngue, por neste piloto ser possível recorrer a mais do que um anotador em algumas das línguas analisadas. Embora a Tipologia do Texto de Partida tenha sido elaborada com uma extensão mais reduzida, comparada com as tipologias verificadas no processo inicial, ou as diretrizes tenham uma secção dedicada para dúvidas entre as categorias, suportada ainda com árvores de decisão suplementares, pode, contudo, ocorrer alguma confusão com certas categorias. Sendo a anotação um processo de avaliação humana, já é expectável haver variação no acordo entre os anotadores, não só uns com os outros, mas também consigo mesmos (Artstein, 2017). Mesmo ao auxiliar com diretrizes e definições detalhadas, a subjetividade do anotador é um fator que pode influenciar o IAA. Lommel *et al*. (2014a) explicam que uma das causas de discórdia no processo de anotação de erros ou de pós-edição é a diferente análise e interpretação de erros. O próprio anotador pode não concordar com a definição de erro estabelecida nas diretrizes e ter a sua própria noção em mente ao anotar. Este tipo de resultado não é incomum, tal como foi realçado também por Lommel *et al.* (2014a), nas suas próprias experiências.

O IAA é calculado através de vários coeficientes, neste caso específico iremos apenas focar-nos em dois deles — *Cohen's Kappa* e *Pearson*. *Kappa* é um coeficiente quantitativo que mede o acordo entre dois avaliadores, neste caso dois anotadores, que estão a avaliar o mesmo conteúdo.

Os valores do *Kappa* vão desde números negativos (menos que 0) até 1. Um valor negativo significa que existe muito pouco acordo e 1 significa que existe acordo total.

O coeficiente *Pearson* é representado por *r* e é uma medida de associação linear entre duas variáveis. Os seus valores vão desde -1 a 1. Se o seu valor é inferior a 0, há uma ligação negativa entre as variáveis, quando uma aumenta a outra decresce. Se o valor é 0, não há qualquer associação entre as duas variáveis. Se o valor é mais elevado que 0, então a sua associação é positiva.

Enquanto os coeficientes de acordo, como o Kappa, são utilizados para estudos iniciais de melhoramento de diretrizes e análise de dados que ilustram quão distantes são as interpretações dos anotadores, coeficientes de correlação, como Pearson, indicam até que ponto os anotadores foram consistentes entre si (Amidei et al., 2019).

Devido a restrições de tempo, apenas os dados do utilizador vão ser utilizados para a realização de IAA. Em primeiro lugar, serão analisados os resultados do acordo que os anotadores tiveram consigo próprios. Este tipo de acordo também é importante, pois irá permitir verificar se os anotadores tiveram alguma dificuldade na aprendizagem da tipologia. Não ter concordância interna no processo de anotação é um indicador do possível grau de complexidade de uma tipologia. Na Tabela 9 iremos apresentar os resultados do acordo dos anotadores com eles mesmos. De forma a identificar o anotador e a língua, os anotadores serão apresentados de acordo com o seu código de linguagem seguido de 1 ou 2, o que significa que há dois anotadores para essa língua.

Anotador	Valor de acordo a nível de frase
PL	0.9
ES-1	1.0
ES-2	0.7
RO	0.9
NL	1.0
PT-BR	0.9
IT-1	1.0
IT-2	0.9
DE-1	1.0
DE-2	1.0

Tabela 9: Acordo dos anotadores consigo mesmos ao nível de frase

Artstein (2017) informa que a recomendação de Carletta, em que valores de coeficientes acima de 0.8 são fiáveis (Artstein, 2017), foi aceite na comunidade de linguística computacional. Ao reunir vários artigos de pesquisa sobre o IAA, foi possível determinar uma média entre os valores máximos e mínimos (Amidei *et al.*, 2019). Deste modo, o valor mínimo de *r* de *Pearson* é de 0.20. Como é possível verificar na Tabela 9, grande parte dos anotadores foi consistente consigo mesmo, resultando em valores de 0.9 e 1.0, exceto o anotador ES-2, com o valor de 0.7. Uma vez que os resultados estão maioritariamente acima de 0.8, o resultado pode ser interpretado como as categorias e materiais de apoio foram simples e explícitos para os anotadores perceberem o funcionamento da tipologia.

Embora haja consistência dos anotadores consigo próprios, o mesmo pode não se traduzir no IAA. Em seguida, iremos apresentar as Tabelas 10 e 11 com valores *Pearson* de IAA a nível de frase e documento nas línguas com dois anotadores, que neste caso são apenas espanhol, italiano e alemão.

Língua	Valor de IAA a nível de frase
ES	0.5
IT	0.5
DE	0.1

Tabela 10: Acordo interno a nível da frase

Língua	Valor de IAA a nível de documento
ES	0.1
IT	0.5
DE	-0.1

Tabela 11: Acordo interno a nível do documento

Em ambas as tabelas, podemos verificar que, embora os valores a nível de frase não sejam bastante elevados, especialmente com os anotadores de alemão, com um valor *Pearson* de 0.1, o IAA foi positivo. Quanto ao IAA tomando como base todo o documento, apenas os anotadores de italiano mantiveram a sua consistência, enquanto os anotadores das outras línguas diminuíram o grau de acordo. Um exemplo disso foi o acordo no espanhol ainda se encontrar positivo, mas no alemão já não. Sendo estes resultados muito gerais, foi decidido utilizar o coeficiente *Kappa* para verificar o acordo em áreas mais específicas, nas categorias, classes de erros e severidades. Os resultados são apresentados, respetivamente, nas Tabelas 12, 13 e 14.

Língua	Valor de IAA a nível de categorias
ES	0.3
IT	0.2
DE	0.2

Tabela 12: Acordo interno a nível de categorias

Língua	Valor de IAA a nível de classes de erro
ES	0.3
IT	0.3
DE	0.2

Tabela 13: Acordo interno a nível de classes de erro

Língua	Valor de IAA a nível de severidades
ES	0.3
IT	0.2
DE	0.2

Tabela 14: Acordo interno a nível de severidades

Tal como foram decididos os valores máximos e mínimos dos coeficientes (Amidei *et al.*, 2019), o valor mínimo estabelecido para o *Kappa* foi de 0.10. Como é possível verificar nas Tabelas 12, 13 e 14, os valores foram sempre acima do valor mínimo de *Kappa*. Em conclusão, os resultados do IAA foram satisfatórios para uma primeira testagem da tipologia apresentada. Como é confirmado em Lommel *et al.* (2014a), a compreensão humana da linguagem é bastante variável, até mesmo em tradutores profissionais. Uma vez que os anotadores não estão habituados a analisar o TP, isto só comprova que é necessário mais tempo no treino dos anotadores numa futura testagem.

7. Conclusões e trabalho futuro

Através deste estudo foi possível comprovar o impacto do TP na qualidade do TC, em que não só os erros como também estruturas linguísticas influenciam a TA.

Apesar de não haver uma tipologia que considerasse exclusivamente erros em textos monolingues, a verificação de tipologias prévias que tomavam em consideração alguns desses erros auxiliou bastante a construção de uma nova tipologia. Foi fundamental tomar consciência das diferenças entre os dois interlocutores presentes na assistência ao cliente e compreender os desafios linguísticos provenientes delas. A linguagem de *chat* também apresenta desafios adicionais pela sua especificidade, como por exemplo a sua urgência e a mistura entre o discurso escrito e o oral.

Ao realizar três estudos distintos, foi possível verificar o impacto da escrita de agentes e utilizadores, especialmente quanto ao seu uso de estruturas linguísticas caraterísticas da linguagem de *chat* e de conversas *online*.

Após este estudo, foi possível adaptar e melhorar a Tipologia do Texto de Partida e as suas diretrizes de forma que estivessem prontas a ser utilizadas num contexto comercial e de produção na área de apoio ao cliente. Foram realizadas novas experiências com dados reais de clientes da Unbabel para comprovar a eficácia da tipologia e o impacto que os erros e estruturas linguísticas no TP têm na qualidade do TC (Gonçalves *et al.*, 2022). A Unbabel também tem a intenção de testar a Tipologia do Texto de Partida para que fique alinhada com a Tipologia de Erros da Unbabel e, por fim, produzir uma Tipologia de Erros na Unbabel que tenha em consideração tanto os erros do TP, como também do TC.

8. Agradecimentos

Os autores desejam agradecer a João Graça e Alon Lavie, por terem acreditado na importância da qualidade do texto de partida e o seu impacto na qualidade do texto de chegada produzido por sistemas de Tradução Automática. Também desejamos agradecer às equipas de engenharia da Unbabel, pelo seu auxílio, e aos voluntários que aceitaram testar a Tipologia do Texto de Partida numa primeira fase inicial. Este trabalho foi desenvolvido no âmbito do projeto MAIA, número de contrato 045909 com referência UIDB/50021/2020.

9. Referências

Amidei, Jacopo, Piwek, Paulo & Willis, Alistair (2019) Agreement is overrated: A plea for correlation to assess human evaluation reliability. *Proceedings of The 12th International Conference on Natural Language Generation*, pp. 344-354.

Artstein, Ron (2017) Inter-annotator Agreement. In. Ide, Nancy & Pustejovsky, James. *Handbook of Linguistic Annotation.* Springer Nature, pp. 297-314.

Cabarrão, Vera, Moniz, Helena, Batista, Fernando, Ferreira, Jaime, Trancoso, Isabel & Mata, Ana Isabel (2018) Cross-domain analysis of discourse markers in European Portuguese. *Dialogue & Discourse* 9 (1) pp. 79-106.

German Research Center for Artificial Intelligence (DFKI) & QTLaunchPad, [https://www.qt21.eu/mqm-definition/definition-2015-12-30.html] disponível em endereço [consultado em novembro de 2020].

Gonçalves, Madalena, Buchicchio, Marianna, Stewart, Craig, Moniz, Helena & Lavie, Alon (2022) Agent and User-Generated Content and its Impact on Customer Support MT. *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 199-208.

Hammarberg, Björn & Grigonytė, Gintarė (2014) Non-Native Writers' Errors – a Challenge to a Spell-Checker. *P1st Nordic workshop on evaluation of spellchecking and proofing tools (NorWEST2014)*.

Höhn, Sviatlana, Pfeiffer, Alain & Ras, Eric (2016) Challenges of error annotation in native/non-native speaker chat. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pp. 114-124.

Kraichoke, Casey (2017) *Error Analysis: A Case Study on Non-Native English Speaking College Applicants' Electronic Mail Communications.* Tese de mestrado, University of Arkansas.

Lee, John, & Stephanie, Seneff (2009) An analysis of Grammatical Errors in Non-Native Speech in English. *MIT Computer Science and Artificial Intelligence Laboratory*.

Lommel, Arle, Popović, Maja & Burchardt, Aljoscha (2014a) Assessing Inter-Annotator Agreement for Translation Error Annotation. *Conference: LREC Workshop on Automatic and Manual Metrics for Operational Translation Evaluation.*

Lommel, Arle, Uszkoreit, Hans & Burchardt, Aljoscha (2014b) Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció.* (12), pp. 455-463.

Mattiello, Elisa (2013) Extra-grammatical Morphology in English. Abbreviations, Blends, Reduplicatives, and Related Phenomena. De Gruyter Mouton.

Miyata, Rei & Fujita, Atsushi (2021) Understanding Pre-Editing for Black-Box Neural Machine Translation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1539–1550.

Nars, Alexis, Damnati, Geraldine, Guerraz, Aleksandra & Bechet, Frederic (2016) Syntactic parsing of chat language in contact center conversation corpus. *Proceedings of the SIGDIAL 2016 Conference*, pp. 175–184.

Oliveira, Luanna de Sousa do Nascimento (2017) *Expressões Fixas do Português Formadas a partir de Nomes Gerais: aspectos lexicais e variacionistas.* Tese de mestrado, Faculdade de Letras da UFMG.

Otemuyiwa, Abosede Adebola (2017), «A Linguistic Analysis of WhatsApp Conversations among Undergraduate Students of Joseph Ayo Babalola University», *Studies in English Language Teaching*, 5, 3, pp. 393-405, disponível em https://www.researchgate.net/publication/317414364_A_Linguistic_Analysis_of_WhatsApp_Conversations_among_Undergraduate_Students_of_Joseph_Ayo_Babalola_University [consultado em 2021].

Rozovskaya, Alla & Roth, Dan (2010) Annotating ESL Errors: Challenges and Rewards. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 28-36.

Sanchez-Torron, Marina & Philipp, Koehn (2016) Machine Translation Quality and Post-Editor Productivity. *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track* 18, pp. 16-26.

Sirbu, Anca (2015) Language Transfer Triggered by Bilingualism. *"Mircea cel Batran" Naval Academy Scientific Bulletin* 18 (1), pp. 374-376.

Tezcan, Arda, Hoste, Véronique & Macken, Lieve (2017) SCATE Taxonomy and Corpus of Machine Translation Errors. In. Pastor, Gloria Corpas & Durán-Muñoz, Isabel, James. *Trends in E-Tools and Resources for Translators and Interpreters.* Brill, pp. 219-244.