Mestrado em Tradução

Especialização em Inglês

Relatório de estágio

Variantes do inglês e inteligência artificial para tradutores: estudo de caso na empresa *Unbabel*

Ester Maria da Silva Piedade

a2022103331

Maio de 2024

Relatório de Estágio apresentado para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Tradução realizado sob a orientação científica do Prof. Doutor Marco Neves e da Doutora Marina Sánchez Torrón.

Agradecimentos

Quero agradecer, antes de mais, os meus orientadores Prof. Doutor Marco Neves e Doutor Marina Sánchez Torrón por todo o apoio no decorrer deste estágio.

Ao Professor Marco por estar sempre disponível para esclarecer todas as dúvidas, por ser sempre compreensivo e, no geral, pela sua empatia e disponibilidade que tiveram um profundo impacto na minha experiência académica.

À Marina, por estar sempre presente, pela infinita paciência e disponibilidade. Cujo o apoio e orientação constantes foram fundamentais para a realização desta investigação. Agradeço também a sua capacidade de proporcionar sempre um ambiente positivo, pois transformou todo este percurso desafiante numa experiência gratificante.

À Unbabel, e à *Community Team* por me acolherem de braços abertos, pela oportunidade de realizar este estágio e por toda a ajuda.

À Cláudia, por me ouvir pacientemente durante meses a falar do mesmo tópico e por me conseguir animar e trazer sempre alguma tranquilidade e conforto nos dias mais turbulentos.

Ao resto dos meus amigos (vocês sabem quem são), obrigada por serem sempre uma luz na minha vida.

E os meus pais, por me ouvirem, por todo o apoio e compreensão durante todo o meu percurso académico.

VARIANTES DO INGLÊS E INTELIGÊNCIA ARTIFICIAL PARA TRADUTORES: ESTUDO DE CASO NA EMPRESA UNBABEL

RESUMO

Com a evolução da tecnologia e da inteligência artificial (IA), em particular com o recente fenómeno, GPT-4, o papel do tradutor volta a estar no centro de debate e discussão. De facto, a eficácia destes sistemas de IA é inegável. Contudo, é possível encarar este progresso como uma oportunidade para considerar a IA como uma ferramenta de apoio ao tradutor. Será que os tradutores poderiam melhorar os seus conhecimentos linguísticos com o auxílio do GPT? E se as empresas decidissem investir na formação dos seus tradutores e linguistas internos?

Este relatório apresenta um estudo abrangente sobre a criação de testes para o treino especializado de falantes não nativos em inglês britânico, visando a avaliação e formação de editores em tradução e localização intralinguística. Utiliza-se o sistema Generative Pretraining Transformer 4 (GPT-4) para automatizar a criação de testes. A investigação centrar-se-á na distinção entre o inglês americano e o inglês britânico, procurando responder à questão se o GPT-4 é eficaz no treino de linguistas não nativos em inglês britânico.

Para responder a esta pergunta, foi necessário explorar as capacidades do GPT-4 na tarefa mencionada, incluindo a criação de *prompts*, geração de bancos de questões e a subsequente avaliação das questões geradas por um linguista britânico. Os dados recolhidos mostram a infiabilidade do GPT e a necessidade de supervisão humana.

PALAVRAS-CHAVE: Inglês Britânico; Tradução; Inteligência Artificial; GPT-4; Treino; Inglês Americano

VARIANTES DO INGLÊS E INTELIGÊNCIA ARTIFICIAL PARA TRADUTORES: ESTUDO DE CASO NA EMPRESA UNBABEL

ABSTRACT

With the evolution of technology and artificial intelligence (AI), particularly with the recent phenomenon, GPT-4, the role of the translator is once again at the center of debate and discussion. Indeed, the effectiveness of these AI systems is undeniable. However, it is possible to view this progress as an opportunity to consider AI as a support tool for the translator. Could translators improve their linguistic skills with the help of GPT? And what if companies decided to invest in training their internal translators and linguists?

This report presents a comprehensive study on the creation of tests for specialized training of non-native speakers in British English, aimed at the assessment and training of editors in intralinguistic translation and localization. The Generative Pretrained Transformer 4 (GPT-4) system is utilized to automate the creation of tests that assist in the training and evaluation of translators in this specific variant of English. The research will focus on distinguishing between American English and British English, seeking to answer the question of whether GPT-4 is effective in training non-native linguists in British English.

To answer this question, it was necessary to explore GPT-4's capabilities in the mentioned task, including creating prompts, generating question banks, and the subsequent evaluation of the questions generated by a British linguist. The collected data shows GPT's unreliability and the need for human supervision.

KEYWORDS: British English; Translation; Artificial Intelligence; GPT-4; Training; American English

ÍNDICE

Conteúdo

Lista de acrónimos

ALPAC - Automatic Language Processing Advisory Committee

API - Application Programming Interface

CNN - Convolutional Neural Networks

GPT - Generative Pre-Transformers

IA - Inteligência Artificial

MQM - Multidimensional Quality Metrics

LLM - Large Language Model

PLN - Processamento de Linguagem Natural

QE - Quality Estimation

RNN - Recurrent Neural Networks

1. INTRODUÇÃO

O presente relatório visa ilustrar o período de estágio realizado na empresa Unbabel que decorreu entre os meses de Setembro de 2023 e Fevereiro de 2024 no âmbito da componente não-letiva do mestrado de Tradução. A Unbabel é uma empresa que atua no setor da tradução automática suportada pela inteligência artificial. Combinando esta tecnologia avançada com uma comunidade global de tradutores humanos, a empresa oferece traduções rápidas e de alta qualidade em vários idiomas, permitindo superar barreiras linguísticas e promovendo uma comunicação eficiente entre empresas e clientes de origens linguísticas e culturais distintas.

A oportunidade de realizar um estágio nesta empresa revelou-se extremamente valiosa, tendo em conta o seu avanço tecnológico e o empenho em promover a investigação em diversas áreas que afetam a perceção e, consequentemente, a prática da tradução. Face ao rápido desenvolvimento da inteligência artificial e da tradução automática, novas questões emergem. Em vez de persistirmos em debater sobre técnicas ou métodos de tradução, o nosso foco começa a deslocar-se para o que designamos por "pós-tradução": revisão pós-tradutória, controlos de qualidade, entre outros.

Nesta perspetiva, o papel do tradutor mantém-se fundamental. Com efeito, os avanços linguísticos recentes suscitam inquietações quanto à posição do tradutor no seio da indústria. Efetivamente, este caso não se limita apenas à prestação de serviços linguísticos, mas a uma variedade de carreiras como escritores, designers, ilustradores, analistas, etc. Um artigo do jornal *Washington Post* (Verma & De Vynck, 2023) ilustra experiências pertinentes a esta realidade: "*she was let go without explanation,* (...) *using ChatGPT was cheaper than paying a writer, the reason for her layoff seemed clear.*" No entanto, à medida que surgem mais reportagens e investigações sobre este assunto, os académicos realçam a importância de encarar a IA como um complemento à inteligência e criatividade humanas (Tu et al., 2023). É exatamente por esta razão que se revela imprescindível o avanço da pesquisa nesta área. Um entendimento mais profundo destas ferramentas pode propiciar a melhoria da qualidade do trabalho humano.

Tendo tudo isto em consideração, este relatório terá como tema a investigação do impacto que um *Large Language Model* como GPT-4, é útil ou não no treino e avaliação dos tradutores ingleses não nativos da variante britânica, mais especificamente nativos de

inglês americano, para lidar com a variante alvo deste relatório: inglês britânico. Esta investigação decorre de uma necessidade de investigação da própria empresa, que, devido a desenvolvimentos recentes na sua organização, necessita especificamente de treinar tradutores no uso dessa variante. Considerando que o mestrado visa formar tradutores e prepará-los para o futuro na sua área, torna-se essencial que fiquem ao corrente do progresso tecnológico, desenvolvendo a literacia em meios que têm um impacto significativo na indústria.

Com este projeto, pretende-se analisar como, no futuro, poderemos considerar a IA como um recurso para o treino de tradutores que ambicionam aperfeiçoar as suas competências linguísticas.

Sendo assim, o presente relatório começará a sua exposição com uma contextualização da empresa que acolheu este estágio, a Unbabel, percorrendo o seu historial de atividades (secção 2) e as metodologias adotadas no domínio da tradução automática, inteligência artificial e *machine learning* (secção 2.1., secção 2.2., secção 2.2.1., secção 2.2.2. e secção 2.2.3.), finalizando esta parte com os objetivos do estágio (secção 2.4.). Prosseguir-se-á com uma revisão literária (secção 3), na qual serão apresentados e analisados os mais recentes desenvolvimentos no campo de *Large Language Models* (Secção 3.1., secção 3.1.1, secção 3.1.2, secção 3.1.3), incluindo o GPT-4 e os seus primórdios (Secção 3.2.). Abordar-se-á, ainda, o domínio da Engenharia Prompt (secção 3.3.), destacando os impactos significativos decorrentes das metodologias avançadas desenvolvidas por especialistas. Este segmento visa explorar estratégias de otimização aplicadas ao GPT-4, visando aprimorar a sua performance (secção 3.3.1.). Subsequentemente, focar-se-á no uso da Inteligência Artificial (IA) analisando as oportunidades e desafios apresentados em contextos educacionais (secção 3.4.). Em seguimento haverá uma exposição do modelo andragógico (secção 3.5.)

Prossegue-se com uma análise comparativa entre o inglês britânico e o americano, cujas distinções foram essenciais para os testes iniciais com o GPT-4. Esta secção visa elucidar as características linguísticas específicas que influenciam os sistemas de IA no processamento da língua inglesa (secção 3.6., secção 3.6.1., secção 3.6.2, secção 3.6.2.1., secção 3.6.2.2.).

No segmento subsequente, introduzir-se-á a metodologia (secção 4), juntamente com a questão de investigação que norteia este relatório (secção 4.1.1), seguida da

delineação dos resultados ideais. Esta secção detalhará o processo metodológico adotado (secção 4.2), visando eliminar viés e outras limitações inerentes à investigação. Seguidamente, haverá um desenvolvimento de todas as descobertas, processos, obstáculos e testes elaborados que mapearam a busca da resposta à pergunta de investigação (secção 4.3., secção 4.3.1., secção 4.4., secção 4.5., secção 4.5.1., secção 4.5.2., secção 4.5.3., secção 4.5.4., secção 4.5.5., secção 4.6.,secção 4.7.).

A fase final deste relatório é dedicada à apresentação dos resultados (secção 5), que serão correlacionados com a resposta à pergunta de investigação inicialmente proposta (secção 5.1., secção 5.2., secção 5.3., secção 5.4., secção 5.5., secção 5.6.). Ainda haverá uma proposta de novos *prompts* corelacionados com os resultados da investigação (secção 6). Concluir-se-á com uma discussão sobre as conclusões obtidas (secção 7.1), limitações do estudo (secção 7.2.) e sugestões para pesquisas futuras (secção 7.3).

2. APRESENTAÇÃO DA EMPRESA

Como referido anteriormente, o estágio em questão decorreu na Unbabel, uma empresa portuguesa fundada em 2013 por Vasco Pedro, João Graça, Hugo Silva, Sofia Pessanha e Bruno Prezado. A Unbabel dedica-se a serviços de tradução assistida por IA e tradução automática. Atualmente, a empresa possui escritórios em cidades como São Francisco, Nova Iorque, Lisboa, Edimburgo, Londres, Cebu e Timisoara. Atualmente os seus serviços oferecem traduções entre 130 pares de línguas.

Dado que a Unbabel realiza a totalidade das suas traduções com recurso à IA, a implementação de controlos de qualidade torna-se essencial. A despeito dos avanços significativos, a inteligência artificial e, consequentemente, a tradução automática ainda apresentam erros e imperfeições que, se ignorados, podem comprometer radicalmente a qualidade de uma tradução. Neste contexto, a Unbabel recorre a processos de controlo de qualidade que combinam a supervisão humana com *machine learning*.

Antes da aquisição das empresas EVS e Lingo24, a oferta de serviços da Unbabel estava exclusivamente focada no Apoio ao Cliente. Com isto em mente, os procedimentos de controlo de qualidade foram especificamente desenvolvidos para atender a esse segmento, abarcando o tipo de conteúdo habitualmente associado ao suporte ao cliente, como *chat*, *tickets* e perguntas frequentes (FAQs). Contudo, à medida que a empresa expandiu, tornou-se necessário implementar várias *pipelines* especializadas para diferentes finalidades. A razão para esta diversificação prende-se com as distintas exigências de velocidade e qualidade de tradução para cada serviço.

2.1 Métodos de Tradução

A Unbabel tem métodos de tradução apenas com tradução automática e processos de tradução automática com intervenção humana.

No contexto da tradução de conversas de *chat*, que são empregadas para suporte e interação com os clientes, a tradução é realizada exclusivamente recorrendo à tradução automática. Para traduções apenas com tradução automática, o processo revela-se curto dado que a comunicação, e consequentemente a tradução, realiza-se frequentemente em tempo real.

No início do processo, logo após o cliente requisitar um serviço de tradução, procede-se a uma fase de anonimização dos dados, assegurando que todas as informações pessoais do cliente são ocultadas ou convertidas em equivalentes semânticos que preservam a sua privacidade e segurança. Essa conversão é realizada por um sistema de *machine learning* especialmente desenvolvido para executar tais tarefas. O sistema identifica e substitui elementos como nomes, endereços, números de telefone e palavras-passe, utilizando um módulo de Reconhecimento de Entidades Nomeadas (REN) fornecido pela Unbabel. Porém, quando a tradução é entregue ao cliente, as entidades são restauradas no seu formato e conteúdo original, preservando assim, todas as suas informações. Uma vez finalizado este processo, avança-se para a fase de tradução automática, E finalmente, a tradução é imediatamente fornecida ao cliente. Posteriormente, o trabalho prossegue para uma equipa de anotadores, incumbidos de

registar e avaliar todos os erros identificados, contribuindo, assim, para o enriquecimento do conjunto de dados utilizados no treino de algoritmos de machine learning.

Para a tradução especializada, destinada a textos de maior complexidade, como tradução de áreas específicas como medicina, financeira, marketing, conteúdos de web etc. que habitualmente albergam terminologia específica e requisitos de qualidade mais exigentes, o processo difere ligeiramente. Assim, todo o processo conserva as similitudes da *pipeline* anterior, passando pela anonimização dos dados, seguindo-se para a tradução automática que, ao ser concluída, dá lugar ao procedimento de Estimativa de Qualidade (em inglês, Quality Estimation ou QE). Este sistema, também adestrado por meio de técnicas *machine learning*, com os dados das anotações (ver secção 2.2.1), é encarregado de calcular uma pontuação que reflete a qualidade da tradução e identifica os segmentos de texto que necessitam de revisão. Caso a pontuação de qualidade seja considerada baixa, a comunidade de tradutores procede à revisão das traduções, antes de estas serem entregues ao cliente. Todavia, no término da cadeia procedimental, o texto é sujeito à revisão por especialistas do domínio em questão. Deste modo, aferir-se-á que esta modalidade de tradução está sob um controlo qualitativo mais rigoroso e, ao contrário das traduções efetuadas em diálogos ao vivo (como os *chats*) este processo revela-se mais demorado. Adicionalmente, mesmo após a finalização do processo, realiza-se um controlo de qualidade por meio do Métricas de Qualidade Multidimensionais (em inglês, Multidimensional Quality Metrics ou MQM). Este passo envolve a avaliação e categorização de eventuais erros por um anotador, e os dados gerados são posteriormente utilizados para o aperfeiçoamento contínuo dos sistemas de QE.

2.2. Controlos de Qualidade

Tendo agora um entendimento mais profundo acerca dos métodos de tradução empregados pela Unbabel, torna-se pertinente uma análise detalhada dos mecanismos de controlo de qualidade implementados e do seu contexto operacional, oferecendo métodos automáticos e manuais de controlo de qualidade. Com efeito, a empresa recorre a diversos sistemas para assegurar a qualidade das traduções, sendo um deles o MQM. Estas métricas representam um sistema estruturado para apreciação da qualidade da tradução e

são versáteis o suficiente para serem aplicadas tanto à tradução humana quanto à automática[1].

Na Unbabel, a equipa de anotadores é incumbida de identificar e classificar todas as imprecisões conforme estabelecido pela tipologia de tradução da Unbabel[2], categorizando os erros em níveis: *neutral*, *minor*, *major* e *critical*. Posteriormente, com base na severidade dos erros identificados e na extensão do texto traduzido, procede-se ao cálculo de uma pontuação, onde, por exemplo, uma tradução sem erros tem 100 ponto. Consequentemente, quantos mais erros a tradução possuir, mais a pontuação vai diminuindo.

2.2.1. Anotação

Conforme mencionado anteriormente, o sistema MQM não opera exclusivamente com tecnologias de inteligência artificial e aprendizagem *machine learning*; este está intrinsecamente associado à atuação de anotadores humanos, cujo papel é fundamental para aferir a qualidade das avaliações e, por conseguinte, das traduções em si. Todas as anotações que contribuem para aprimorar a precisão dos dados de tradução são efetuadas numa plataforma dedicada, denominada *Annotation Tool*. Estas anotações são feitas por linguistas nativos da língua alvo, onde existe uma seleção de categorias correspondentes aos erros encontrados.

O seu trabalho reveste-se de uma importância fundamental, visto que procede não apenas à avaliação e classificação de tipologias de erros decorrentes de traduções efetuadas por sistemas de tradução automática, mas também incide sobre a análise das intervenções realizadas pelos editores integrantes da equipa da Community (Unbabel, n.d. a).

2.2.2. Edição

Os editores constituem os profissionais responsáveis pela edição e revisão das traduções geradas por sistemas de tradução automática e outros tradutores. Em virtude da

[1] ver: https://themqm.org/

[2] para mais detalhes ver: https://help.unbabel.com/hc/en-us/articles/6444304419479-Annotation-Guidelines-Typology-3-0

relevância da sua função na edição e revisão dos textos, a manutenção da qualidade das traduções é assegurada não só por meio de um processo de seleção criterioso, mas também através da avaliação dos profissionais de forma contínua (Unbabel, n.d. b) e das anotações anteriormente mencionadas.

Aquando da manifestação de interesse em assumir funções de editor na Unbabel, após a inscrição na plataforma da empresa, e na eventualidade de o par de línguas desejado se encontrar disponível, o candidato é elegível para proceder à realização de testes de línguas. Tais testes englobam 12 questões de escolha múltipla, solicitando ao candidato a seleção da edição apropriada para um determinado segmento. Estes exames visam assegurar o cumprimento dos padrões de proficiência e qualidade exigidos pela companhia. Após esta fase, o candidato avança para a etapa subsequente, que implica uma introdução à interface da plataforma e a execução de um teste de aptidões. Este último consiste em cinco tarefas de edição, as quais, após serem finalizadas, são submetidas a avaliadores para apreciação. Uma performance considerada satisfatória neste teste de competências conduz à promoção do indivíduo ao cargo de editor (Unbabel, n.d. b).

Para assegurar uma constante apreciação da qualidade, os candidatos, uma vez promovidos a editores, o que implica a receção de uma remuneração pelo seu trabalho, são submetidos a um regime de avaliações regulares. Este processo visa garantir a manutenção dos padrões de qualidade exigidos. Os testes aplicados são selecionados ou designados de forma aleatória. Na eventualidade de um editor manter um desempenho exemplar e níveis de qualidade elevados durante estas avaliações periódicas, existe a possibilidade de ser promovido a revisor. Por outro lado, resultados insatisfatórios ou uma adequação insuficiente aos padrões de qualidade nas avaliações recorrentes podem resultar na reclassificação da posição do editor e, por sua vez perde o acesso a tarefas remuneradas. As referidas avaliações periódicas são avaliadas por especialistas, representando oportunidades para os editores refletirem sobre e aprenderem com erros anteriormente cometidos. É também uma das razões que tais avaliações são realizadas com regularidade.

2.2.3. Revisão

Quando os editores ascendem à categoria de revisores, não lhes são confiadas apenas tarefas de pós-edição, mas também a responsabilidade de efetuar revisões. Usualmente, este tipo de incumbência destina-se a textos com padrões de qualidade mais elevados, incluindo textos de natureza especializada.

A função do revisor consiste em analisar e aprimorar o trabalho realizado pelos editores, o que implica cumprir uma série de normas linguísticas e proceder à comparação entre o texto-fonte e o texto-alvo, aplicando as modificações necessárias para assegurar os mais elevados padrões de qualidade. Entre os vários aspectos aos quais os revisores precisam dedicar especial atenção no decorrer da revisão de um texto, destacam-se:

- O cumprimento da terminologia apropriada, especialmente relevante no caso de textos que utilizam terminologia especializada, e, adicionalmente o cumprimentos dos requisitos particulares de cada cliente;
- A fluência do texto, que deve refletir um caráter "nativo", assegurando que o mesmo seja percetível e natural para o leitor;
- A consistência ao longo do texto, uma vez que, dada a intervenção de diversos editores em segmentos diferentes do texto, podem surgir discrepâncias. Cabe ao revisor garantir que o texto mantenha uma coerência e uniformidade em toda a sua extensão (Unbabel, n.d. d).

Este conjunto de responsabilidades sublinha a importância do papel desempenhado pelos revisores no processo de garantia da qualidade textual, potenciando assim a entrega de conteúdos que cumpram rigorosamente os requisitos estabelecidos pelo cliente (Unbabel, n.d. d).

2.4. Objetivos do estágio

Tendo em consideração o exposto, emergem indagações acerca da integração desta experiência de estágio no contexto da empresa. Efetivamente, até ao advento da recente aquisição de Lingo24 e EVS pela Unbabel, a diferenciação das variantes do inglês não constituía uma exigência. O critério prevalecente para tradutores, anotadores e editores residia apenas na adesão consistente à variante do inglês selecionada, em

alinhamento com as diretrizes linguísticas estabelecidas pela empresa. No decorrer deste estágio estas diretrizes foram retiradas e substituídas por diretrizes específicas de inglês britânico e inglês americano. É importante notar que esta seleção das variantes era puramente arbitrária, sendo apenas exigida consistência linguística.

Com a integração dos clientes oriundos da Lingo24 e EVS, emergiu na Unbabel a necessidade de reconhecer o inglês britânico como uma variante linguística específica. Consequentemente, tornou-se imperativo encarar esta variante não apenas como parte dos pares de línguas para tradução, mas também para a localização intralinguística. Este contexto delineou o projeto cuja finalidade se centrou em capacitar os editores para o conhecimento adequado do inglês britânico. Dada a composição plural do corpo de anotadores, editores e revisores da Unbabel, que englobava indivíduos de distintas proveniências linguísticas, onde alguns não eram falantes nativos de inglês e aqueles cuja língua materna é o inglês estão dispersos por diversas nações anglófonas, a formação especializada mostrou-se essencial.

O presente estudo concentra-se exclusivamente na questão supracitada, buscando explorar métodos escaláveis para treinar falantes nativos de diferentes variantes do inglês na variante britânica.

3. A EXPLORAÇÃO DE PLN, VARIAÇÕES LINGUÍSTICAS E ENSINO DE ADULTOS: UMA REVISÃO DA LITERATURA

O presente capítulo constitui uma exposição da literatura sobre diversos tópicos abordados neste relatório com uma breve descrição da história de NLP (secção 3.1), como surgiu o GPT e o que é (secção 3.2), os princípios da engenharia *prompt* (secção 3.3), o impacto da IA em contextos educacionais (secção 3.4), uma breve introdução à andragogia (secção 3.5) e, finalmente uma compreensão sobre as características do inglês britânico e americano (secção 3.6).

Este relatório abrange uma ampla gama de tópicos, tornando complexa a tarefa de apresentar uma revisão exaustiva da literatura necessária para este estudo, no entanto, esta é uma tentativa de abordar a maioria desses tópicos.

3.1. Nas Raízes da Inteligência: O Alvor da Compreensão Máquina-Linguagem

No âmbito da discussão sobre IA, torna-se imperativo abordar os seus primórdios e traçar a sua evolução até à atualidade, destacando, pelo menos, os marcos mais significativos para a devida contextualização deste estudo.

Assim, para compreendermos a história da IA tal como a conhecemos hoje, é essencial começar por explorar o campo do Processamento da Linguagem Natural (PLN) ou linguística computacional. Atualmente, o PLN pode ser definido como uma disciplina da engenharia dedicada a desenvolver máquinas capazes de manipular ou replicar a linguagem humana, seja ela escrita, oral ou estruturada de maneira similar à linguagem humana (DeepLearning.AI, 2023).

3.1.1. Primeira era

Atendendo à vinculação intrínseca desta disciplina com a linguagem, não é de estranhar que, nas suas origens, tenha emergido precisamente no campo da tradução. De facto, é possível identificar quatro eras distintas que demarcaram a evolução do PLN (Manning, 2022)**.** A primeira era estende-se dos anos 1950 a 1969, iniciando-se, com efeito, com o desenvolvimento da tradução automática. Nas conceções iniciais, considerava-se que a tradução poderia facilitar a decodificação de mensagens durante a Segunda Guerra Mundial, momento este que foi crucial para o avanço da computação. Ademais, durante a Guerra Fria, as partes envolvidas no conflito desenvolveram sistemas capazes de traduzir os trabalhos científicos de outras nações (Manning, 2022).

Adicionalmente, em 1957 Chomsky publicou a sua obra *Syntactic Structures* (Chomsky, 1957) que revolucionou o conhecimento linguístico da época, dado que até então não havia maneira de incorporar gramática nas máquinas de maneira que compreendessem o seu significado (Shrivastava, & Jain, 2020). Conclui que para um

computador poder compreender uma língua, a estrutura da frase teria de ser alterada. Com isto, introduziu uma gramática gerativa denominada *Phase Structure-Grammar*, que envolvia a tradução de linguagem natural para um formato adaptado à compreensão dos computadores.

Contudo, conforme se poderia antecipar, a quantidade de informação disponível sobre a linguagem humana, bem como o estudo e compreensão desta, era incrivelmente limitada na época. Os primeiros sistemas de tradução automática realizavam traduções palavra por palavra, mostrando-se incapazes de lidar com princípios linguísticos mais complexos. De facto, os progressos em tradução automática sofreram um revés quase terminal em 1966 com a publicação do relatório *Automatic Language Processing Advisory Committee* (ALPAC), que recomendava o cessar do financiamento destinado à investigação em tradução automática (Jones, 1994).

Ainda assim, este período, apesar das suas óbvias falhas, foi muito importante nesta área, como afirma Jones:

> But it is important to recognise what these first NLP workers did achieve. They recognised, and attempted to meet, the requirements of computational language processing, particularly in relation to syntactic analysis, and indeed successfully parsed and characterised sentences. They investigated many aspects of language, like polysemy, and of processing, including generation. They addressed the issues of overall system architectures and processing strategies, for example in direct, interlingual or transfer translation. They began to develop formalisms and tools, and some influential ideas first appeared, like the use of logic for representation.
>
> Jones, 1994, p. 6

3.1.2. Segunda Era

A segunda era ocorreu de 1970 a 1992. Este período foi marcado pela origem de uma série de sistemas de PLN, nomeadamente: o SHRDLU de Terry Winograd, o LUNAR de Bill Woods, os sistemas de Roger Schank, como o SAM, os de Gary Hendrix, como o LIFER, e o GUS de Danny Bobrow. Estes sistemas destacaram-se pela sua capacidade de lidar com aspetos complexos da linguagem, tais como a sintaxe. Foram

desenvolvidos manualmente e representaram uma tentativa de modelar e aplicar a complexidade inerente à compreensão da linguagem humana. (Manning, 2022).

Algumas destas soluções foram implementadas em contextos práticos, como por exemplo, nas consultas a bases de dados (Harris, 1979 citado em Manning, 2022). Este desenvolvimento destacou a viabilidade e a utilidade potencial do PLN em contextos reais. Paralelamente, verificou-se também um avanço significativo no campo das teorias linguísticas, o que possibilitou estabelecer uma clara distinção entre o conhecimento linguístico de natureza declarativa e os processos procedimentais da sua manipulação. Importa realçar que esta distinção constituiu um marco relevante, uma vez que promoveu uma abordagem sistemática e eficiente na conceção e implementação destes sistemas. (Manning, 2022).

3.1.3. Terceira Era

Com a chegada da Internet e a disponibilização de texto em formato digital, tudo se transformou na terceira era, que se estendeu de 1993 a 2012. Neste período, tornou-se crucial focar nos algoritmos e na exploração da compreensão da linguagem escrita. (Manning, 2022). Na década de 90, estimulou-se a pesquisa em *corpora* anotados, "(...) ou seja, conjuntos de textos sobre um domínio de conhecimento onde cada uma das suas palavras foram identificadas segundo sua função sintática." (Vieira & Lopes, 2010).

Foi igualmente durante este período que os modelos estatísticos adquiriram popularidade (Foote, 2023). Assim, a combinação das abordagens simbólicas, fundamentadas em conceitos linguísticos, com os modelos estatísticos, conduziu a resultados superiormente eficazes (Manning e Shütze, 1999 citado em Vieira & Lopes, 2010). No ano de 1997, foram introduzidas as RNN, que continuam a ser consideradas, até aos dias de hoje, um marco relevante na pesquisa em PLN (Foote, 2023).

Em 2011, sistemas de assistência por voz, tal como a Siri da Apple, assinalaram o surgimento dos primeiros sistemas de PLN de maior sucesso. No seu artigo "*Advances in Natural Language Processing*", Hirschberg e Manning atribuem estes avanços significativos na linguística computacional a:

> Four key factors enabled these developments: (i) a vast increase in computing power, (ii) the availability of very large amounts of linguistic data, (iii) the development of highly successful machine learning (ML) methods, and (iv) a much richer understanding of the structure of human language and its deployment in social contexts.
>
> Hirschberg & Manning (2015)

3.2. Os primeiros passos do futuro: Desenvolvimento da arquitetura do Transformer

Desde 2013 até à atualidade, os desenvolvimentos em PLN sofreram uma considerável reorientação com a introdução do *deep learning* ou dos métodos neurais artificiais (Manning, 2022), os quais, por sua vez, pavimentaram o caminho para o desenvolvimento do que hoje conhecemos como GPT.

Apesar de a fama dos Large Language Models (LLM) ter crescido sobretudo em 2020 com o lançamento público do ChatGPT, o seu desenvolvimento teve início anos antes. Em 2017, foi publicado o artigo "*Attention is All You Need*" (Vaswani et al). De facto, este artigo impulsionou em grande parte o desenvolvimento dos LLM, uma vez que foi o primeiro a abordar o modelo *Transformer*. Este modelo se tornou um dos artigos mais importantes para fornecer contextualização sobre os LLM. O modelo Transformer possui uma arquitetura de *deep learning* que não requer as camadas utilizadas por modelos precedentes. Depende exclusivamente de mecanismos de atenção.

Por conseguinte, o modelo Transformer destacou-se dos modelos anteriores, uma vez que a sua arquitetura manipula dados sequenciais, como por exemplo, a linguagem, de forma mais eficiente do que as Redes Neurais Recorrentes (RNN) e as Redes Neurais Convolucionais (CNN), alcançando um desempenho *state of the art* em tarefas de tradução automática sem recorrer a quaisquer camadas recorrentes. No entanto, o que revolucionou este modelo foram os mecanismos de autoatenção. Estes mecanismos permitiam que o modelo avaliasse a importância de diferentes partes de uma sequência de entrada enquanto processava cada palavra.

A *Neural Machine Translation*, com sua arquitetura de codificador-decodificador, apresentava limitações ao tentar traduzir sequências longas, pois apresentava dificuldade

em reter informações dos primeiros elementos da sequência. Tal ocorria porque o decodificador só tinha acesso ao estado oculto final do codificador, perdendo, assim, informações relevantes apresentadas no início da sequência.

O *Transformer* distingue-se dos outros modelos por não processar os dados sequencialmente, ou seja, um texto não é tratado de forma tradicional, que implicaria o processamento de partes de dados uma de cada vez, ou palavra por palavra. Em vez disso, processa a totalidade da informação em simultâneo através de uma técnica de processamento paralelo. Graças ao mecanismo de autoatenção, o *Transformer* permite que cada posição no codificador aceda a todas as posições na camada anterior do descodificador. Isto possibilita que o modelo considere a totalidade da sequência ao processar cada palavra, contribuindo significativamente para a compreensão do significado semântico das frases.

Outro mecanismo de atenção presente no *Transformer* é a atenção *multi-head*. Este método permite a execução simultânea de múltiplos processos de autoatenção, avaliando diferentes pontuações de atenção para "cabeças" diferenciadas. Este procedimento habilita o modelo a considerar diversos aspetos das relações semânticas, incluindo o contexto e as informações sintáticas.

Ao mesmo tempo, o modelo é composto por uma estrutura de camadas idênticas, cada uma contendo duas subcamadas: o mecanismo de autoatenção com atenção *multi-head* e uma rede *feed-forward*. Todas estas camadas são interligadas por conexões residuais, que por sua vez são seguidas de normalização de camadas. Todos estes mecanismos conferem ao *Transformer* uma grande eficiência. Contudo, para gerar saídas de elevada qualidade, é também necessário considerar a ordem das palavras. Para tal, os autores introduziram o conceito de "codificação posicional" (positional encoding), um método para incorporar informação sobre a posição de cada palavra numa frase no conjunto de dados processado pelo modelo. Isso permite que o modelo leve em conta a posição das palavras ao realizar previsões. Confira-se a seguinte figura 1 ilustração da arquitetura:

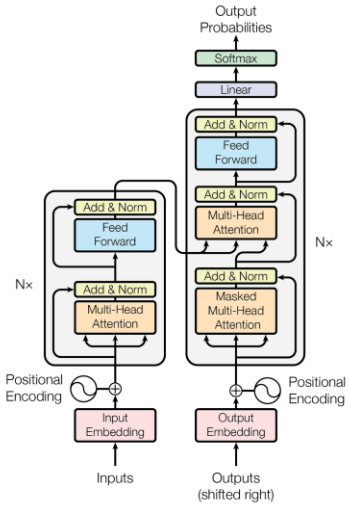


Figura 1 – Arquitetura do *Transformer*

Vaswani et al, 2017

Este artigo representou um marco no campo do Processamento de Linguagem Natural (PLN), não só ao estabelecer um ponto de partida para o desenvolvimento de outros modelos, mas também pelos seus notáveis resultados na área da tradução.

> For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. On both WMT 2014 English-to-German and WMT 2014 English-to-French translation tasks, we achieve a new state of the art. In the former task our best model outperforms even all previously reported ensembles.
>
> Vaswani et al, 2017, p.10

De facto, um dos desenvolvimentos gerados por esse estudo foi o modelo de inteligência artificial GPT (Generative Pre-trained Transformer). Considerando que o foco principal deste relatório recai sobre o modelo GPT-4, torna-se imperativo fornecer uma contextualização da sua evolução histórica. A OpenAI alcançou notoriedade junto do público com o lançamento do ChatGPT, um *chatbot* avançado e gerador de texto, que foi disponibilizado pela empresa no dia 30 de novembro de 2022. Este modelo faculta aos utilizadores a capacidade de interagir com um modelo de LLM, oferecendo a flexibilidade para personalizar o tamanho da resposta, o formato, o estilo de escrita, o idioma, entre outros parâmetros, com respostas semelhantes às de um humano.

Desde então, a OpenAI tem continuado o desenvolvimento de novos modelos, sendo o GPT-4 um dos mais recentes e avançados. Este modelo, semelhante ao GPT-3, no que se refere à interatividade em conversação, distingue-se pelo facto de que a sua versão de chat pode ser acessada tanto mediante um serviço do *chatbot* pago como através da API (Application Programming Interface) da OpenAI (OpenAI, n.d. a).

Este modelo do GPT ultrapassou o desempenho do seu predecessor, o GPT-3.5-turbo, em exames, alcançando um lugar entre os primeiros 10% no que toca a pontuações, em contraste com o GPT-3.5, que se situou na faixa dos últimos 10% (OpenAI, n.d. b). Estes exames e testes *benchmark*, originalmente concebidos para avaliar a competência humana, incorporaram questões provenientes de Olimpíadas Científicas. O GPT-4 demarca-se do modelo anterior não só pela maior consistência e criatividade, mas também pela sua capacidade reforçada de abordar questões de maior complexidade (OpenAI, 2023).

3.3. Engenharia *Prompt*

Tendo já clarificado o percurso histórico da IA, torna-se necessário compreender claramente o que se designa por "*prompt*". Importa igualmente entender o que constitui a área de engenharia de *prompts*, o aperfeiçoamento (conhecido em inglês por "*fine-tuning*"), assim como todos os aspetos relacionados que se revelaram pertinentes para a condução desta investigação, dado que se trata de uma disciplina/área essencial e central.

Conforme anteriormente observado, os modelos de IA incluem interfaces de chat, permitindo aos utilizadores uma experiência bastante imersiva e interativa com estas tecnologias. Assim, não é segredo que modelos de LLM pré-treinados adquirem conhecimentos linguísticos fundamentais (Liu et al., 2019; Amrami and Goldberg, 2018 citado em Sorensen et al., 2022) necessários para a execução de tarefas e conhecimento do mundo factual (Petroni et al., 2020; Bosselut et al.; Bouraoui et al.; Zuo et al., 2018 citado em Sorensen et al., 2022). Por essa razão, estes modelos são capazes de responder a perguntas colocadas pelos utilizadores ou de realizar outras tarefas relacionadas com o processamento de linguagem sem a necessidade de procurar informações *online* (Radford et al., 2019; Devlin et al., 2019; Raffel et al., 2019 citado em Sorensen et al., 2022).

Efetivamente, os modelos de IA mais avançados conseguem aprender rapidamente a desempenhar tarefas quando lhes são fornecidos exemplos ou instruções em linguagem natural claras e compreensíveis também para humanos (Brown et al., 2020 citado em Sorensen et al., 2022).

Confira-se portanto a seguinte definição de *prompt*:

> Prompt engineering is a set of techniques and methods to design, write, and optimize instructions for LLM, called prompts, such that the model answers will be precise, concrete, accurate, replicable, and factually correct [8,9,18]. Prompts are understood as a form of programming because they can customize the outputs and interactions with an LLM [9]. They involve adapting instructions in natural language, obtaining the desired responses, guaranteeing contextually accurate results, and increasing the usefulness of generative language models in various applications [13]. Its applications include fields such as medical education, radiology, and science education [11,12,34].
>
> Velásquez-Henao et al., 2023

Observou-se que o desempenho da IA sofre alterações significativas em função da qualidade das instruções fornecidas, sem que seja necessário alterar ou atualizar o treino do próprio modelo. Tais instruções, frequentemente designadas como *prompts*, caracterizam-se pela descrição das tarefas em linguagem escrita e/ou pela inclusão de exemplos ilustrativos (Brown et al., 2020 citado em Lester et al., 2021).

Prompts eficazes distinguem-se através da incorporação de princípios fundamentais, a saber: claridade e precisão, contextualização da informação, especificação do formato desejado e modulação da verbosidade (Lo., 2023 citado em Velásquez-Henao et al., 2023). Dada a sua importância, a elaboração de *prompts* configura-se como um processo intrinsecamente criativo, que exige não só intuição como também um contínuo refinamento interativo (Lo., 2023 citado em Velásquez-Henao et al., 2023). É crucial sublinhar que todos esses aspectos devem ser meticulosamente adaptados à tarefa específica solicitada à IA, enfatizando a necessidade de uma informação rigorosamente alinhada com o contexto (Velásquez-Henao et al., 2023).

Caso os *prompts* não observem os princípios anteriormente mencionados, os resultados obtidos poderão ser afetados, manifestando-se através de respostas vagas e

imprecisas, incorreções factuais ou inadequação contextual (Jha et al., 2023 citado em Velásquez-Henao et al., 2023).

3.3.1. Guia prático para engenharia *prompt*

Dada a limitação temporal associada a este estágio, a investigação em torno desta disciplina revelou-se sucinta, com um foco particular nos métodos práticos da engenharia de *prompts*, destacando-se, mais especificamente, o guia prático fornecido pela OpenAI (OpeanAI, n.d. c).

De acordo com as informações disponibilizadas neste guia, identificam-se seis estratégias primordiais destinadas a aprimorar a performance de sistemas baseados em IA, mais especificamente referentes ao modelo GPT-4.

1. A primeira estratégia consiste em fornecer instruções claras, análogo a como um ser humano mais eficientemente desempenha uma tarefa quando esta é explicada de forma detalhada e inequívoca. Da mesma forma, o mesmo ocorre com os modelos de IA. Estratégias para proporcionar instruções claras com o objetivo de melhorar o seu desempenho incluem solicitar ao modelo para adotar uma persona, usar delimitadores para enfatizar partes específicas do *prompt*, fornecer todos os passos necessários para a conclusão de uma tarefa, fornecer exemplos dessa mesma tarefa, e ainda, especificar a dimensão pretendida do resultado.
2. A inclusão de textos de referência revela-se vantajosa, tendo em conta que o GPT-4 pode gerar informações falsas e apresentá-las de forma convincente nas suas respostas. Por conseguinte, é benéfico proporcionar uma referência que o modelo possa consultar, de forma a fornecer respostas mais precisas e fundamentadas.
3. Segmentar tarefas de grande envergadura em sub-tarefas mais elementares e simples revela-se de crucial importância. Importa notar que tarefas de complexidade elevada tendem a suscitar um maior número de erros comparativamente às tarefas de menor complexidade. Deste modo, este método pode oferecer um roteiro mais claro e precisão aumentada para o modelo prosseguir.

4. Conceder tempo ao modelo para refletir pode ser benéfico. Solicitar ao modelo que desenvolva uma "linha de raciocínio" pode auxiliar na obtenção de conclusões próprias de forma mais robusta. Métodos para implementar esta prática incluem instar o modelo a deduzir conclusões independentes sem recorrer imediatamente a uma solução e questionar sobre possíveis lacunas informativas ou erros que possam não ter sido mencionados na resposta anterior.
5. Recorrer a ferramentas externas. Este método é aplicado com o intuito de suprir as limitações do modelo através da complementação com *outputs* gerados por outras ferramentas. Existem sistemas especificamente desenvolvidos para se destacarem em determinadas tarefas. Portanto, caso outro modelo ou software apresente um desempenho superior ao do GPT-4 numa tarefa específica, revela-se proveitoso incorporar esse *output* no *prompt*.
6. Implementar alterações de maneira sistemática. A justificativa para este método reside no princípio de que a otimização da performance só é alcançável através de processos que possam ser mensurados. Em determinadas circunstâncias, ajustes pontuais num *prompt* podem conduzir a melhorias notáveis em casos específicos, mas resultar em desempenho inferior quando aplicados a um conjunto mais amplo de exemplos. Assim, revela-se vantajoso realizar testes sistemáticos dos *outputs*. Este procedimento pode incluir a comparação dos resultados obtidos com respostas consideradas *Gold Standard*.

3.4. Inteligência Artificial na Sala de Aula: Uma nova era de aprendizagem com o Apoio da Tecnologia

É inegável que a IA tem permeado diversas disciplinas. No entanto, emerge como um tópico relevante de discussão e um campo de interesse questionar qual será o papel da IA na educação e de que forma poderá influenciá-la, dada a sua ampla gama de

utilidades (Kasneci et al., 2023). De facto, têm sido realizados estudos focados no suporte da IA ao ensino. Embora não esteja diretamente relacionado à elaboração de testes, o desempenho dos LLM em tarefas de índole educativa revela-se pertinente para a argumentação presente neste relatório.

Em determinadas circunstâncias, o emprego da IA para propósitos educacionais pode não ser recomendado, salientando-se possíveis desafios associados à sua utilização. Entre estes, destacam-se a insuficiência de recursos disponíveis, evidenciando-se frequentemente que a tecnologia, por si só, não é capaz de satisfazer as exigências das tarefas designadas. Acresce a isso a escassez de enquadramentos regulatórios adequados, a insuficiente proteção de dados pessoais, assim como a falta de competências digitais por parte dos utilizadores, entre outros aspectos relevantes (Jassen et al., 2021, citado em Tlili et al., 2023).

Por outro lado, a utilização da IA para fins educacionais desvenda um leque de oportunidades e novos modos de suporte e abordagens para os utilizadores. Househ et al. (2023) destacam que, no âmbito da educação médica, a IA de facto oferece múltiplas oportunidades para enriquecer o processo de aprendizagem. Entre estas, incluem-se planos personalizados e materiais didáticos adaptados, permitindo, mediante o reconhecimento dos pontos fortes e fracos dos alunos, a adaptação dos conteúdos de maneira mais individualizada às suas necessidades. Os autores adicionam ainda que o feedback obtido pode ser aplicado para refinamento dos *outputs* de LLM a longo prazo. A IA disponibiliza igualmente ferramentas de análise e avaliação, onde os LLM possam efetuar avaliações dos alunos, permitindo analisar seus pontos fortes e fracos, fornecendo feedback construtivo e, inclusive, sugerir métodos de avaliação.

Sistemas de IA podem também prestar apoio aos docentes na organização das aulas, através da introdução de conteúdos específicos da disciplina, no desenvolvimento de aulas, atividades e questões. Revelam-se igualmente valiosos na aprendizagem de línguas estrangeiras, oferecendo explicações gramaticais, resumos, sugestões de melhorias e facilitando a prática de conversação (Kasneci et al., 2023). Adicionalmente, constatou-se que o recurso a LLM pode contribuir para a redução da ansiedade associada à aprendizagem de línguas estrangeiras, fomentando a motivação dos estudantes e estimulando a resolução de problemas (El Shazly, 2021).

No entanto, segundo Househ et al. (2023) LLM demonstraram possuir limitações, conforme evidenciado em estudos acerca dos impactos da inteligência artificial na educação. Entre essas limitações, destaca-se a questão da fiabilidade ou a possibilidade de fornecerem informações inverídicas, uma vez que, apesar dos avanços tecnológicos, ainda existe a probabilidade de produzirem conteúdos falsos ou enganosos, particularmente devido ao seu estilo de comunicação autoritário e persuasivo. Outra lacuna observada nas LLM é a sua inconsistência: a capacidade de gerar respostas divergentes à mesma solicitação pode conduzir a variações na qualidade e a desafios relacionados com a credibilidade das informações prestadas. Esta problemática da fiabilidade pode ser atribuída, em parte, às suas bases de conhecimento restritas, tendo em vista que LLM são formadas com base em conjuntos de dados que podem conter informações desatualizadas e não especializadas (Househ et al., 2023).

Portanto, é admissível afirmar que a IA tem evidenciado um notável potencial como ferramenta de apoio ao setor educativo. No entanto, sublinha-se a necessidade de uma utilização cautelosa, recomendando-se, possivelmente, a elaboração e implementação de diretrizes para a sua integração efetiva nestes contextos (Tlili et al., 2023). Adicionalmente, torna-se imprescindível o conhecimento aprofundado destes sistemas. Tlili et al. (2023) argumentam que é fundamental proporcionar formação tanto a estudantes como a mentores no manejo destes modelos, capacitando-os para lidar eficazmente com os seus avanços. Isso implica não apenas uma compreensão dos aspectos operacionais, mas também das implicações destas tecnologias no ambiente educacional.

3.5. Andragogia

Dado que este relatório incide sobre o desenvolvimento de testes destinados ao treino e avaliação de tradutores em uma variante da língua inglesa da qual não são falantes nativos, é imperativo que os testes incorporem elementos educacionais de modo a contribuir ativamente para a aquisição de competências e conhecimentos de forma mais eficaz.

Neste capítulo, proceder-se-á à análise de alguns princípios de aprendizagem que podem revelar-se fundamentais na elaboração de testes didáticos para adultos.

Naturalmente, nem todos os princípios de aprendizagem serão aplicáveis, visto que o escopo da investigação se limita à elaboração de testes. Não obstante, subsistem teorias pertinentes que podem ser devidamente aproveitadas.

De facto, a psicologia tem investigado metodologias e princípios didáticos destinados aos adultos. Apesar de a psicologia ser uma ciência de surgimento relativamente recente, foram alcançados progressos significativos. Estes estudos revelaram-se cruciais para a formulação de princípios e teorias sobre a aprendizagem. Ormrod, na sua obra "*Human Learning*", discute a diferença entre princípios e teorias da aprendizagem.

Os princípios de aprendizagem são generalizações relativas a fatores que influenciam a aprendizagem, descrevendo os resultados específicos desses fatores. Ormrod (2011) exemplifica com o princípio segundo o qual comportamentos que são seguidos por recompensas tendem provavelmente a aumentar em frequência, um conceito amplamente aplicável a diferentes espécies, tipos de aprendizagem e variedades de recompensas. Este princípio é também conhecido como reforço, ou mais especificamente, reforço positivo. Desta forma, os princípios oferecem padrões estáveis e observáveis através de diversos contextos (Ormrod, 2011, p. 5).

As teorias de aprendizagem, por sua vez, examinam os mecanismos subjacentes à aprendizagem e explicam a razão pela qual certos fatores são importantes. A título de exemplo, a teoria social cognitiva propõe que as recompensas potenciam a aprendizagem ao encorajar a atenção sobre a informação que está a ser assimilada, oferecendo compreensão acerca das razões pelas quais as recompensas influenciam a aprendizagem. Deste modo, entende-se que as teorias se desenvolvem ao longo do tempo, à medida que novas investigações são realizadas, fornecendo explicações mais aprofundadas acerca dos princípios observados (Ormrod, 2011, p. 5).

Foi no início da década de 70 que a aprendizagem adulta foi, pela primeira vez, diferenciada da aprendizagem direcionada a crianças (pedagogia). Esta distinção foi estabelecida por Malcolm Knowles (1998). Introduzindo o termo andragogia, a sua proposta foi considerada revolucionária, suscitando debates acerca da sua definição: "*It has been alternately described as a set of guidelines (Merriam, 1993), a philosophy (Pratt, 1993), a set of assumptions (Brookfield, 1986), and a theory (Knowles, 1989).*" (Knowles et al., 1998).

Knowles (1998) propôs uma série de princípios ou premissas do modelo andragógico:

1. Necessidade de saber
2. Conceito de si
3. Papel da experiência
4. Vontade de aprender
5. Orientação da aprendizagem
6. Motivação[3]

Esta diferenciação é de suma importância, uma vez que as necessidades dos adultos no que concerne à aprendizagem diferem substancialmente do modelo tradicionalmente utilizado para crianças. Tal distinção deve ser considerada meticulosamente na conceção de testes.

É imperativo dar uma atenção especial, sobretudo à premissa do "papel da experiência". Knowles argumenta que os adultos aprendem melhor através da experiência e sublinha a importância de integrar as vivências pessoais no ensino dirigido a adultos. Além disso, no que concerne à "vontade de aprender" (do inglês, "*readiness to learn*"), o autor defende que esta vontade está intrinsecamente relacionada com a necessidade de compreender o motivo pelo qual a aprendizagem é relevante, e a capacidade de aplicar os conhecimentos adquiridos em situações reais da vida:

> Readiness to learn. Adults become ready to learn those things they need to know and be able to do in order to cope effectively with their real-life situations. An especially rich source of “readiness to learn” is the developmental tasks associated with moving from one developmental stage to the next. The critical implication of this assumption is the importance of timing learning experiences to coincide with those developmental tasks. For example, a sophomore girl in high school is not ready to learn about infant nutrition or marital relations, but
> let her get engaged after graduation and she will be very ready.
>
> Knowles et al., 1998, p.67

[3] Tradução retirada de: *A Andragogia: Nova arte de formação* (n.d.) https://repositorio.ul.pt/bitstream/10451/2840/19/ulfp037529_tm_anexo16_A%20Andragogia%2C%20Nova%20Arte%20de%20Forma%C3%A7%C3%A3o.pdf

Uma consequência crucial desta ideia é a relevância de alinhar as experiências de aprendizagem com tarefas de desenvolvimento pessoal e profissional. Adicionalmente, Knowles aponta que existem métodos para estimular essa vontade, através de, por exemplo, orientação profissional.

Para a elaboração de testes destinados a tradutores internos, isso implica que pode ser vantajoso customizar os testes de modo a que reflitam experiências reais que os tradutores possam vir a enfrentar ou já tenham enfrentado em contextos profissionais. Isto é, embora seja relevante, para os testes de línguas, incluir vários contextos e, para formas de seleção ou mesmo na monitorização de profissionais, abordar diferentes contextos linguísticos, com estas noções e ideias propostas, pode-se inferir que para alcançar melhores resultados, essa formação das variantes (ou de qualquer outro par de línguas) com exemplos de tarefas realizadas na empresa ou resolução de problemas adaptados contextualmente à empresa seria mais eficaz.

Deste modo, pode-se assumir que se as pessoas que realizam os testes estiverem cientes de que estes são contextualmente relevantes, haverá um maior interesse pessoal e, consequentemente, uma melhor retenção dos conteúdos, dado que compreendem que, inevitavelmente, irão aplicar essa aprendizagem na prática.

3.6. Inglês Britânico e as suas Características Linguísticas

O inglês britânico distingue-se por possuir ortografia, vocabulário e convenções gramaticais específicas, diferenciando-se assim dos demais dialetos da língua inglesa. Este capítulo dedica-se a explorar algumas das diferenças mais marcantes entre o inglês britânico e o inglês americano.

3.6.1. Contexto histórico

Para obter uma compreensão aprofundada das diferenças linguísticas do inglês britânico, é crucial considerar o seu contexto histórico. O inglês, pertencente à família das línguas germânicas, foi introduzido nas Ilhas Britânicas durante os séculos V e VI pelos

Anglos, Saxões e Jutos. Ao longo das gerações, a língua evoluiu, sofrendo alterações significativas (Millward, 1988, p.66).

Millward (1988) afirma que um marco "cataclísmico" na história da língua inglesa foi a Invasão Normanda de 1066, que representou uma vitória francesa, sobretudo do ponto de vista linguístico. Este acontecimento introduziu na Inglaterra uma vasta quantidade de vocabulário de origem francesa. Além disso, a presença normanda influenciou diversos aspetos do país, incluindo política, economia, religião e cultura. Contudo, este evento marcou o início de um período bilingue, durante o qual uma considerável parcela da população se expressava tanto em inglês quanto em francês (Millward, 1988, p.121). A língua inglesa foi enriquecida com palavras francesas como "*tax, estate, trouble, duty, and pay. English household servants would have learned French words like table, boil, serve, roast, and dine.*" (Millward, 1988, p.122).

A imprensa desempenhou um papel fundamental na disseminação e padronização do dialeto de Londres. Considerando que a maioria das tipografias se localizava na capital, a maioria dos livros publicados empregava esse dialeto. Este fato contribuiu significativamente para a sua adoção em todo o território inglês (Millward, 1988, p.124).

Em 1558, subsequente à derrota da Armada Espanhola, a Inglaterra emergiu como uma potência marítima dominante na Europa. Tal estatuto propiciou a conquista de colónias, incluindo aquela que viria a ser conhecida como Estados Unidos da América (Millward, 1988, p.195). De facto, a modalidade de inglês falado durante o período das explorações e colonização desempenhou um papel crucial no desenvolvimento do Inglês Americano tal como é reconhecido atualmente. Este processo de evolução representa o culminar de múltiplos componentes linguísticos, influenciado por diversas vagas de imigração para a América do Norte, como por exemplo a imigração dos *Scots-Irish* em 1740 (Bailey, 2012, p. 93) , e da africana em 1619 (Bailey, 2012, p. 23).

No entanto, quando estas novas vagas de imigração chegaram ao continente, já existia um *koine* americano distinto. Os imigrantes que chegaram posteriormente integraram-se neste ambiente linguístico, enquanto, em paralelo, conservavam os seus próprios dialetos em contextos privados. Em consequência, quando a Declaração da Independência foi assinada, em 1776, já estava estabelecido nos EUA um dialeto nacional único. Este fenómeno resultou de um processo multicultural e transgeracional, marcando

assim a sua diferenciação em relação ao inglês britânico e delineando uma clara identidade linguística (Luu, 2022).

3.6.2. Diferenças linguísticas

Embora a globalização esteja a contribuir para que as diferenças entre o inglês britânico e o inglês americano se tornem mais subtis, é relevante explorar as divergências existentes entre estas variantes linguísticas. Para tal, selecionaram-se duas fontes principais, as quais analisam aspetos cruciais da linguística inglesa: D'Amelio (2021), que apresenta informações sobre diferenças lexicais e ortográficas entre as duas variantes; e Akan (2017), que efetua uma análise das diferenças gramaticais entre o Inglês Britânico e o Americano.

Neste contexto, propõe-se explorar e sintetizar estas informações, com o objetivo de compreender melhor as nuances linguísticas destas variantes.

3.6.2.1. Diferenças Lexicais

D'Amelio (2021) aborda diferenças notáveis entre as duas variantes linguísticas, as quais decorrem das reformas propostas por Webster. Tais alterações visavam simplificar a grafia das palavras para que estas espelhassem mais fielmente a pronúncia, por exemplo, a transformação de "*musick*" para "*music*", ou a alteração da terminação do inglês britânico "-ou" para "-o" no inglês americano, uma mudança ainda evidente hoje em palavras como "*colour*" e "*color*", ou "*flavour*" e "*flavor*" (Lahey, 2000 citado em D'Amelio, 2021). Outras modificações incluem a substituição de "*-re*" por "*-er*", como nas palavras "*theatre*" e "*theater*"; a preferência por "*-ze*" em lugar de "*-se*", em palavras como "*realize*"; bem como a alteração das terminações de "*-gue*" para "-g" em palavras como "analogue" no inglês britânico para "analog" no inglês americano.

Contudo, a autora também sublinha que determinados termos adquiriram uma complexidade acrescida no inglês americano, exemplificado pelo uso duplicado do "I" em palavras como "*enrol*" no inglês britânico e "*enroll*" no inglês americano. Apesar das diversas variações mencionadas, existem palavras que se mantiveram inalteradas em ambas as variantes, como é o caso de "*advise*".

D’Amelio (2021) explora ainda as divergências vocabulares que transcendem a mera forma escrita, sobretudo em domínios como o transporte, as compras e a alimentação, onde se destacam vários exemplos que ilustram a representação de um mesmo objeto ou conceito através de dois termos distintos. Entre estes, destacam-se pares lexicais como “*cookie*” e “*biscuit*”; “*subway*” e “*underground*”; “*candy*” e “*sweets*”; “*cotton candy*” e “*candy floss*”, entre outros.

3.6.2.2. Diferenças Gramaticais

O ensaio de Akan (2017) proporciona uma análise aprofundada das diferenças gramaticais existentes entre o Inglês Britânico e o Americano, contemplando uma vasta gama de aspetos linguísticos, incluindo substantivos, verbos, pronomes, artigos, pontuação e convenções de pontuação e datação, entre outros. Tal análise destaca a complexidade intrínseca a ambas as variantes linguísticas.

Nomes e concordância verbal

Uma diferença significativa abordada por Akan (2017) relaciona-se com a concordância sujeito-verbo, especialmente no que diz respeito a nomes coletivos. O inglês britânico demonstra uma flexibilidade que permite a utilização de verbos tanto no singular quanto no plural, dependendo da perceção da unidade ou da pluralidade do grupo (Modiano, 1996; Quirk et al., 1985 citado em Akan, 2017). Por outro lado, no inglês americano, verifica-se uma consistência no uso de um verbo no singular com nomes coletivos. A título de exemplo no inglês britânico::

“*My team is winning. Vs. His team are all sitting down.*”

Enquanto no inglês americano:

“Which team is losing?

**Which team are losing?”*

(Akan, 2017)

Genitivos

Relativamente às construções possessivas, ambas as variantes linguísticas, tipicamente, recorrem ao genitivo com “s” para designar posse em nomes animados. No

entanto, Akan (2017) assinala um crescimento na preferência pelo uso do genitivo com "s" também com nomes abstratos no inglês americano (Hundt, 1997 citado em Akan, 2017).

Artigos

Registam-se algumas diferenças notáveis no uso dos artigos: no inglês britânico, observa-se a utilização de "a" e "an" dependendo do som inicial da palavra seguinte, enquanto que, em contextos informais do inglês americano, pode ocorrer o uso de "a" mesmo perante palavras que iniciam com som de vogal. (Tottie, 2002 citado em Akan, 2017).

Pronomes

As diferenças no uso de pronomes manifestam-se numa marcada preferência pelo pronome indefinido "one" na mesma frase no inglês britânico, ao passo que, no inglês americano, pode-se optar pela substituição por "he", evidenciando, assim, uma dicotomia entre padrões de uso formal e informal. Considere-se os seguintes exemplos:

"If one loses one's temper, one should apologize." - Inglês Britânico
"If one loses his temper, he should apologize." - Inglês Americano

(Akan, 2017)

Verbos e Modalidade

Akan (2017) aborda os diferentes usos dos verbos auxiliares modais, observando que "*shall*" e "*should*" são empregados com maior frequência no inglês britânico, frequentemente refletindo um nível de formalidade ou convites de ação, em contraste com "*will*" e "*would*", preferidos no inglês americano. A título de exemplo:

"Inglês Britânico: I shall go. (will- usado maioritariamente em inglês falado)
Inglês Americano: I will go. (will- usado em inglês escrito e falado)"

(Akan, 2017)

No que diz respeito à expressão de eventos futuros, o inglês britânico evidencia uma utilização mais frequente de "*be going to*" em contextos informais, enquanto que no inglês americano é comum a ocorrência da forma contraída "*gonna*". Akan (2017)

observa que, no tocante às formas do particípio passado, o inglês americano tende a adotar "*get*", ao passo que "*gotten*" é mais usual no inglês britânico.

Adicionalmente, indica-se uma diferença no uso das terminações verbais: "t" no inglês britânico em contraste com "ed" no inglês americano, uma distinção que influencia tanto a pronúncia quanto a ortografia, exemplificada em "*burnt*" *versus* "*burned*". Outros verbos que exemplificam este fenómeno incluem "*dream*", "*learn*" e "*spoil*".

O autor também sublinha que o inglês britânico recorre com frequência ao tempo presente perfeito para referir ações passadas que têm impacto no presente, ao passo que no inglês americano observa-se uma tendência para o uso do passado simples ou do presente perfeito, conforme o contexto específico, como exemplificado a seguir:

> *"Inglês Britânico- Najin feels ill. He's worked a lot.*
> *Inglês Americano- Najin feels ill. He worked a lot. ou Najin feels ill. He has worked a lot."*
>
> (Akan, 2017)

Pontuação e datas

Observam-se igualmente divergências em algumas normas de pontuação, nomeadamente no uso da vírgula: no inglês britânico, geralmente não se utiliza vírgula antes do último item de uma lista (conhecida como Oxford Comma). No que se refere a abreviações e acrónimos, na Inglaterra, costuma-se adicionar um ponto final após tais elementos, como por exemplo: "Dr.", "Mr.", "Mrs.", etc. Em contraste, o inglês americano faz uso da vírgula de forma típica após estas abreviações: "Dr,", "Mr,", "Mrs,", etc.

Akan (2017) também ressalta que o uso do hífen frequentemente difere entre as variantes, com o inglês britânico a recorrer ao hífen para unir prefixos ao núcleo da palavra, como em "*post-war*" ou "*co-operation*". Por outro lado, o inglês americano tende a omitir o hífen, resultando em formas como "*postwar*" ou "*cooperation*". As convenções para a apresentação de datas também divergem, sendo a preferência britânica pelo formato "dia-mês-ano", enquanto a variante americana favorece a sequência "mês-dia-ano".

Esta exposição das características linguísticas ilustra algumas das várias semelhanças entre ambas as variantes do inglês, evidenciando a sua versatilidade e adaptabilidade. Demonstrando ainda que ambas as variantes coexistem, oferecem um portal para observar a complexidade da língua inglesa e a forma como esta acomoda diferentes modalidades de comunicação.

3. METODOLOGIA

Neste capítulo proceder-se-á à descrição detalhada da metodologia adotada no âmbito do projeto de estágio. Tem-se como finalidade proporcionar uma compreensão abrangente acerca dos métodos empregues que conduziram aos resultados elencados neste relatório. A seleção criteriosa dos métodos e a organização rigorosa do projeto revelam-se cruciais para o seu êxito.

4.1. Pergunta de Investigação

Considerando toda a informação previamente delineada neste relatório, ressalta-se que a presente investigação emerge da necessidade de desenvolver treino linguístico especializado para falantes de inglês americano (ou de outras variantes do inglês, embora, conforme anteriormente mencionado, para uma análise mais objetiva, a comparação se restrinja ao inglês americano e ao inglês britânico) em contextos de localização para o inglês britânico. Assim sendo, torna-se imperativo desenvolver uma solução escalável que permita treinar falantes nativos de inglês americano na variante do inglês britânico. Isso implica, inevitavelmente, a automatização desse processo. Desta forma, a questão que orienta esta investigação é: Qual é a eficácia do GPT-4 no treino de falantes nativos de inglês não britânico em inglês britânico?

4.1.1. Resultados Ideais

Caso o GPT-4 demonstre eficácia nesta função, antecipa-se um cenário onde a automatização do processo seja integral. Isso equivaleria a uma base de questões altamente fiável, com um nível de correção e precisão superior a 80% e uma incidência

baixa de erros por questão. Seria ainda indispensável que tal automatização resultasse na minimização, ou mesmo na eliminação, da necessidade de intervenção humana para realizar correções. Além disso, seria crucial assegurar que os termos ou expressões selecionadas para os testes fossem relevantes e apresentassem um grau de dificuldade apropriado ao contexto empresarial em questão.

4.2. Conceção da investigação

Para assegurar a eficácia deste estudo, delineou-se uma metodologia desde o seu início:

O primeiro passo envolveu a realização de um estudo exploratório sobre as funcionalidades do GPT-4, com especial foco na sua aplicação como recurso de apoio ao ensino da língua inglesa, e mais precisamente na sua utilidade para o treino focado em variantes linguísticas. Durante esta fase, procedeu-se ao estudo comparativo entre as relevantes variantes do inglês, o que culminou na elaboração de uma tabela ilustrativa das discrepâncias entre o inglês britânico e o americano. Essa compilação serviu, inicialmente, para documentar de forma sistemática as divergências identificadas; em seguida, a tabela foi empregada nas interações com o GPT-4, fornecendo-se ao sistema *prompts* cuidadosamente elaborados com o intuito de gerar questões que explorassem a diversidade entre as formas variantes.

Após a conclusão da pesquisa preliminar, tornou-se imperativo delinear a metodologia do estudo de forma mais detalhada. Dada a abrangência e a relativa novidade do tema em questão, impunha-se a necessidade de estabelecer parâmetros claros para circunscrever a investigação, especialmente considerando o período de tempo limitado disponível para a realização do estágio. Foi nesta etapa que se procedeu à definição das categorias de perguntas a serem utilizadas nos testes.

Com a metodologia estabelecida e a contextualização necessária, deu-se início à fase prática do projeto, que consistiu na utilização de *prompts* para gerar conjuntos de perguntas. Este processo foi conduzido ao longo de vários dias, uma vez que envolveu a experimentação iterativa com diferentes *prompts*. Ao mesmo tempo, prosseguiu-se com a investigação sobre o campo de *prompt engineering*, cujas descobertas revelaram-se de grande valia para o refinamento e eficácia da criação destes.

Durante este processo, procedeu-se à anotação sistemática de todas as problemáticas identificadas bem como das soluções encontradas, as quais serão discutidas em maior detalhe em seções subsequentes. Assim, ao longo de vários dias, realizou-se uma avaliação contínua de todas as questões geradas.

Por meio de um processo iterativo de tentativa e erro, diversos *prompts* foram testados ao longo dos primeiros meses. O refinamento destes ajustava-se em função dos resultados obtidos ao longo do processo. Após conseguir os *prompts* desejados, procedeu-se à formulação do conjunto de questões. No entanto, mesmo após o aprimoramento dos *prompts*, as questões geradas ainda continham múltiplos deslizes, imprecisões, entre outros erros.

Assim sendo, tornou-se imprescindível proceder a uma avaliação mais meticulosa das questões para aferir a sua qualidade. Para tal, foi necessário recorrer a um linguista nativo de inglês britânico, a fim de rever a dimensão linguística das questões; as anteriores pesquisas, não obstante, persistiam nuances linguísticas com erros que, não sendo eu falante nativa, não conseguia identificar no projeto.

Para esta análise foi necessária a criação de uma tabela de avaliação, que incorporava todas as questões geradas pelo GPT-4, juntamente com um conjunto de itens de avaliação selecionados para permitir uma análise minuciosa da qualidade das questões. Essa tabela foi então remetida a uma linguista nativa de inglês britânico, que teve a incumbência de avaliar todas as questões de acordo com os critérios estabelecidos. Esta linguista tem um extenso historial de tradução e experiência no ramo da educação, como tal, os seus conhecimentos são valiosos na otimização das perguntas geradas.

Com a tabela devidamente preenchida, seguiu-se a análise dos resultados e o respectivo feedback. Os comentários recebidos foram fundamentais para determinar a eficácia do GPT-4 na criação de bancos de questões destinados ao treino ou à avaliação de falantes não nativos de inglês britânico. Uma exploração mais detalhada destas conclusões será apresentada adiante neste documento.

4.3 Primeiros testes com GPT-4

Todo o processo iniciou-se com um *prompt* básico, no qual se solicitava ao GPT-4 que gerasse questões sobre o tema selecionado. Importa referir que esta foi a minha iniciação no uso do GPT-4; assim, inicialmente, não me foquei no aperfeiçoamento dos *prompts*. Com isso em mente, nesses, pedia-se que o sistema criasse bancos de questões de múltipla escolha com uma resposta correta e, em outros casos, questões de preenchimento de espaços. Testando-se repetidamente a geração desses *outputs*, com o objetivo de aferir a qualidade incipiente sem qualquer refinamento, identifiquei que o principal defeito residia na repetição e na inadequação dos distratores. Ou seja, nas questões de múltipla escolha, as opções incorretas eram facilmente identificáveis. Por vezes o GPT-4 tende a criar questões excessivamente simples, em algumas das quais as respostas corretas eram flagrantemente óbvias devido à evidente discrepância de significado entre as opções, ou porque os distratores incluíam opções agramaticais. Veja-se o exemplo:

```
Question 3: Auxiliary Verbs In British English, which sentence
uses the auxiliary verb correctly?
A) I might could go to the football match later.
B) I haven't eaten any chocolates.
C) We don't can open this bottle.
D) She's ran the London Marathon.
```

O facto de algumas das alternativas de resposta (C e A) apresentarem conjugações agramaticais, facilita que o avaliado chegue à resposta correta através da eliminação dos distratores. Importa igualmente salientar que o próprio enunciado da pergunta denota uma estruturação atípica, resultando numa falta de fluidez. É importante, no entanto, salientar que a ocorrência destes erros pode ser atribuída ao *prompt* utilizado, considerando que, sendo dos primeiros empregados, estes apresentavam uma carência em termos de instruções adequadas. Da mesma forma, os parâmetros relativos à temperatura e à penalidade por frequência permaneceram inalterados, seguindo os valores padrão.

Ainda também, nos resultados produzidos, as questões incidiram sobre as diferenças entre termos comuns nos dialetos do inglês: *garage/car park*; *flat/apartment*; *gasoline/petrol*; *candy/sweets*; *wallet/handbag*; *diaper/nappy*; *flashlight/torch*; "*trash/rubbish*", entre outros.

O propósito inicial consistia em assegurar que, após cada geração de *output* gerada pelo GPT-4, o histórico correspondente fosse eliminado. Isto é, o intuito não residia em criar um contexto conversacional no chat, mas antes estabelecer um processo completamente automatizado, desprovido da necessidade de qualquer informação precedente. Dessa maneira, permitiria a qualquer usuário adotar um método de "*copy and paste*"" dos *prompts*, simplificando assim a geração de conteúdo. Devido a essa necessidade de eliminar o histórico era inviável instruir o modelo a evitar repetições de termos, pois faltaria a continuidade contextual necessária para tal função.

4.3.1 Recolha e Implementação de Diferenças Linguísticas

Tendo em consideração os desafios mencionados, procedeu-se à otimização dos *prompts*. A primeira etapa desta otimização focou-se na mitigação do problema da repetição. Como solução inicial, compilou-se toda a informação obtida durante a pesquisa linguística, isto é, as discrepâncias identificadas entre o inglês britânico e o inglês americano, incluindo variações lexicais previamente ilustradas. O propósito desta abordagem era dispor de um leque mais extenso de diferenças linguísticas e, em seguida, disponibilizar essa compilação ao GPT-4 juntamente com o *prompt*. Deste modo, ao utilizar os dados congregados, esperava-se que o modelo evidenciasse menor propensão à repetição.

Esta informação resultou da consulta a uma gama selecionada de bibliografias e recursos eletrónicos especializados. Realça-se a escassez de estudos recentes enfocando as discrepâncias entre as variantes linguísticas em questão de forma atualizada para além das fontes mencionadas anteriormente (ver secção 3.6.2). Assim, complementando a discussão empreendida nesse capítulo, recorreu-se também às obras de McArthur (1992), Peters (2004) e Scott (2004), para alguma informação adicional. No entanto, as diferenças coligidas visaram um propósito específico: avaliar o nível de conhecimento dos tradutores relativamente ao inglês britânico. As diferenças escolhidas foram, por conseguinte,

aquelas que permitiram um confronto direto com o inglês americano, este último mais predominante e, portanto, um referencial comparativo robusto para as questões elaboradas.

A obra de John Algeo (2006) destacou-se pela sua aplicabilidade prática, ao elencar divergências substantivas entre ambas as variantes, complementadas por percentagens que indicavam a frequência de utilização, permitindo discernir sobre a pertinência de cada uma para a inclusão nos testes. Salienta-se a importância de selecionar essas diferenças não só pela relevância, mas também pela objetividade e consistência que conferem aos testes no âmbito da empresa. Deste modo, diferenças a nível gramatical, ortográfico e lexical foram consideradas como as mais significativas para o presente estudo.

Identificámos ainda como relevante a distinção terminológica, particularmente considerando que a Unbabel fornece serviços de tradução especializada. A busca por bibliografia que abordasse a terminologia específica das variantes revelou-se complexa, motivando-me, desta feita, a explorar recursos *online* em *websites* focados nas diferenças entre o inglês britânico e o americano. Tal pesquisa possibilitou a compilação de terminologias em âmbitos médico, financeiro e jurídico.

Este procedimento visou não apenas a redução das taxas de repetição nas questões, como também o empenho em incorporar informações precisas e credíveis na elaboração das mesmas. A escolha de fontes digitais, procurou assegurar a integridade e a adequação do material reunido à natureza do trabalho realizado na empresa.

4.4. Processo e obstáculos encontrados

Aquando da implementação da tabela com as diferenças entre as variantes linguísticas, os resultados iniciais evidenciaram uma série de problemas. O mais notório foi novamente a ineficácia dos distratores (Carr, 2011 p. 92). A título ilustrativo, apresenta-se o seguinte *prompt*:

```
Using this as an example:
```

British	American

Biscuit	Crackers
Chips	Fries/French Fries
Flat	Apartment
Candyfloss	Cotton Candy
Football	Socker
University	College
Holiday	Vacation
Timetable	Schedule
Postbox	Mailbox

```
Create multiple choice tests to train a native american speaker
in british english.
```

Considera-se a primeira parte do *prompt*, introduzida pela expressão "*using this as an example*:", na qual a tabela foi diretamente inserida no *prompt* antes das instruções subsequentes, localizadas imediatamente abaixo desta.

Considerando a simplicidade do *prompt*, era previsível que surgissem problemas. Apesar do GPT-4 ser uma versão mais avançada do que o modelo publicamente disponível, o GPT-3.5-turbo subsiste a necessidade de realizar diversos ajustes para otimizar os resultados obtidos. A qualidade do *output* pode, então, ser avaliada como segue:

```
1. What is the British equivalent of the American word
"Crackers"?
    a) Chips
    b) Timetable
    c) Biscuit
    d) Football

2. In American English, "Cotton Candy" is similar to which
British term?
    a) Chips
    b) Candyfloss
    c) University
```

```
    d) Flat

3. The British term "Chips" is similar to which American
English word?
    a) Fries/French Fries
    b) Crackers
    c) Schedule
    d) Football
```

Analisando estes exemplos, um aspeto que se sobressai é a qualidade dos distratores. Como método de correção deste problema, incluí explicitamente no *prompt* instruções para garantir que todos os distratores apresentassem proximidade de significado. Esta estratégia mostrou-se significativamente eficaz para aprimorar a qualidade dos resultados obtidos. Ao enfatizar a semelhança semântica entre os distratores (Carr, 2011) observou-se um aumento na dificuldade dos testes, o que por sua vez, contribuiu para a elevação do padrão de exigência. A diferença alcançada é apresentada a seguir:

```
2. In British English, "Candyfloss" corresponds to which term
in American English?
   a) Cotton Candy
   b) Bubble Gum
   c) Lollipop
   d) Jelly Beans
```

No entanto, o que pareceu resultar num aspeto começou a desencadear problemas noutro. Com estas novas tentativas, emergiram padrões de problemas. Um desafio que persistiu durante o aperfeiçoamento dos *prompts* foi a repetição de termos.

Durante os testes iniciais com a tabela de diferenças linguísticas, o GPT-4 mostrava preferência pelas primeiras colunas da tabela, de tal forma que, a cada novo *output* gerado, testavam-se os mesmos termos e expressões, comprometendo a qualidade

dos testes. Mesmo após detetar este problema e modificar as instruções dos *prompts* para o GPT-4, no sentido de este usar uma maior variedade de exemplos da tabela em vez de se cingir aos primeiros termos e exemplos, a repetição persistiu.

É verdade que, após a inclusão dessas instruções nos *prompts*, se notou uma ligeira tendência em selecionar outros termos para as interrogações. No entanto, esta pequena alteração não resolveu completamente o problema. Além do mais, como anteriormente referido, após todos esses testes, era necessário eliminar qualquer conteúdo no histórico do GPT-4, de modo a evitar que houvesse qualquer contexto prévio quando da geração de novo conteúdo. Por essa mesma razão, o mesmo padrão de repetição formou-se novamente. Esta tentativa de aleatorizar a escolha de informação da tabela não foi bem-sucedida.

Com o objetivo de solucionar este problema, optei por dividir a tabela original em subconjuntos focados em áreas terminológicas e aspetos gramaticais. Consequentemente, a tabela foi segmentada em várias tabelas menores, distribuídas por diferentes folhas. Em vez de integrar a tabela completa no *prompt*, procedi à adição de uma tabela mais específica e solicitei a geração de um número de questões equivalente ao total de exemplos ou termos presentes na tabela. Assim, por exemplo, se uma tabela contivesse 20 exemplos ou termos, seria pedido que se gerassem 20 questões correspondentes. Esta abordagem eliminaria a possibilidade de repetição.

De facto, esta abordagem mostrou-se bastante eficaz, exceto no caso da tabela de gramática. A ausência de exemplos concretos, resultou em várias instâncias onde foi apresentada apenas a estrutura gramatical sem exemplos ou unicamente com exemplos ilustrativos. A intenção por detrás da apresentação da estrutura gramatical era incitar o GPT-4 a gerar os seus próprios exemplos, fomentando assim o processo automático desejado. Contudo, tal expectativa não se concretizou, resultando em outputs defeituosos provenientes dessa tabela.

Outra solução eficaz para o problema da repetição foi a preservação do histórico de conversação. Ao optar por não eliminar o histórico e iniciar um novo diálogo na interface de chat, foi possível introduzir um *prompt* como "*Now give me a set of 5 more questions and do not repeat examples*". O objetivo deste *prompt* era prevenir a repetição de exemplos previamente utilizados e incentivar a incorporação de novas diferenças linguísticas no conteúdo gerado.

A utilização da tabela de diferenças linguísticas mostrou-se parcialmente ineficiente. A escassez de fontes fiáveis e detalhadas limitou a capacidade de recolher uma ampla gama de diferenças, o que, por sua vez, restringiu a produção de um conteúdo variado. Diante dessa situação, optámos por prescindir da tabela na maior parte do processo de geração das perguntas. Esta foi empregada exclusivamente no banco de questões de escolha múltipla para elaborar questões focadas em terminologia específica. Nos demais bancos de questões, a tabela deixou de ser utilizada. A figura 2 ilustra o exemplo do processo de teste dos *prompts*:

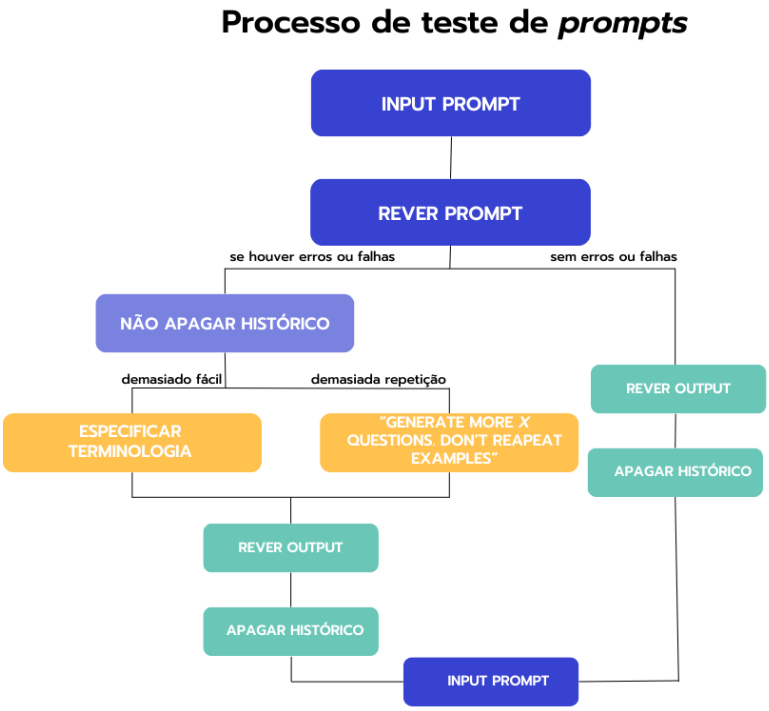


Figura 2 - Processo de teste de *prompts*

Um dos desafios identificados residiu na degradação da qualidade de desempenho. Isto sugere que, à medida que a extensão dos *prompts* aumentava, ocasionalmente, o GPT-4 deixava de cumprir determinadas instruções. Uma das manifestações mais claras deste fenómeno verificou-se nas questões de revisão. A minha intenção era usar exercícios de revisão no banco de questões, contudo, com os resultados obtidos, não foi possível. Durante o processo de desenvolvimento dos *prompts*,

deparámo-nos com múltiplos obstáculos e, em resposta a cada um deles, implementámos instruções adicionais com o propósito de prevenir reincidências dessas falhas. Ao tomar consciência desta situação, verifiquei que os *prompts* se tinham tornado excessivamente extensos, no entanto tais acréscimos não se traduziram na melhoria de desempenho. Para exemplo ilustrativo, veja-se o seguinte *input* e *output*:

```
Take this as an example:
```

British	American
cannot	cannot
musn't	shouldn't
needn't/need not	doesn't have to/didn't have to
ough not to/oughtn't to	shouldn't
used not to/usen't to	didn't use to
will/shall,be going to	be going to
have been going on	was going on
hadn't had	didn't have
has certainly	certainly has
done beautifully	beautifully done
made oddly	oddly made
proper(ly)	really/very
a print-out presentation	a printout presentation
Skipping Rope	Jump rope
The artists enjoyed the sing-along on the stage.	The singers had fun with the singalong on the campfire.
dialing tone	dial tone
racing car	racecar
rowing boat	rowboat
barber's shop	barbershop
cookery book	cookbook

skimmed milk	skim milk
The committee were unable to agree.	The committee was unable to agree.
Muse are a great band.	Muse is a great band.
The Beatles are playing at Wembley.	The Beatles are playing at Wembley.
One Direction are a great band.	One Direction is a great band.
At the weekend	On the weekend
Play in a team	Play on a team
In hospital	In the hospital
Monday to Friday	Monday through Friday
Fill in a form	Fill out a form
Write to someone	Write someone
At the back	In the back
Biscuit	Cookie
Chips	Fries/French Fries
Flat	Apartment
Candyfloss	Cotton Candy
Football	Socker
University	College
Holiday	Vacation
Timetable	Schedule
Postbox	Mailbox
wash up (wash dishes)	wash up (wash someone's body)
pavement (side of the road)	pavement (the road itself where vehicles drive - sidewalk is the side of the road)
faculty (university school or college)	faculty (teaching staff)
purse (what americans refer as a wallet)	purse (a woman's handbag)
Managing Director	CEO (Chief Executive Officer)

Estate Agent	Realtor
Cheque	Check
Chemist / Chemist’s	Pharmacist, Drugstore / Pharmacy
Clinical Trial	Clinical Study
General Practitioner	Family Practitioner / Physician
P.I.D (protruded intervertebral disc)	P.I.D (pelvic inflammatory disease,’ a decidedly female ailment)
aetiology	etiology
anaemia	anemia
anaesthetic	anesthetic
dyslipidaemia	dyslipidemia
glycaemic index	glycemic index
gynaecology	gynecology
haemoglobin	hemoglobin
haemorrhage	hemorrhage
paediatric	pediatric
coeliac	celiac
dyspnoea	dyspnea
manoeuvre	maneuver
oedema	edema
oesophagus	esophagus
oestrogen	estrogen
menorrhoea	menorrhea
homoeopath	homeopath
debtors	receivables
creditors	payables
fixed assets	property/plant and equipment
financial year	fiscal year
turnover	sales
stocks	inventory

equity capital	common stock
Operating profit/loss	Income/loss from operations
Profit and loss account	Income statement
dialogue	dialog
speciality	specialty
colour	color
Catalogue	Catalog
travelling	traveling
jewellery	jewelry
honour	honor
labour,labor	labor
programme	program
Pardon?	Excuse me?
not one bit	not at all
excuse me	sorry
DD/MM/YYYY	MM/DD/YYYY
ordinal month (9th november)	november 9/9 november
9.30 am (AM, PM/a.m., p.m.)	9:30 am (AM, PM/a.m., p.m.)
pounds (£)	dollars ($)
123 – one hundred and twenty-three	123 - one hundred twenty-three
2,200 – two thousand two hundred	2,200 - twenty-two hundred
124 - one two four	124 one twenty-four

I'm planning to conduct training for American English speakers on British English writing. Please assemble a reviewing exercise with the following structure: only one text written both in American English and British English using the examples from the table; the American English and British English parts must be highlighted like *word or expression*; then create a set of multiple choice questions about the highlighted parts of the text, the questions must consist of asking if the highlighted parts are correct and then the person must choose the right answer. Make sure the distractors

```
are similar in meaning to the correct answer. The goal is to
evaluate American to British writing skills, thus the options
should involve choosing the British English equivalent when
necessary. Avoid mentioning the theme being assessed within
the question. Ensure to clearly identify the right answer for
each query.


Text:

American English:

Last *weekend*, I *filled out a form* at the *drugstore* to
join a *clinical study*. The *pharmacist* said I *certainly
has* a chance to be selected. The study is about the effects
of the new *anesthetic* on patients with *dyslipidemia*. I
also bought some *skim milk* and *crackers*. On my way home, I
saw a beautiful *racecar* that was oddly made. When I got
home, I found out that my favorite band, One Direction, *is*
performing at Wembley.

British English:

Last *at the weekend*, I *filled in a form* at the *chemist's*
to join a *clinical trial*. The chemist said I *has certainly*
a chance to be selected. The trial is about the effects of the
new anaesthetic on patients with dyslipidaemia. I also bought
some skimmed milk and biscuits. On my way home, I saw a
beautiful racing car that was made oddly. When I got home, I
found out that my favourite band, One Direction, are
performing at Wembley.

Questions:

1) In the American English text, is "weekend" used correctly?
A) Yes
B) No - Correct answer (it should be 'on the weekend')

(...)
```

É observável que, no início do *prompt*, fornecemos a tabela de diferenças linguísticas, seguida pela instrução "*Take this as an example*", e subsequentemente acompanhada por uma lista de diretrizes. É relevante salientar que, neste contexto

específico, não disponibilizei um exemplo concreto da tarefa a ser executada pelo GPT-4, o que pode ter sido um fator contribuinte para o resultado defeituoso. No entanto, este exemplo particular realça tanto o excesso de comprimento como o não cumprimento das instruções fornecidas. Solicitei explicitamente "*only one text written both in American English and British English*", e o GPT-4 produziu dois textos distintos, um em inglês britânico e outro em inglês americano.

Neste ponto, tornou-se evidente que o GPT-4 enfrentava alguns problemas consistentes: mostrava limitações quanto à aderência completa a *prompts* que continham múltiplas instruções e, a este respeito, observou-se que quanto mais extenso o *prompt*, menor a qualidade do output gerado. Assim, ficou claro que era necessário encontrar um equilíbrio na formulação dos *prompts*, de modo a evitar extensões exageradas, mas, simultaneamente, assegurar que continham informações e comandos relevantes que preservassem a qualidade das questões propostas. Essa constatação levou-nos a desencorajar o uso repetido da tabela e, por consequência, a abandonar a ideia de elaborar exercícios de revisão, uma vez que, em todos os testes realizados, o modelo nunca foi capaz de gerar um resultado satisfatório.

4.5. Bancos de Questões e *Prompts* Utilizados

Até este ponto da investigação, embora estivesse a experimentar diversos tipos de questões, ainda não havia tomado uma decisão sobre quais seriam incluídas nos bancos finais. Da mesma forma que estava a avaliar os *prompts* e a determinar os comandos mais adequados para este tipo de tarefa, encontrava-me igualmente a testar quais os tipos de questões mais apropriados para este projeto. Naturalmente, os critérios de decisão relativamente à sua inclusão nos bancos finais dependiam inteiramente do desempenho do GPT-4 e da medida em que era necessário intervir no *prompt*.

Tendo em conta todos os problemas identificados, tornou-se evidente que era crucial manter um certo equilíbrio nos *prompts*. Ou seja, mostrou-se imperativo fornecer exemplos das tarefas realizadas, pois, como observado anteriormente, a performance melhorava significativamente quando bons exemplos das tarefas eram fornecidos. Adicionalmente, seria necessário selecionar exercícios mais simples e coesos para otimizar o desempenho do GPT-4. Dessa forma, seria possível utilizar *prompts* mais

objetivos e concisos, que não exigissem extensas instruções para alcançar uma performance satisfatória.

Assim sendo, nesta secção, será abordado cada *prompt* utilizado, os comandos específicos de cada um deles e os bancos de questões empregados nesta investigação.

4.5.1. *Prompt* Escolha Múltipla

Decidimos desde início manter os tipos de questões padrão ou mais frequentes em qualquer tipo de teste: escolha múltipla e verdadeiro ou falso, ambos bastante semelhantes em sua estrutura. A razão para esta escolha deve-se ao facto de serem os tipos de questões mais comuns encontrados em diversos testes, mas principalmente devido à sua prevalência em testes de língua[4]. Similarmente, com o intuito de fornecer um exemplo adequado no *prompt*, procedi a uma avaliação de outputs anteriores e selecionei a questão que melhor demonstrava a estrutura e desempenho desejados, incorporando-a ao *prompt*. Apresenta-se, assim, o *prompt* final:

```
    Generate a set of 20 multiple choice exercises. The
goal is to help Americans learn  British English. The
questions should ask for the British equivalent of  American
English expressions and grammar. Make sure the distractors
are always similar in meaning to the correct answer. Always
point out the right answer. Follow the next structure:
    Which of the following British English expressions is
most frequently used as the equivalent to the American
English expression "He shouldn't press the red button."?
a) He's not going to press the red button.
b) He needn't to press the red button.
c) He is not allowed to press the red button.
d) He mustn't press the red button. (correct)
```

[4] Testes como IELTS, TOEFL, TOEIC, CELPIP. Como guia inicial consultou-se o manual *IELTS 16 Academic Student's Book with Answers with Audio with Resource Bank.* (2021). Cambridge English.

Assim, pode-se observar que as únicas instruções que mantive foram: o número de questões necessárias (20); o propósito da tarefa ("t*o help Americans learn British English*"); a descrição das questões e especificidades desejadas, como, por exemplo, "*always point out the correct answer*"; e a estrutura pretendida.

4.5.2. *Prompt* Verdadeiro ou Falso

Relativamente ao banco de questões de verdadeiro ou falso, a estrutura do *prompt* foi semelhante à utilizada para as questões de escolha múltipla. Contudo, não considerei a necessidade de fornecer um exemplo, dado que, devido à simplicidade da estrutura destes exercícios, a performance do GPT-4 mostrou-se satisfatória desde o início, embora ocasionalmente variasse na produção de explicações para as respostas.

```
True or False: A British person would understand "E.R" as an
abbreviation for Emergency Room.
FALSE (In British English they call it "A&E" for Accidents
and Emergencies.)
```

Deste modo, o *prompt* usado foi:

```
Generate a set of 20 true or false questions. The goal is to
help Americans learn  British English. The questions should
ask for the British equivalent of  American English
expressions and grammar. Make the questions about medical,
tech, financial and law terminology. Always point out the
right answer. Randomize the answers.
```

Este *prompt*, semelhante ao de escolha múltipla, contém o número de questões, o propósito da tarefa e as instruções. Contudo, aqui, nota-se que especifiquei o tipo de terminologia, numa tentativa de evitar a repetição de termos como "*apartment/lift*", "*vacation/holiday*", "*biscuit/cookie*", "*gasoline/petrol*", algo que se tornava bastante frequente sem uma indicação explícita do tipo de terminologia alvo desejada para as

perguntas. Adicionalmente, incluí também o comando "*Randomize the answers*", uma vez que, em alguns casos, as respostas eram maioritariamente falsas.

4.5.3. *Prompt* Preenchimento de Espaços

Optámos pelo tipo de questão de preenchimento de espaços, dado que este é também um dos formatos mais frequentes em testes de línguas e são exercícios com os quais praticamente todas as pessoas estão familiarizadas, pois fazem parte dos métodos educacionais adotados nas escolas. Devido à sua prevalência, o desempenho do GPT-4 na geração destas questões também se mostrou satisfatório. Verifique-se o *prompt* utilizado:

```
     Generate a set of 20 fill-in-the-blank exercises. The
goal is to help Americans learn  British English. The
questions should ask for the British equivalent of  American
English vocabulary. Each question should only have one
solution and this should be evident from reading it. Follow
the structure of these examples:
     They rode the ___ to get from one floor to another,
known as a "elevator" in the US.
British English equivalent: lift

     She put on her ___ to protect against the rain, but in
the US, they'd call it a "raincoat".
British English equivalent: mac

The baby wore a ___, known in the US as a "diaper".
British English equivalent:  nappy
```

É relevante sublinhar que a necessidade de prover um exemplo concreto da tarefa mostrou-se fundamental, visto que, em testes anteriores, quando eu apenas fornecia explicações sobre como desejava que as questões fossem elaboradas, verificaram-se falhas ocasionalmente. Apresenta-se um exemplo representativo dessa situação:

```
1. In American English, the end of the week is often referred to
as _____. (Hint: British English Prepositions)
   - Answer: On the weekend
```

Neste caso, foi disponibilizada a tabela de diferenças entre variantes mencionada anteriormente. No *prompt*, especificámos que desejava que fosse incluído o espaço em branco seguido de uma pista da resposta correta. Assim, observou-se que o GPT-4 utilizou o termo "*prepositions*" como dica. Tal pista revela-se ineficaz, pois sugere que a resposta correta poderia ser qualquer proposição da língua inglesa, o que pode levar a interpretações ambíguas ou erradas. Para além disso, é de salientar que a formulação da frase em si apresentou-se atípica e pouco clara.

4.5.4. *Prompt* Transformação

Os exercícios de transformação, apesar de possuírem uma estrutura similar à dos de preenchimento de espaços, representaram uma tentativa de elaborar exercícios de preenchimento, porém numa versão mais extensa e, possivelmente, mais complexa, com um maior número de espaços a preencher, pois envolve frases inteiras. No *website Teaching British* os exercícios de transformação são descritos da seguinte forma: "*A transformation exercise is an exercise where learners are given one sentence and need to complete a second sentence so that it means the same.*" Contudo, nem sempre foi possível cumprir este objetivo, visto que, em algumas questões, não se observou um aumento no número de espaços por preencher.

No entanto, o *prompt* aderiu aos mesmos princípios observados nos anteriores:

```
Create a set of  30 sentence transformation exercises.
Provide a sentence in American English and ask the test-
```

```
taker to rewrite it using the appropriate British English
equivalents. The goal is to help Americans learn British
English. Example:
     Transform the following American English sentence using
British English words.
""I bought a candy bar at the store.""
British English equivalent: ""I bought a ______ at the
_______."" _ Make sure to always provide the correct answer."
```

Neste *prompt*, especificamos o número de exercícios desejados, neste caso 30, elucidamos a natureza do exercício ou a sua estrutura, isto é, a apresentação de uma frase seguida do pedido dirigido ao indivíduo que realiza o teste para a reescrever, posteriormente, definimos o objetivo do *prompt* e, finalmente, providenciamos um exemplo ilustrativo do formato ideal do exercício.

4.5.5. *Prompt* Identificação da Variante

No banco de questões dedicado à identificação da variante linguística, empreendeu-se uma tentativa de adotar uma nova abordagem na sua criação. Assim, este conjunto de questões distinguiu-se metodologicamente dos demais ao procurar incorporar "exemplos reais" nas questões propostas. Deste modo, estes exercícios incluíram frases extraídas de um documento que continha traduções autênticas, revistas por profissionais na Unbabel. Tendo em conta que as traduções evidenciavam traços tanto do inglês britânico quanto do americano, optamos por selecionar segmentos que demonstrassem sinais claros de ambas as variantes linguísticas.

Numa tentativa de adotar um método mais inovador, abstive-me de recorrer inteiramente às capacidades do GPT-4 para a geração automática de todas as perguntas. Em alternativa, as frases selecionadas foram incluídas no *prompt* e, a partir destas, formaram-se perguntas de identificação da variante linguística. Esta decisão metodológica foi sustentada pelo quadro andragógico proposto por Knowles (1998) (consultar secção 3.5). Nesse quadro, Knowles (1998) defende que a aprendizagem de

adultos é significativamente potencializada quando o conhecimento recém-adquirido se associa a um propósito concreto na vida real. De acordo com Knowles, a motivação para aprender é consideravelmente reforçada quando os materiais didáticos se encontram alinhados com as perspetivas profissionais e pessoais dos aprendizes (consultar secção 3.5).

Assim sendo, a elaboração deste banco de questões constituiu uma tentativa de avaliar o desempenho do GPT-4 na criação de perguntas mais contextualmente relevantes, visando potencialmente diminuir a lacuna entre o conhecimento teórico e a aplicação prática por parte dos tradutores, além de proporcionar questões mais imersivas e contextualmente fundamentadas. É igualmente relevante salientar que quaisquer dados pessoais ou informações específicas contidas nos segmentos utilizados foram alteradas ou omitidas com o intuito de preservar ao máximo a confidencialidade de informações sensíveis.

Veja-se o *prompt* do banco em questão:

> Create a set of questions where you ask what variety of english the phrase is. The goal is to help Americans learn British English. The goal is to test if they can distinguish the varieties of English. Use this phrases:
> Characterized by uncoordinated muscle contractions, lasting around 2 to 4 minutes, with loss of consciousness and often accompanied by sphincter incontinence. -American
> According to the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-V), a manifest condition is defined as a distinct period of abnormally and persistently elevated, expansive or irritable mood and abnormally and persistently increased goal-directed activity or energy, lasting at least a week and present most of the day, almost every day (or any duration, if hospitalization is necessary). -This will make it possible to customize the treatment afterwards to meet the individual needs of each patient. - American

Short-sleeved viscose dress in a beautiful plain colour design. - British
You found something that you had been planning to read for a while and spent most of the afternoon looking at that. - British
Recruitment must be conducted in a fair and non-discriminatory manner. - British
We'll measure the answer to this question using a Visual Analogue Scale (VAS) ranging from 0 (Not at all critical) to 10 (Very critical). - British
Invited National Researcher and Coordinator of the Center for Eating Disorders at the North University Hospital Centre: Dr. John [EMAIL-0] - American
The trousers are designed with belt loops, side pockets and a comfortable fit with a straight leg and turn-up hem. - British
Comfortable sweatshirt with a beautiful gold text print. - British
Randomized Controlled Trial of Transcranial Magnetic Stimulation in Adults with Anorexia Nervosa – Neurobiology Through Perceived Criticism – THE STUDY - American
The consent statement for fiscal year 2001 has been filed at 21-03-2001. - British
Fill out the form below to participate - American
Zizag will report the prizes to SKAT and pay any prize taxes. How to participate: Fill in the registration form and rank the prizes you would like to see in the Christmas calendar this year. You will participate in the draw for 6 gift cards valued at kr. 500. Participation is free and the competition does not require a purchase By participating, you also agree to sign up for Club Zigzag (see membership conditions), and will receive a welcome email (unless you are already a member of Club Zigzag) Drawing of winner The competition ends on 30/11/03 The winner will be found by random draw on 01.12.2003. The winner will be notified directly at the email

```
address provided when registering for the competition By
participating in the competition, it is accepted that Zigzag
may publish the winners' names on the media used by Zigzag
Prize 6 gift cards valued at kr. 500 for URL-0. - British
In other situations, if you do not pay the entire balance,
TAEG of 14.5% shall be applied. - British
Always provide the right answer
```

Nesta situação, o conjunto de instruções fornecidas no *prompt* foi consideravelmente reduzido, uma vez que a maior parte do conteúdo necessário para a elaboração das questões já estava providenciado. As instruções limitaram-se apenas a clarificar o objetivo do *prompt* e a especificar que se desejava a utilização das frases fornecidas para a conceção de exercícios de identificação de variante.

4.6. Criação de um quadro de avaliação*: Scorecard*

Quando os bancos de questões estavam completamente elaborados, tornou-se essencial a criação de uma tabela *scorecard* destinada à avaliação da qualidade dos outputs gerados pelo GPT-4. Este procedimento revelou-se crucial para assegurar a qualidade e exatidão das perguntas formuladas, sobretudo considerando a sua aplicação em avaliações das variantes linguísticas. Para levar a cabo essa avaliação, era imperativo desenvolver um quadro de critérios que abrangesse todos os possíveis erros inerentes ao conjunto de questões. Este *framework* necessitava de ser extensivo, com o objetivo de garantir que quaisquer falhas identificáveis nas questões fossem devidamente identificadas e corrigidas, assegurando, assim, a relevância e eficácia dos testes linguísticos.

A complexidade desta tarefa era acentuada pelo facto de que, apesar da minha investigação inicial sobre algumas das distinções linguísticas, a minha falta de especialização aprofundada em inglês britânico e, adicionalmente, o fato de não ser

falante nativa, impunham limitações à minha capacidade de avaliar as questões com a necessária precisão. Muitas das fontes online disponíveis para consulta encontravam-se desatualizadas ou, mesmo quando eram contemporâneas, como é o caso de alguns *websites*, apresentavam o risco de conter informações incorretas. Dado este contexto, revelava-se evidente a indispensabilidade de envolver um linguista nativo de inglês britânico no processo. A perspetiva e a expertise de um profissional com profundo conhecimento sobre as nuances linguísticas das variantes do inglês seriam cruciais para assegurar que os bancos de questões alcançassem o padrão de qualidade necessário para a elaboração de testes eficazes e precisos.

Portanto, era imperativo que esta tabela *scorecard* incluísse questões capazes de explorar todas as nuances de falhas ou erros que as perguntas pudessem apresentar. Consequentemente, tornava-se necessário avaliar cada questão individualmente, assim como os bancos de questões de forma geral. Efetivamente, um passo inicial deste processo consistia em identificar potenciais problemas nas questões e determinar o que tornaria cada uma delas insatisfatória ou inadequada para utilização em um teste.

Considerando que, de facto, tínhamos selecionado os melhores *prompts* na altura para a tarefa em questão, tornava-se claro que o *output* gerado não era perfeito e que, apesar disso, persistiam falhas em algumas questões. Por essa razão, era essencial realizar correções para os erros mais significativos. Isto implicava, portanto, a criação de perguntas *Gold Standard*. Assim, tornava-se indispensável que, para cada pergunta que necessitasse, houvesse algum elemento na resposta das perguntas de avaliação desencadeasse um alerta para quem estivesse a avaliar, indicando que era necessário elaborar um *Gold Standard*.

Uma solução implementada consistiu na criação de um menu suspenso contendo uma série de declarações que mencionavam problemas específicos. Conforme a escolha efetuada pela pessoa responsável pela avaliação das questões, o *Gold Standard* seria, ou não, ativado conforme necessário. A referida configuração pode ser visualizada na figura abaixo:

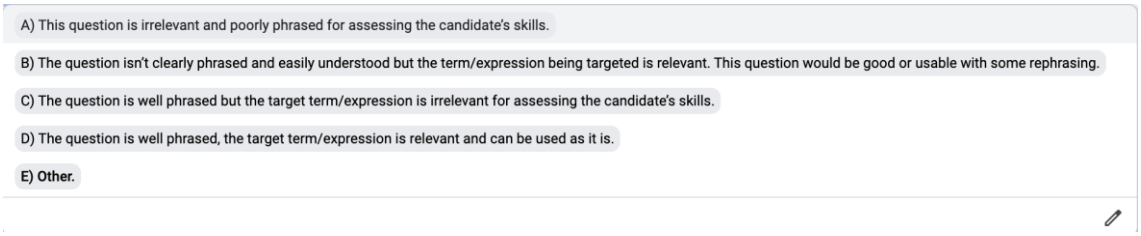


Figura 3- Menu *dropdown*

Este menu tinha como finalidade avaliar ou verificar diversos aspetos:

A. Esta opção indica que a pergunta é considerada irrelevante, ou seja, o termo ou expressão-alvo a ser testado não se mostra pertinente no contexto do teste. Adicionalmente, identifica-se que a estrutura ou a forma escrita da pergunta não está clara e contém erros. Caso o avaliador selecione esta opção, a necessidade de criação de uma questão *Gold Standard* será ativada.
B. Esta indicação refere-se a situações onde a estrutura e a forma escrita da questão apresentam defeitos e imprecisões, embora o termo ou expressão-alvo seja considerado pertinente para avaliar os conhecimentos dos indivíduos que realizam o teste. A seleção desta opção pelo avaliador resultará na ativação do processo de criação de uma questão *Gold Standard*.
C. Diferentemente da opção B, esta alternativa aponta que a estrutura e redação da questão são consideradas satisfatórias, não apresentando defeitos. No entanto, o termo ou expressão-alvo selecionado não se mostra pertinente no contexto proporcionado; por exemplo, pode representar uma diferença linguística inexistente. A escolha desta opção pelo avaliador desencadeará a ativação do processo para a criação de uma questão *Gold Standard*.
D. Se o avaliador optar por esta alternativa, significa que a questão não apresenta irregularidades e não necessita de qualquer tipo de alteração. A

seleção desta opção não implicará a ativação do processo para a criação de uma questão *Gold Standard*.

E. A opção "*Other*" serve para os casos em que o avaliador identifique erros que não estiveram contemplados nas opções anteriores. A escolha desta alternativa não resultará na ativação do processo de criação de uma questão *Gold Standard*.

Se a criação de uma questão *Gold Standard* fosse ativada, isso indicaria que as questões apresentavam erros graves. Nessa situação, o linguista não necessitaria responder às restantes questões de avaliação, pois a ação prioritária seria a criação da questão *Gold Standard*. Caso contrário, ou seja, se a questão *Gold Standard* não fosse ativada, ele deveria prosseguir respondendo às questões remanescentes, que abordariam outros problemas identificados. Esse conjunto subsequente de questões consistiria em perguntas de natureza mais objetiva, mais especificamente perguntas de "sim" ou "não", além de questões qualitativas. Estas últimas convidavam o avaliador a fornecer contexto para as suas respostas, permitindo uma compreensão mais detalhada das questões avaliadas. As restantes perguntas de avaliação sã

- [Q4.1] Is the suggested right answer actually the most frequent?
- [Q6.1] Is there more than one frequently used answer?
- [Q4] Is the suggested right answer correct?
- [Q5] If your response to Q4 was "no", please specify what the correct answer should be.
- [Q6] Is there more than one correct answer?
- [Q7] If your response to Q6 was "yes", please specify what are the other correct answers.
- [Q8] Are the distractors close in meaning to the correct answer?
- [Q9] Was the phrase/term relevant and up-to-date with modern usage in American English?
- [Q10] If your response to Q9 was "no", please state a more current term or expression equivalent used in American English.
- [Q11] Was the phrase/term relevant and up-to-date with modern usage in British English?

- [Q12] If your response to Q11 was "no", please state a more current term or expression equivalent used in British English.
- [Q13] Is the correct answer mentioned or indicated in the stem itself?
- [Q14] Kindly share any thoughts, feedback or suggestions you have regarding this question and its answer options. Any commentary or insights you provide are highly valuable and appreciated.
- [Q15] Was the phrase provided an unequivocally clear example of British or American English without ambiguity?
- [Q16] If your response to Q15 was "no", what are the main confusing elements you can identify?

Também se solicitou uma avaliação dos bancos de questões em sua totalidade, visto que era essencial determinar a eficácia desses para futura utilização. Dessa forma, para proceder à avaliação dos bancos de questões em sua inteireza, foram pedidos os seguintes elementos:

- [Q17] On a scale of 1 (Not useful at all) to 5 (Very Useful), please rate how appropriate you believe this question bank is for evaluating the knowledge of non-native British English speakers about British English.
- [Q18] Is the question bank comprehensive enough to cover key areas necessary for training non-native translators on British English?
- [Q19] In your opinion, was there a crucial difference between UK and US English not covered in the question bank?
- [Q20] If yes, what?
- [Q21] Kindly offer a comprehensive evaluation of the question bank specifically addressing any potential shortcomings or areas of improvement that have not been previously discussed. Your professional insights are fundamental to the continuous refinement and progression of this examination process.

Estas questões foram concebidas para oferecer perspetivas abrangentes sobre todas as nuances linguísticas possíveis do inglês britânico, e, além disso, permitir a

apresentação de sugestões sobre possíveis lacunas ou insuficiências que poderiam ser benéficas para o treino e avaliação de candidatos ou tradutores. Técnicas de formulação de questões ou métodos de classificação como estes facilitam a inclusão de feedback qualitativo mais aprofundado e esclarecedor, que abrem caminho para a exploração de contextos até então não considerados.

Após a finalização deste quadro de avaliação, o mesmo foi enviado a uma linguista nativa de inglês britânico. Devido a restrições de tempo, foi possível contar apenas com a avaliação de uma única pessoa para analisar os bancos de questões.

4.7. Proposta de categorias de erros

A justificação para estas categorias emergiu da tentativa de recolher os dados da investigação de forma mais coesa. A organização dos dados é crucial para a sua ilustração visual através de gráficos e formas mais quantitativas para resultados mais conclusivos. Nesta investigação, é fundamental identificar os erros mais frequentes tanto nos resultados como nos bancos de questões para promover a correção dos *prompts*. Segue-se uma justificação para as categorias adotadas neste estudo.

Dada a especificidade do tema, revelou-se um desafio encontrar referências ou padrões para avaliar as questões ou testes, especialmente considerando a grande diversidade dos tipos de questões abordadas e a sua origem (considerando que são output do GPT-4). A maior parte da bibliografia existente centra-se em questões de escolha múltipla e não abrange o contexto único destas.

É ainda importante salientar que estas categorias são aplicáveis à totalidade das questões, o que significa que a sua aplicação não se restringe apenas à avaliação das respostas, do enunciado ou dos distratores, mas estende-se ao conjunto integral da questão.

Tendo em conta todos estes aspetos, apresentam-se as categorias propostas para a análise dos resultados deste relatório. Estas serão detalhadas nas secções de resultados, proporcionando uma visão mais visual e abrangente deste estudo.

Accuracy ou precisão: Refere-se à correção das informações apresentadas na questão; isto é, verifica se a afirmação feita é factualmente correta. Por exemplo, se uma resposta indica que certos termos são sinónimos mas verifica-se que não significam o

mesmo, ou não existe uma distinção clara entre eles porque a expressão americana também é utilizada em inglês britânico, isso é classificado como um erro de precisão. Avalia-se, portanto, a veracidade do enunciado. O mesmo princípio pode ser aplicado às respostas incorretas, isto é, sempre que a resposta proposta pelo GPT-4 se revele errada. Tome-se como exemplo:

```
Which of the following British English expressions is most
frequently used as the equivalent to the American English
expression "She's looking for a cookbook."?
   a) She's looking for a cookery book. (correct)
   b) She's looking for a cooking book.
   c) She's looking for a chef's book.
   d) She's looking for a kitchen book.
```

Para mais contexto, considere-se a avaliação da linguista: "*(...) the expression cookbook has been adopted in BE. We do, however, often use the expression recipe book.".*

Lack of Context ou falta de contexto: Indica que o enunciado não fornece informação ou contexto suficiente para resolver a questão adequadamente. Esta lacuna pode levar o examinando a errar não por falta de conhecimentos linguísticos, mas devido à ausência de um cenário completo da questão. Exemplos disso incluem situações em que, por exemplo, a avaliadora observou a existência de múltiplas respostas possíveis para uma determinada questão, variando conforme o contexto, seja ele formal ou informal. Uma situação semelhante ocorreu em relação a diferentes usos de certos vocábulos, indicando que há várias respostas potenciais dependendo dos padrões linguísticos específicos, que geralmente são explicitamente solicitados aos tradutores. Como exemplo ilustrativo:

```
How do the British refer to a medical "Seizure"?
   a) Convulsion
   b) Fit (correct)
```

```
c) Swoon
d) Shock
```

Este exemplo específico apresenta o termo "*seizure*", que é de terminologia médica, contudo, sem qualquer tipo de contexto, existem várias respostas possíveis, segundo a avaliadora: "*(...) convulsion, fit and seizure are all used, depending on the context. I would consider not using this question unless a very specific medical context is provided."*

Ambiguity ou ambiguidade: Similar à falta de contexto, a ambiguidade diz respeito à possibilidade de múltiplas interpretações de uma frase. Pode causar confusão ou levar ao erro por exigir que o examinando adivinhe ou interprete de forma ambígua o significado. Este tipo de erro complica a extração de conclusões claras. Da mesma forma, uma questão também pode ser classificada com esta categoria nas instâncias, por exemplo, em que dois termos podem ser aplicáveis no inglês britânico, sendo a sua escolha condicionada por diretrizes específicas que normalmente são pedidas aos tradutores.

Difere da falta de contexto, pois avalia a clareza e precisão da questão em si, enquanto a outra avalia as lacunas de informações necessárias para a realização da questão. Segue-se como exemplo ilustrativo:

```
What variety of English is this phrase? "In other situations,
if you do not pay the entire balance, TAEG of 14.5% shall be
applied."
British
```

Relevance ou relevância: Avalia o quão pertinente é a informação; isto é, mede o quanto um termo ou expressão é útil para avaliar conhecimentos de inglês britânico. Por exemplo, questionar sobre a localização em inglês britânico de "*computer*" quando ambos os dialetos utilizam o mesmo termo não seria relevante. Embora não seja uma ocorrência comum em todas as instâncias, um dos exemplos mais evidentes que ocasionalmente se verifica nesta investigação é quando o vocábulo ou expressão utilizada revela-se de tal forma irrelevante que a avaliadora opta por não estabelecer um *Gold Standard*,

considerando-os impertinentes devido à ausência de diferenças significativas entre as variantes linguísticas. A exemplo desta situação, verifique-se:

> Don't forget to recite the ___, or as they'd call it in the US "Pledge of Allegiance". British English equivalent: Oath of Loyalty.

Distractor quality ou qualidade dos distratores: Esta categoria examina a eficácia dos distratores em perguntas de escolha múltipla, assegurando que não forneçam pistas inadvertidas acerca da resposta correta através de inconsistências de significado ou outros indicadores. Verifique-se o exemplo:

> What's the equivalent of biscuit?
>
> a) Mouse
> b) Cookie
> c) Green
> d) Sky

É também relevante mencionar que a construção destas categorias teve por base as questões apresentadas à avaliadora. Dado que se procuram identificar os mesmos erros, torna-se lógico converter algumas dessas questões em categorias de avaliação.

4. RESULTADOS

Após a conclusão deste processo, foi possível proceder à análise dos resultados obtidos na avaliação. Para esse efeito, realizou-se uma contabilização sistemática e quantitativa dos erros, com o objetivo de tornar os resultados o mais conclusivos possível. Neste capítulo, serão apresentados os resultados decorrentes do estudo em questão. Conforme mencionado previamente, o estudo contou com a participação única de uma linguista, falante nativa de inglês britânico. A apresentação dos resultados estará organizada em cinco secções distintas: Escolha Múltipla (secção 5.1), Preenchimento de Espaços (secção 5.2), Transformação (secção 5.3), Identificação da Variante Linguística (secção 5.4) , e Verdadeiro ou Falso (secção 5.5).

5.1. Escolha Múltipla

O banco de questões de escolha múltipla compreendeu um total de 39 questões. Este demonstrou ser o mais propenso a erros, apresentando a maior taxa de erros por questão entre todos os bancos analisados. Notavelmente, foi o único banco de questões que incluiu uma pergunta com três erros, predominando, na maior parte das questões, a ocorrência de um único erro.

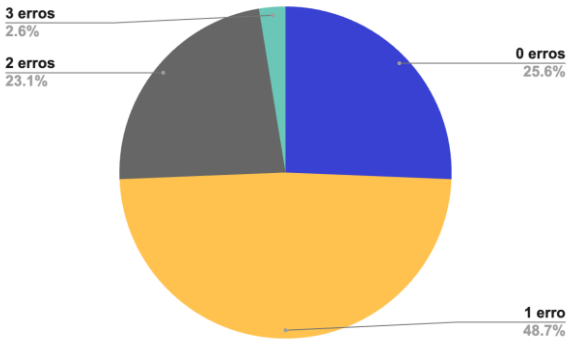


Figura 4 - Frequência de erros por questão

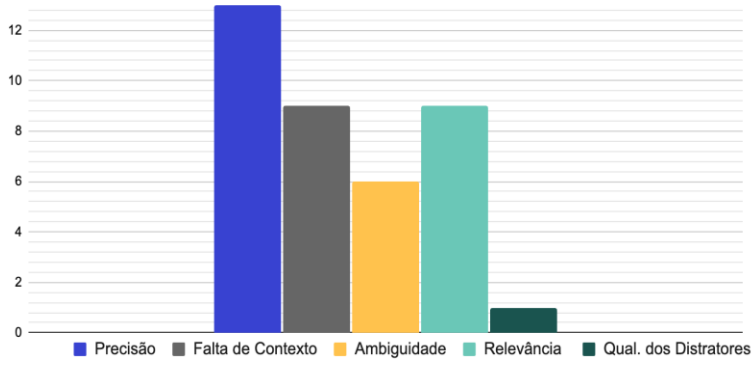


Figura 5 - Contagem dos tipos de erros

Como visto na figura 5, o erro mais frequentemente identificado está diretamente associado à precisão, representando 34% de 38 questões, o que sinaliza falhas

fundamentais intrínsecas às próprias questões. Adicionalmente, registou-se um problema significativo relacionado com a ambiguidade, correspondendo a 16%, e uma notória falta de contexto, abrangendo 24% dos casos. Esta última questão surge porque muitos dos termos foram apresentados isoladamente, sem estarem integrados em frases que facilitassem a compreensão das expressões alvo. A irrelevância dos conteúdos também foi um aspecto notório, afetando gravemente a qualidade do banco de questões, com uma incidência de 24%.

A avaliadora destacou que as diferenças observadas entre as variantes linguísticas são frequentemente dependentes de nuances específicas; isto é, a participante indicou que tais disparidades são determinadas por padrões linguísticos distintos, os quais são frequentemente estabelecidos pela empresa ou cliente. Estes padrões são esperados que o tradutor siga e mantenha rigorosamente. Consideremos o exemplo a seguir:

```
19. What is "Inventory" known as in Britain?
   Collection
   Assortment
   Stocks (Correct)
   Holdings
```

Esta questão solicita a tradução do termo "*Inventory*" para o inglês britânico, com o objetivo de avaliar o conhecimento específico em terminologia de contabilidade. No entanto, dada a ausência de contextualização do termo em uma frase, a pessoa que realiza o teste pode optar pela opção A como resposta correta, dado que o sinónimo mais conhecido é, de facto, "*collection*". Técnica e conceitualmente, essa escolha pode não ser considerada incorreta. Confirme-se o comentário da avaliadora à pergunta em questão:

> The issue with this question is a lack of context. The most common use of inventory in British English is the same meaning as in American English: a written list of all objects in a specific place. That is when it is used as a countable noun. The other meaning is when it is used as a variable noun and in US English is used to mean a supply or stock of something, at which point Answer C, if made

singular, would be correct. However, without context this is an ambiguous and misleading question.

Apesar dessas limitações, a necessidade de um *Gold Standard* revelou-se moderada, isto significa que a incidência de erros críticos manteve-se a menos de metade; no entanto, uma proporção de 41% permanece substancial e significativa no contexto da avaliação da qualidade de testes. A seguir, apresenta-se o gráfico correspondente:

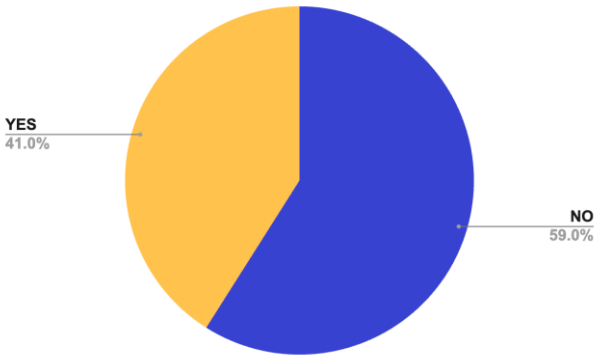


Figura 6 - Precisa de Gold Standard?

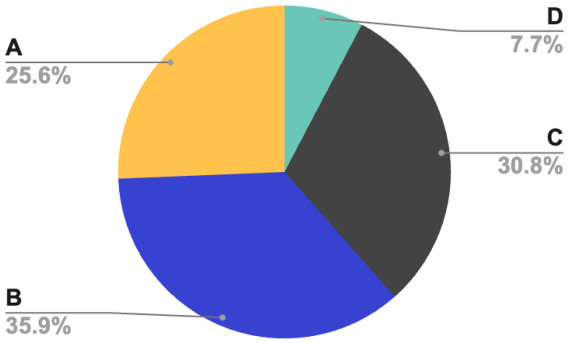


Figura 7 - Frequência de escolha de resposta correta

Com o intuito de aprimorar o processo de avaliação, foi elaborado um gráfico de frequência para as respostas corretas e, como é possível observar na figura 7, há uma maior aderência às opções A, B e C como resposta correta, ao contrário da opção D. Tal procedimento foi adotado dado que este tipo de análise gráfica fornece *insights* valiosos sobre a distribuição das respostas corretas, introduzindo, assim, uma camada adicional de análise que contribui para o refinamento e melhoria do output gerado pelo GPT-4.

5.2. Preenchimento de Espaços

O conjunto de questões de preenchimento de espaços foi composto por um total de 28 questões. Observaram-se melhorias significativas neste conjunto comparando com os resultados do banco de escolha múltipla, evidenciadas por uma distribuição em que 42,9% das questões apresentaram apenas um erro cada, enquanto 57,1% das questões não apresentaram erros.

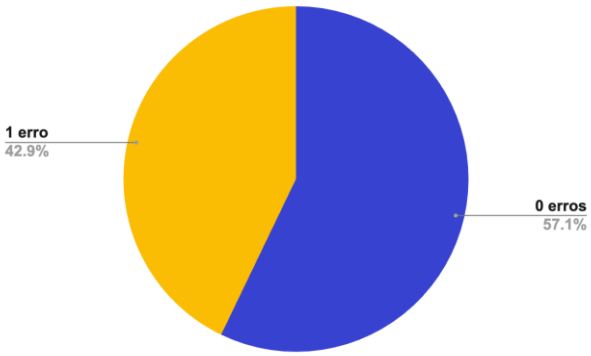


Figura 8- Frequência de erros por questão

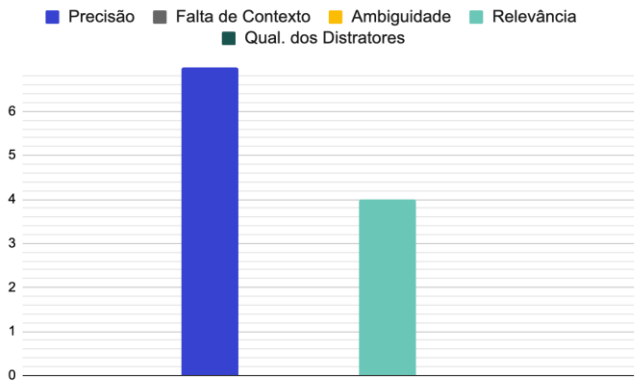


Figura 9 - Contagem dos tipos de erros

Estes resultados corroboram as observações da avaliadora, as quais podem ser atribuídas à estrutura das questões: a presença de uma frase contextual com um espaço em branco e uma pista facilitou a compreensão da função semântica dos termos alvo neste conjunto de questões por parte da avaliadora. No entanto, alguns erros relacionados com a precisão persistiram em 7 perguntas, enquanto no que toca à relevância, persistiram em 4 perguntas. Esses resultados sugerem que várias das diferenças linguísticas propostas pelo GPT-4 podem estar desatualizadas ou podem mesmo não existir.

Apesar disso, a necessidade por um *Gold Standard* manteve-se em níveis semelhantes, com 42,9% das respostas classificadas como positivas e 57,1% como negativas. Tal situação deve-se à gravidade dos erros identificados.

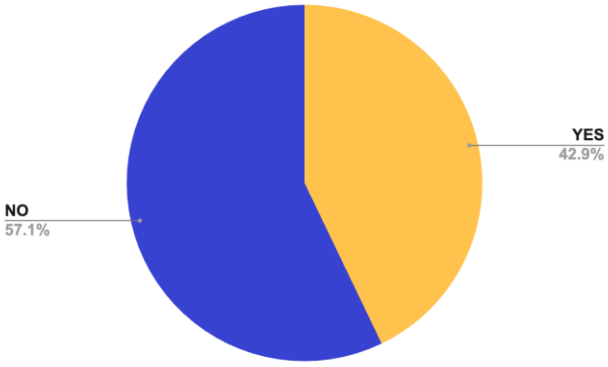


Figura 10 - Precisa de Gold Standard?

5.3. Transformação

As questões de transformação, que partilham semelhanças com o formato de preenchimento de espaços, apresentaram um desempenho superior em comparação com as questões de escolha múltipla. Contudo, ao contrário das questões de preenchimento de espaços, observou-se novamente a presença de questões com dois erros cada.

Com 70,4% das questões sem nenhum erro, os resultados sobressaem relativamente ao banco de questões de escolha múltipla, principalmente devido à contextualização frásica do termo. No entanto, registou-se ainda uma percentagem de 25,9% de questões com um erro e 3,7% com dois erros por questão. Embora estas percentagens sejam inferiores a metade do total, representam, contudo, uma fração significativa que merece atenção crítica, sobretudo tendo em conta que a preponderância destes erros está vinculada a questões de precisão e ambiguidade.

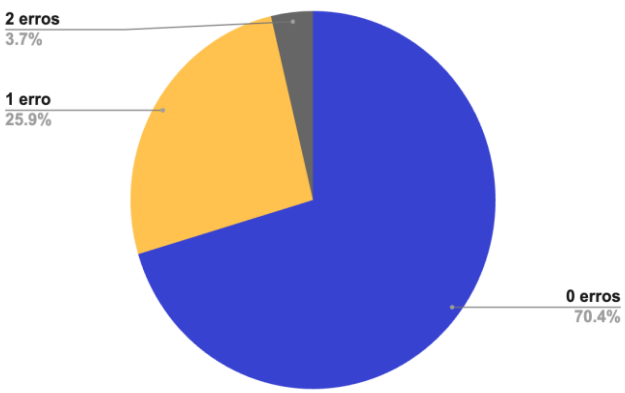


Figura 11- Frequência de erros por questão

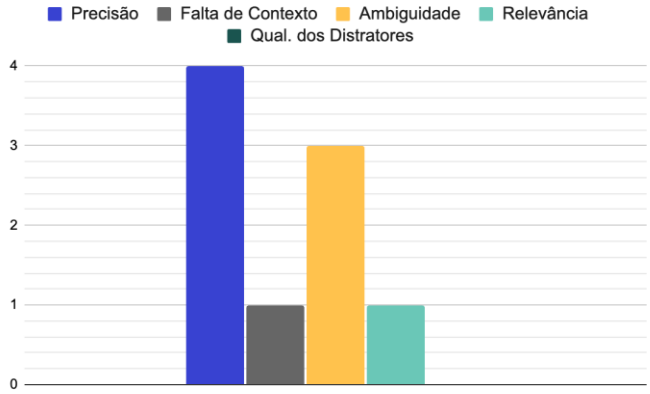


Figura 12 - Contagem dos tipos de erros

A necessidade de *Gold Standard* é evidenciada pelos resultados anteriormente discutidos, nos quais 26,7% das respostas indicaram "sim", enquanto 73,3% apontaram "não".

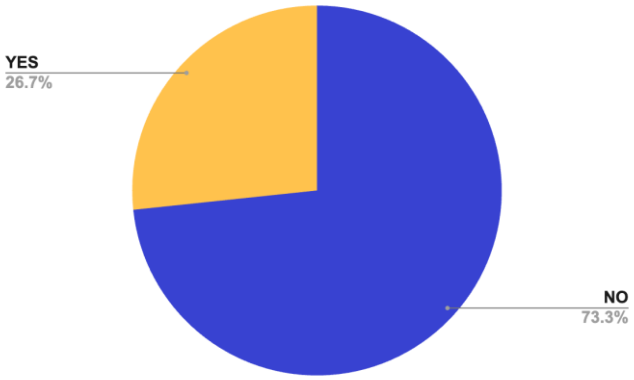


Figura 13 - Precisa de Gold Standard?

Embora os resultados obtidos neste banco de questões e no anterior sejam consideravelmente melhores do que os das questões de escolha múltipla, estes ainda estão longe do ideal, visto que persistem alguns problemas relacionados com a veracidade e relevância das questões. Tais erros são particularmente graves, considerando o propósito esperado destes testes.

5.4. Identificação da variante

O banco de questões focado na identificação da variante linguística compreendeu um total de 15 questões. Ao contrário dos demais conjuntos de questões, como referido na secção 4.5.5, este distinguiu-se por uma tentativa de integrar exemplos concretos realistas nos testes. Para tal, selecionaram-se frases que evidenciavam características distintivas do inglês britânico e americano, a partir de uma compilação de traduções realizadas para inglês na Unbabel. Diferentemente do procedimento adotado nas restantes secções, em que o GPT-4 gerou as questões independentemente, aqui, forneci diretamente essas frases, que foram posteriormente transformadas em exercícios para identificação das variantes linguísticas.

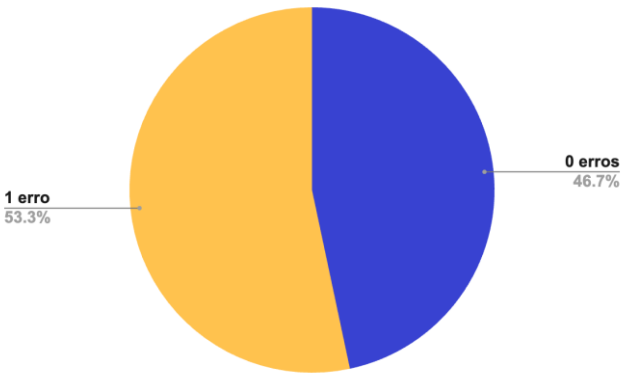


Figura 14- Frequência de erros por questão

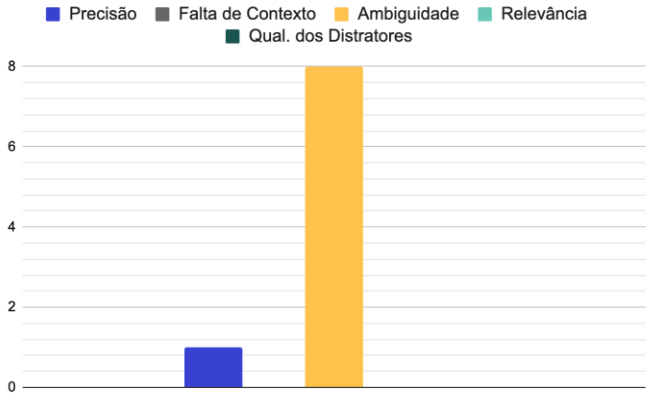


Figura 15 - Contagem dos tipos de erros

Esta metodologia registou os resultados mais favoráveis, em especial na redução de erros associados à precisão, que foi o tipo de erro mais recorrente nos outros bancos de questões. Esta melhoria atribui-se à contextualização aprimorada, em conjunto com o aumento do comprimento das frases. Observa-se que 46,7% das questões estão isentas de qualquer erro, e embora 53,3% das questões contemplem um erro por questão – não representando, por conseguinte, o melhor desempenho observado – nesta ocasião, os erros mais frequentes relacionam-se com a ambiguidade. Este fenómeno decorre da

intervenção do GPT-4, que cortou algumas das frases apresentadas devido ao seu comprimento, eliminado assim a parte da frase que continha o elemento de localização. No total, registou-se 1 questão com erros de precisão e 8 com erros de ambiguidade.

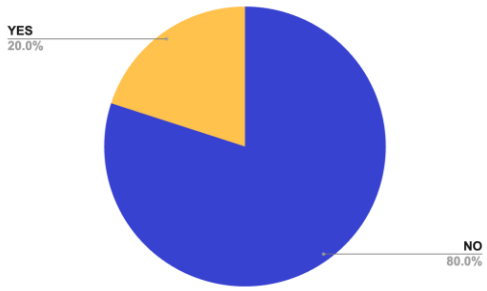


Figura 16 - Precisa de Gold Standard?

Os resultados relativos ao *Gold Standard* estão alinhados com os achados precedentes, registando-se apenas 20% das respostas como "sim" e 80% como "não".

5.5. Verdadeiro ou Falso

O banco de questões de verdadeiro ou falso foi composto por um total de 30 questões.

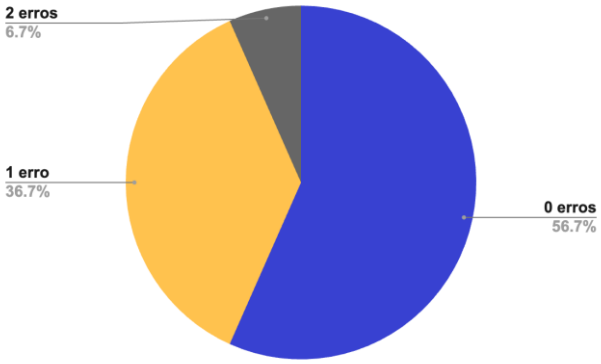


Figura 17 - Frequência de erros por questão

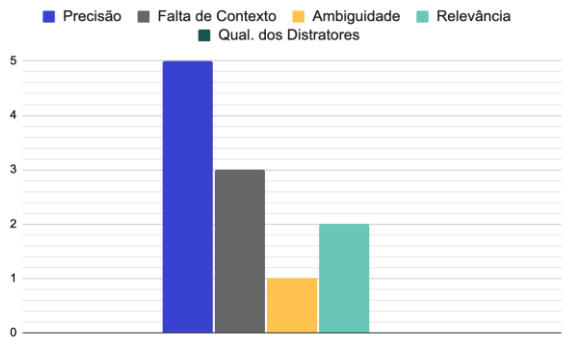


Figura 18 - Contagem dos tipos de erros

Sendo este tipo de questões semelhante em estilo às de escolha múltipla, novamente registaram-se mais problemas, com 36,7% das questões apresentando um erro cada e 6,7% com dois erros por questão. No entanto, observou-se uma percentagem de 56,7% de questões sem qualquer erro.

Observa-se novamente um aumento nos erros de precisão, decorrentes das diferenças linguísticas diretamente geradas e fornecidas nas questões pelo GPT-4, incluindo um total de 5 questões com erros de precisão, 3 questões com falta de contexto, 1 com ambiguidade e 2 com problemas de relevância.

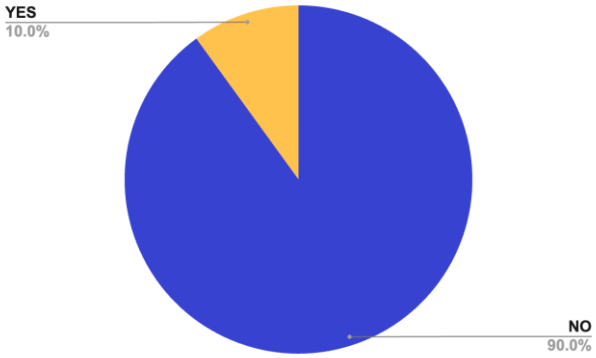


Figura 19 - Precisa de Gold Standard?

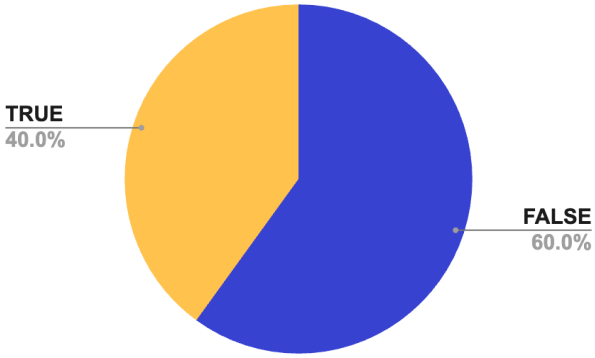


Figura 20 - Frequência de escolha de resposta correta

Este conjunto de questões registou a menor percentagem na necessidade de aplicação de um *Gold Standard*, com apenas 10% das questões a requererem melhorias significativas, enquanto que 90% não necessitou. Tal pode ser atribuído ao facto de este banco de questões incluir explicações para as respostas, o que, segundo a avaliadora, facilitou a compreensão dos exercícios.

Destaca-se ainda que o GPT-4 seguiu eficazmente as instruções para aleatorizar a resposta correta, com 60% das respostas corretas a serem "Falso" e 40% "Verdadeiro".

5.6. Resultados Gerais

TIPOS DE ERROS	Precisão	Falta de Contexto	Ambiguidade	Relevância	Qualidade dos Distratores
Escolha Múltipla	13	9	6	9	1
Preenchimento de Espaços	7	0	0	4	0
Transformação	4	1	3	1	0
Identificação da Variante	1	0	8	0	0
Verdadeiro ou Falso	5	3	1	2	0

Tabela 1 - Contagem de tipos de erros por banco de questão

NÚMERO DE ERROS POR QUESTÃO	0 erros	1 erro	2 erros	3 erros
Escolha Múltipla	10	20	8	1
Preenchimento de Espaços	16	12	0	0
Transformação	19	7	1	0
Identificação da Variante	7	8	0	0
Verdadeiro ou Falso	17	11	2	0

Tabela 2 - Contagem de número de erros por questão

Neste estudo, foram analisadas pela avaliadora um total de 139 questões. O desempenho do GPT-4 nesta tarefa revelou-se satisfatório, com 49,6% das perguntas isentas de erros, 41% apresentando um erro, 8,6% com dois erros e apenas 0,7% com três erros. Apesar de quase metade das questões estarem prontas para uso sem necessidade de correções, é relevante destacar que a maioria dos erros identificados diz respeito à precisão, um problema significativo que pode comprometer a integridade das questões.

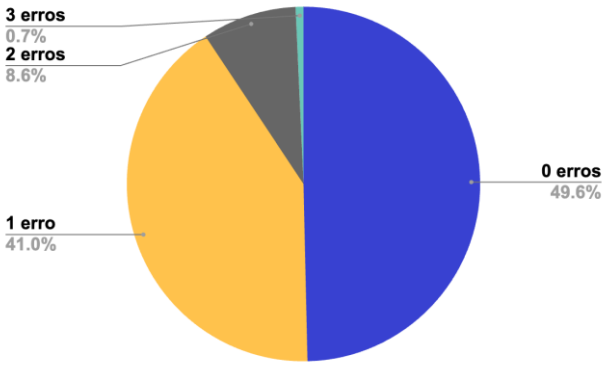


Figura 21 - Frequência de erros por questão geral

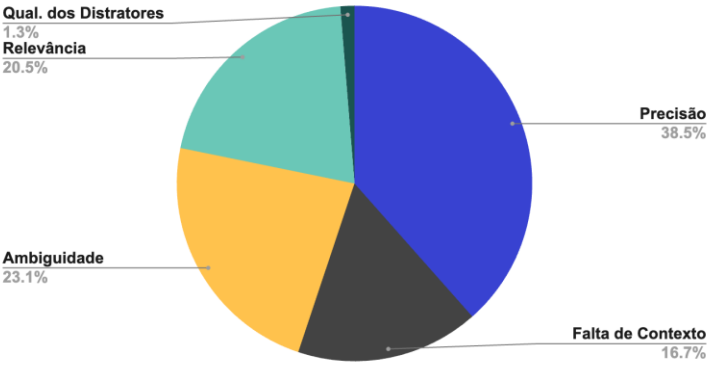


Figura 22 - Percentagem geral dos tipos de erros

Embora os resultados deste estudo tenham sido parcialmente satisfatórios, não se pode considerar que os objetivos ideais foram plenamente alcançados, dado que metade das perguntas apresentou erros e exigiu revisão humana. Assim, para garantir a integridade e atualidade dos bancos de questões, torna-se imperativo a implementação de processos de revisão e correção por parte de linguistas, visando verificar e aprimorar a qualidade dos *outputs* gerados pelo GPT-4.

A avaliadora expressou preocupações relativas à fiabilidade das perguntas, sublinhando a importância crítica de prover um contexto adequado às questões. Esta necessidade revelou-se particularmente notória no banco de questões de “Identificação da Variante”, onde se observou que frases mais extensas contribuíram significativamente para um melhor entendimento por parte da pessoa a fazer o teste e desempenho do modelo.

Adicionalmente, a avaliadora apontou alguns problemas relacionados com o uso de "linguagem definitiva", tal como perguntar "*What’s the equivalent of the term…?*" em vez de algo como "*What’s the more frequent use of the term…?*". Esta metodologia pode revelar-se problemática quando aplicada à avaliação linguística, uma vez que muitas das diferenças linguísticas, hoje em dia, não são percetíveis em termos de "preto e branco". Na realidade, tais diferenças envolvem frequentemente "nuances" de uso ou aderência a padrões linguísticos específicos. Formular questões em termos definitivos não contempla

adequadamente a complexidade e a variação no uso da língua em diferentes contextos, sublinhando a importância de as questões serem desenhadas de forma a acomodar as subtilezas das variações linguísticas.

5. *PROMPTS* MELHORADOS

Através dos resultados desta investigação, não só foi possível responder à questão inicialmente proposta, como também apurar a qualidade dos *prompts* utilizados e identificar potenciais falhas e pontos fortes, de acordo com o *feedback* fornecido pela avaliadora.

Assim sendo, conforme delineado na secção 4.5, os *prompts* empregados incluíam a finalidade da tarefa, proporcionando ao GPT-4 uma visão clara do seu objetivo; instruções detalhadas sobre o que era necessário realizar; um exemplo de elevada qualidade da tarefa para que o modelo pudesse seguir as instruções com precisão; e, quando relevante, a indicação da terminologia a ser empregada nos exercícios.

No entanto, é importante reconhecer que estes *prompts* apresentam limitações. Deram origem a *outputs* que manifestaram problemas de precisão, apresentavam uma carência de contexto e de informação crucial que se revelou confusa até para uma avaliadora nativa de inglês britânico. Assim sendo, é razoável supor que uma pessoa não nativa encontraria dificuldades semelhantes, possivelmente em maior grau.

Considerando os resultados obtidos, chegou-se à conclusão de que uma versão aprimorada dos *prompts* requer a incorporação de especificações e instruções adicionais, que se enumeram como segue:

1. Assegurar que, para cada exercício, seja providenciado um contexto em forma de frase. Este aspeto revela-se particularmente crítico nas questões de escolha múltipla e de verdadeiro ou falso, onde se observou uma maior incidência de termos alvo apresentados fora de contexto. Portanto, é indispensável incorporar o termo alvo de maneira adequada num contexto onde seja evidente o seu uso, de modo a fornecer as informações necessárias para a execução correta dos exercícios;

2. Orientar no sentido de evitar o uso de linguagem definitiva, optando por formulações frásicas mais abrangentes que permitam acomodar os padrões de frequência de uso das diferenças linguísticas que estão a ser avaliadas;
3. Incluir dados pertinentes de modo a garantir que o output seja relevante nos contextos específicos da Unbabel. Assim, nos processos de seleção e avaliação de tradutores, os resultados obtidos serão mais conclusivos e, no âmbito do treino, proporcionará aos editores, anotadores e revisores uma preparação mais aprofundada. Tal como observado no banco de questões de "Identificação de Variante", o fornecimento de dados e informações relevantes contribui positivamente para a performance do GPT-4.
4. Disponibilizar exemplos de referência de alta qualidade (*Gold Standard*): é imperativo que sejam fornecidos ao GPT-4 exemplos exemplares de tarefas de qualidade. Após a realização deste estudo, existem agora questões que podem servir como referências de qualidade para a criação de futuros testes linguísticos, concebidos por um linguista especialista em inglês britânico. A utilização deste tipo de questões tem o potencial de otimizar os *outputs* gerados.

7. CONCLUSÕES, LIMITAÇÕES E TRABALHOS FUTUROS

7.1. Conclusões

Neste relatório, investigou-se a seguinte questão central: Qual a eficácia do GPT-4 na criação de materiais de treino de falantes nativos de inglês não britânico para o domínio do inglês britânico? (Ver a seção 4.1) O propósito, ou o resultado ideal, era criar *prompts* de qualidade suficientemente alta para possibilitar a geração de bancos de questões capazes de testar e treinar tradutores e linguistas num regime completamente automático, minimizando a necessidade de intervenção humana (Ver a seção 4.1.1).

Para alcançar este objetivo, procedeu-se com a criação de *prompts* específicos para cada tipo de pergunta, com a intenção de desenvolver o *prompt* mais eficaz que produzisse os melhores *outputs*. Seguiu-se a compilação de todas as questões geradas pelo GPT-4 e a criação de uma tabela *scorecard*, contendo perguntas para a avaliação de qualidade das questões. Tal avaliação deveria ser realizada por uma linguista especialista em inglês britânico. Finalmente, uma análise consolidada dos dados recolhidos foi conduzida, com o intuito de avaliar a eficiência do GPT-4 nesta específica tarefa.

Após a realização desta investigação, constatou-se que o *output* do GPT-4 é significativamente defeituoso e não se mostra confiável para aplicação direta sem supervisão e ajuste humano adequados. Identificou-se que mais da metade das questões geradas apresentaram um ou mais erros. Além disso, as diferenças linguísticas fornecidas pelo GPT-4 demonstraram-se igualmente não confiáveis, com uma proporção significativa dos erros identificados relacionados à precisão. Problemas adicionais do *output* incluíram a falta de contexto e a ambiguidade nas questões de escolha múltipla, particularmente quando o termo focal estava desprovido de contexto, comprometendo a resolução das questões de maneira correta e eficiente. Similarmente, o uso de linguagem definitiva impactou negativamente a qualidade das questões, dado que muitas distinções entre o inglês britânico e americano não são absolutas.

Estes achados reforçam a necessidade de implementar sistemas de supervisão e mecanismos de *feedback* na geração de questões. No entanto, apesar desses desafios, acredita-se que o emprego do GPT-4 na criação de testes linguísticos pode constituir um método mais rápido e eficiente em comparação à elaboração manual integral das questões de teste. É importante reconhecer que esta perspetiva não é sustentada por dados empíricos e transcende o âmbito desta investigação. No entanto, argumento que a revisão e eventual correção de erros menores em um banco de questões já existente demanda significativamente menos tempo do que a elaboração meticulosa de cada questão de forma individual e detalhada. Baseia-se esta argumentação na premissa de que a integração das capacidades avançadas do GPT-4 com a expertise de um linguista pode resultar na criação de testes linguísticos de alta qualidade, se combinado com *prompts* melhorados (ver secção 6). Da mesma maneira, treinar LLM com dados de design e optimização de testes resultaria igualmente em melhorias nos resultados.

Em um contexto profissional, a sinergia entre as capacidades do GPT-4 e a expertise de tradutores internos pode representar uma oportunidade para as empresas investirem no aprimoramento da aptidão dos seus tradutores, abrindo novos postos de trabalho voltados para a criação automática e revisão de conteúdo gerado pelo GPT-4. Embora este estudo se concentre especificamente na diferenciação intralinguística entre duas variantes do inglês, seria pertinente realizar a mesma investigação virada para outras variantes linguísticas, como, por exemplo, entre o português europeu e o português brasileiro.

A realização de testes, acompanhada por avaliações e monitoramento contínuos, pode garantir o reforço e a atualização dos conhecimentos linguísticos dos tradutores de forma personalizada, contribuindo para assegurar traduções de alta qualidade a longo prazo. Neste sentido, o contexto educacional mostra-se fundamental; os testes devem incorporar elementos didáticos e contextos informativos que enriqueçam o conhecimento avaliado.

7.2. Limitações

Este estudo enfrentou diversas limitações, nomeadamente devido ao seu alcance e complexidade, aliados a restrições temporais. Tal foi evidenciado pelo número de participantes envolvidos, resumindo-se a apenas uma linguista de inglês britânico. A recolha e análise dos dados consumiram um tempo considerável, tornando impraticável a inclusão de mais participantes sem comprometer a conclusão do estudo no prazo estipulado. Adicionalmente, a presença de um único participante pode limitar a robustez e a conclusividade dos resultados.

Para alcançar resultados mais conclusivos, as categorias de avaliação propostas podem revelar-se insuficientes para uma apreciação objetiva e abrangente de todas as questões geradas pelo GPT-4, tendo em conta a especificidade do tema de investigação. A restrição temporal contribuiu para que as categorias definidas permanecessem vagas, sugerindo que uma elaboração de subcategorias focadas na severidade dos erros e em aspetos particulares dentro de cada categoria poderia enriquecer a análise. Isso permitiria

cobrir mais eficazmente as complexidades linguísticas, visando a obtenção de resultados integralmente conclusivos.

7.3. Trabalhos Futuros

Para futuras direções de pesquisa, seria relevante explorar o impacto que estes testes exercem sobre os tradutores e o seu conhecimento linguístico a longo prazo. Igualmente, seria enriquecedor testar estes *prompts* novos e aperfeiçoados, avaliando o seu *output* em comparação com o anterior. Ademais, importa analisar com maior detalhe os tipos de erros presentes nos *outputs* gerados pelo GPT-4, aprimorando as categorias já criadas e, eventualmente, desenvolvendo subcategorias que permitam uma classificação mais granular da gravidade dos erros. Desta forma, poder-se-ia facilitar a resolução dos problemas de maneira semi-automática e a longo prazo.

BIBLIOGRAFIA

Akan, M. F. (2017). *A Profile of the Grammatical Variation in British and American English*. Journal of English Language Teaching and Linguistics, 2(3), 201-214.

Algeo, J. (2006). *British or American English?: A Handbook of Word and Grammar Patterns*. Cambridge University Press.

Bailey, R. W. (2012). *Speaking American: a history of English in the United States*. OUP USA.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, England: Oxford University Press.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton. Reprint. Berlin and New York, 1957

D'Amelio, M. (2021). *British or American English? Awareness, preferences and use among university students of English,* [Università degli Studi di Padova]. https://thesis.unipd.it/handle/20.500.12608/40708

DeepLearning.AI. (2023, January 11). *Natural Language Processing (NLP) [A complete guide]*. https://www.deeplearning.ai/resources/natural-language-processing/

El Shazly, R. (2021). *Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. Expert Systems,* 38(3). doi:10.1111/exsy.12667

Foote, K. D. (2023, July 6). *A Brief history of Natural language processing* - DATAVERSITY. DATAVERSITY. https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/

Harris, L. R. (1979, August). *Experience with robot in 12 commercial, natural language data base query applications.* In Proceedings of the 6th international joint conference on Artificial intelligence-Volume 1 (pp. 365-371).

Hirschberg, J., & Manning, C. D. (2015). *Advances in natural language processing. Science*, 349(6245), 261-266.

Househ, M., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, M. G., Latifi, S., Aziz, S., Damseh, R., Alrazak, S. A., & Sheikh, J. (2023). L*arge Language models in medical Education: opportunities, challenges, and future directions.* JMIR Medical Education, 9, e48291. https://doi.org/10.2196/48291

Jones, K. S. (1994). *Natural Language Processing: A Historical Review*. Current Issues in Computational Linguistics: In Honour of Don Walker, 3–16. doi:10.1007/978-0-585-35958-8_1

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). *ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences*, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Knowles, M. S., Holton, E. F., & Swanson, R. A. (1998). *The Adult Learner : the definitive classic in adult education and human resource development.*

Manning, C. D. (2022, abril). *Human language understanding & reasoning.* American Academy of Arts & Sciences. https://www.amacad.org/publication/human-language-understanding-reasoning

McArthur, T. (1992). *The Oxford companion to the English language*. UK: Oxford University Press.

Millward, C. M. (1988). *A biography of the English language*. http://ci.nii.ac.jp/ncid/BA10555683

Lester, B., Al-Rfou, R., & Constant, N. (2021). *The Power of Scale for Parameter-Efficient Prompt Tuning*. *Conference on Empirical Methods in Natural Language Processing.*

Luu, C. (2022). *When did colonial America gain linguistic independence?* JSTOR Daily. https://daily.jstor.org/colonial-america-gain-linguistic-independence/

OpenAI. (n.d. a) *Pricing*. https://openai.com/api/pricing

OpenAI. (n.d. b) *Models.* OpenAI Platform. https://platform.openai.com/docs/models/overview

OpenAI. (n.d. c) *Prompt Engineering*. OpenAI Platform. https://platform.openai.com/docs/guides/prompt-engineering

OpenAI. (2023, março) *GPT-4*. https://openai.com/index/gpt-4-research

Ormrod, J. E. (2011). *Human learning*. Prentice Hall.

Peters, P. (2004). *The Cambridge Guide to English Usage.* Cambridge University Press.

Scott, J. C. (2004). *American and British Business—related spelling differences.* Business Communication Quarterly, 67(2)

Shrivastava, A., & Jain, J. K. (2020). *Natural language processing: A literature survey*. International Journal of Communication and Information Technology 2, 1(2), 15-20.

Sorensen, T., Robinson, J., Rytting, C. M., Shaw, A. G., Rogers, K. J., Delorey, A. P., & Wingate, D.(2022). *An Information-theoretic Approach to Prompt Engineering Without*

Ground Truth Labels. Political Analysis, *31*, 337 - 351. https://www.semanticscholar.org/reader/d53e70d834243d3d8d4b621c0c52dfec26081155

TeachingEnglish. (n.d.). *Transformation exercise*. https://www.teachingenglish.org.uk/professional-development/teachers/knowing-subject/t-w/transformation-exercise

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). *What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments*, 10(1). https://doi.org/10.1186/s40561-023-00237-x

Tu, X., Zou, J., Su, W. J., & Zhang, L. (2023). *What Should Data Science Education Do with Large Language Models?* https://arxiv.org/pdf/2307.02792.pdf

Unbabel. (n.d. a) *Meet our Community*. Unbabel. https://unbabel.com/our-community/

Unbabel. (n.d. b) *Evaluations: the definitive guide*. Unbabel Support. https://help.unbabel.com/hc/en-us/articles/360048031033-Evaluations-the-definitive-guide

Unbabel. (n.d. c) *Evaluations: the definitive guide*. Unbabel Support. https://help.unbabel.com/hc/en-us/articles/360048031033-Evaluations-the-definitive-guide

Unbabel. (n.d. d) *Essential guidelines for review tasks - before and during your work.* Unbabel Support. https://help.unbabel.com/hc/en-us/articles/13481848238871-Essential-guidelines-for-review-tasks-before-and-during-your-work

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). *Attention is all you need*. arXiv.org. https://arxiv.org/abs/1706.03762

Velásquez-Henao, J.D., Franco-Cardona, C.J., & Cadavid-Higuita, L. (2023). *Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. DYNA*. e Lo, L.S., (2023) The art and science of prompt engineering: a new literacy in the information age, Internet Ref. Serv. Q. DOI: https://doi.org/10.1080/10875301.2023.2227621.

Verma, P., & De Vynck, G. (2023, June 5). *ChatGPT took their jobs. Now they walk dogs and fix air conditioners.* Washington Post. https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/

Vieira, R., & Lopes, L. (2010). *Processamento de linguagem natural e o tratamento computacional de linguagens científicas. Em corpora*, 183.

What is MQM?. MQM (Multidimensional Quality Metrics). (2021, Maio). https://themqm.org/

LISTA DE FIGURAS

LISTA DE TABELAS

ANEXOS

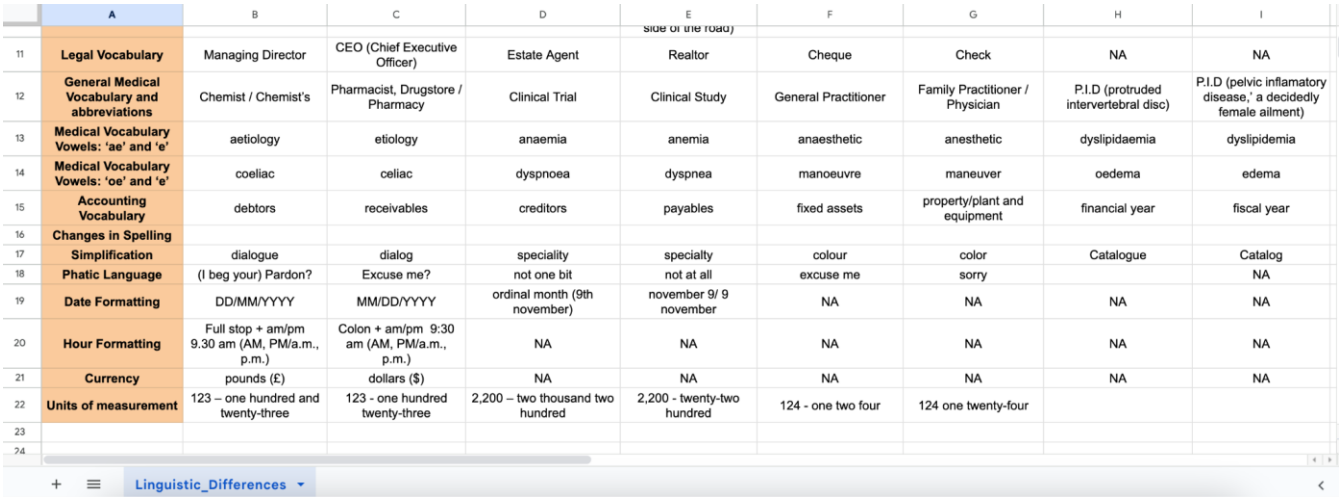
Tabela de diferenças linguísticas

Differences	British v1	American v1	British v2	American v2	British v3	American v3	British v4	American v4
Verb Contractions Involving Modals	can not	cannot	musn't	shouldn't	needn't/ need not	doesn't have to/didn't have to	ough not to/ oughtn't to	shouldn't
Perfect Aspect	have been going on	was going on	hadn't had	didn't have	NA	NA	NA	NA
Adverb Order (Disjuncts)	(uses V+AD) has certainly	(uses AD+V) certainly has	done beautifully	beautifully done	made oddly	NA	NA	NA
Modifying Adjectives and adverbs	proper(ly)	really/very	NA	NA	NA	NA	NA	NA
Compound Nouns	(Always uses hyphens for compound adjectives before a noun) a print-out presentation	(Tends to make compound words) a printout presentation	(Gerund + Noun) Skipping Rope	(Infinitive + Noun) Jump rope	The artists enjoyed the *sing-along* on the stage.	The singers had fun with the singalong on the campfire.	dialing tone	dial tone
Collective Nouns	(Usually plural – A group is typically thought of as a group of individuals) The committee *were* unable to agree.	(Almost always singular, excluding plural sports teams and band names) The committee *was* unable to agree.	Muse *are* a great band.	Muse *is* a great band.	The Beatles *are* playing at Wembley.	The Beatles *are* playing at Wembley.	One Direction *are* a great band.	One Direction *is* a great band.
Prepositions	*At* the weekend	*On* the weekend	Play *in* a team	Play *on* a team	*In* hospital	*In the* hospital	Monday to Friday	Monday through Friday
Noun Synonymy	Biscuit	Crackers	Chips	Fries/French Fries	Flat	Apartment	Candyfloss	Cotton Candy
Homonymy	wash up (wash dishes)	wash up (wash someone's body)	pavement (side of the road)	pavement (the road itself where vehicles drive - sidewalk is the side of the road)	faculty (university school or college)	faculty (teaching staff)	purse (what americans refer as a wallet)	purse (a woman's handbag)

Linguistic_Differences

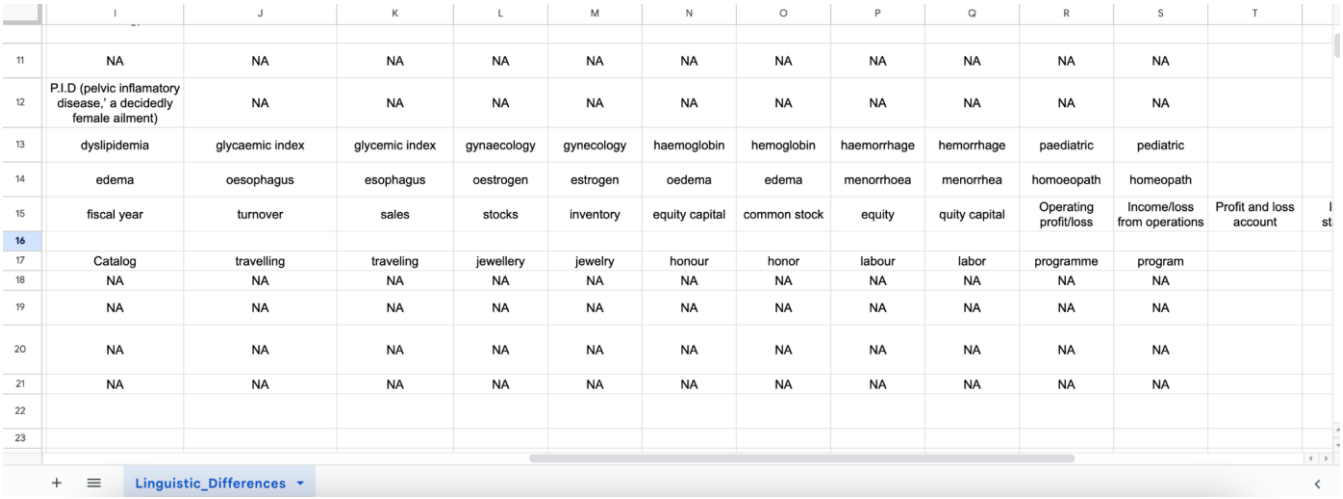
British v5	American v5	British v6	American v6	British v7	American v7	British v8	American v8	British v9	Americanv9
used not to/usen't to	didn't use to	will/ shall	be going to	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
racing car	racecar	rowing boat	rowboat	barber's shop	barbershop	cookery book	cookbook	skimmed milk	skim milk
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Fill in a form	Fill out a form	Write to someone	Write someone	At the back	In the back	NA	NA	NA	NA
Football	Socker	University	College	Holiday	Vacation	Timetable	Schedule	Postbox	Mailbox
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Linguistic_Differences



	A	B	C	D	E	F	G	H	I
					side of the road)				
11	**Legal Vocabulary**	Managing Director	CEO (Chief Executive Officer)	Estate Agent	Realtor	Cheque	Check	NA	NA
12	**General Medical Vocabulary and abbreviations**	Chemist / Chemist's	Pharmacist, Drugstore / Pharmacy	Clinical Trial	Clinical Study	General Practitioner	Family Practitioner / Physician	P.I.D (protruded intervertebral disc)	P.I.D (pelvic inflamatory disease,' a decidedly female ailment)
13	**Medical Vocabulary Vowels: 'ae' and 'e'**	aetiology	etiology	anaemia	anemia	anaesthetic	anesthetic	dyslipidaemia	dyslipidemia
14	**Medical Vocabulary Vowels: 'oe' and 'e'**	coeliac	celiac	dyspnoea	dyspnea	manoeuvre	maneuver	oedema	edema
15	**Accounting Vocabulary**	debtors	receivables	creditors	payables	fixed assets	property/plant and equipment	financial year	fiscal year
16	**Changes in Spelling**								
17	**Simplification**	dialogue	dialog	speciality	specialty	colour	color	Catalogue	Catalog
18	**Phatic Language**	(I beg your) Pardon?	Excuse me?	not one bit	not at all	excuse me	sorry		NA
19	**Date Formatting**	DD/MM/YYYY	MM/DD/YYYY	ordinal month (9th november)	november 9/ 9 november	NA	NA	NA	NA
20	**Hour Formatting**	Full stop + am/pm 9.30 am (AM, PM/a.m., p.m.)	Colon + am/pm 9:30 am (AM, PM/a.m., p.m.)	NA	NA	NA	NA	NA	NA
21	**Currency**	pounds (£)	dollars ($)	NA	NA	NA	NA	NA	NA
22	**Units of measurement**	123 – one hundred and twenty-three	123 - one hundred twenty-three	2,200 – two thousand two hundred	2,200 - twenty-two hundred	124 - one two four	124 one twenty-four		
23									
24									

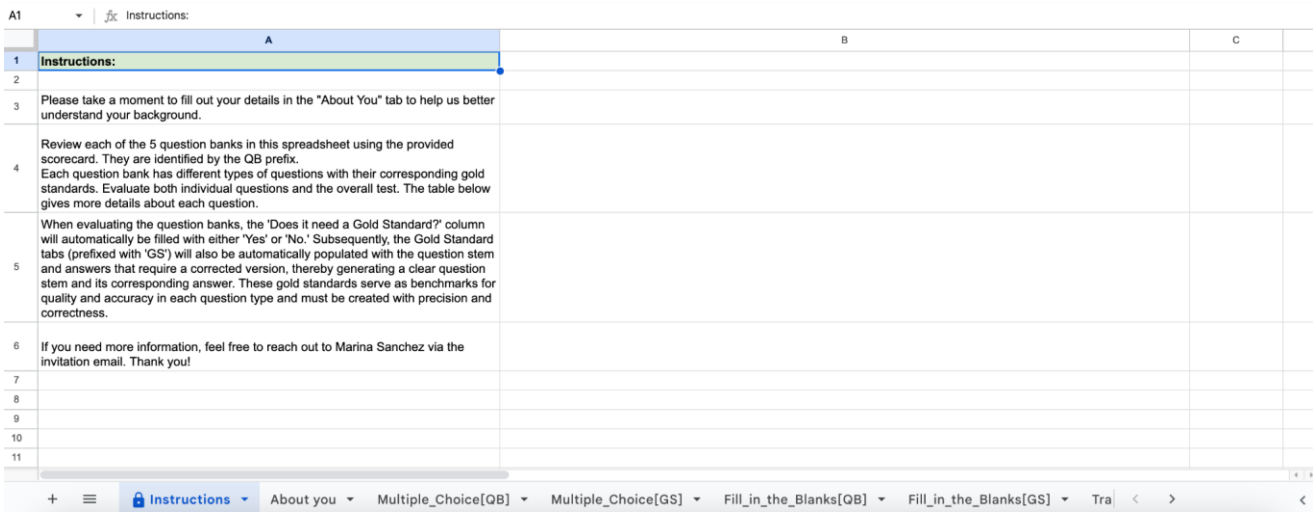
Linguistic_Differences



	I	J	K	L	M	N	O	P	Q	R	S	T
11	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
12	P.I.D (pelvic inflamatory disease,' a decidedly female ailment)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
13	dyslipidemia	glycaemic index	glycemic index	gynaecology	gynecology	haemoglobin	hemoglobin	haemorrhage	hemorrhage	paediatric	pediatric	
14	edema	oesophagus	esophagus	oestrogen	estrogen	oedema	edema	menorrhoea	menorrhea	homoeopath	homeopath	
15	fiscal year	turnover	sales	stocks	inventory	equity capital	common stock	equity	quity capital	Operating profit/loss	Income/loss from operations	Profit and loss account
16												
17	Catalog	travelling	traveling	jewellery	jewelry	honour	honor	labour	labor	programme	program	
18	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
19	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
21	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
22												
23												

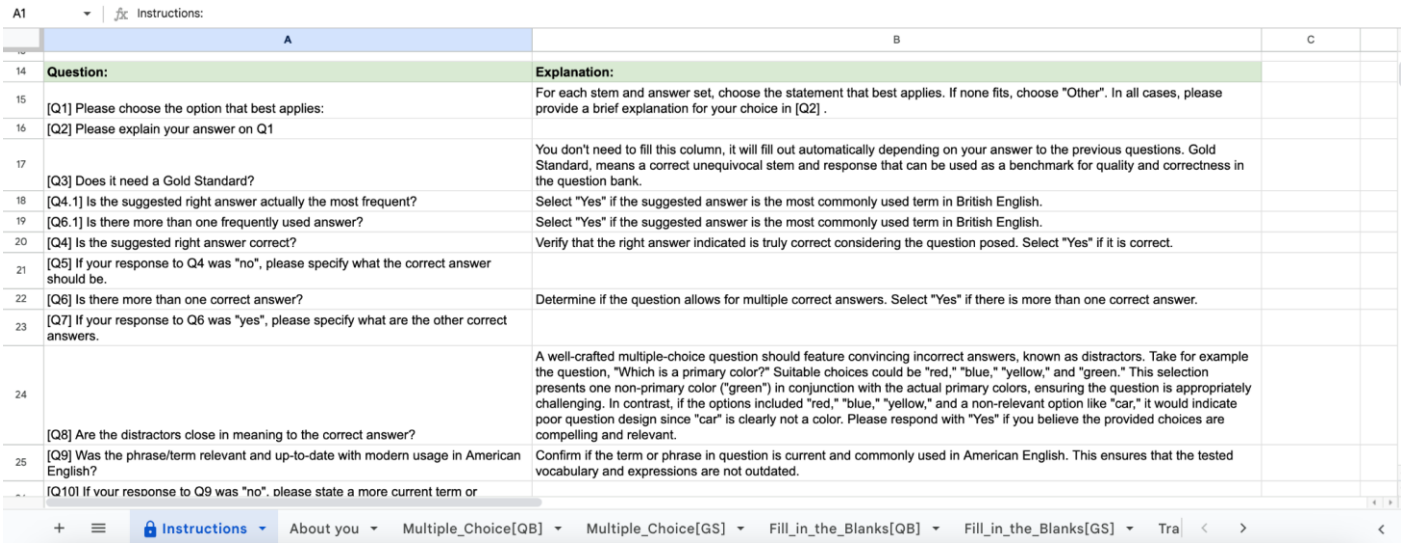
Linguistic_Differences

Scorecard



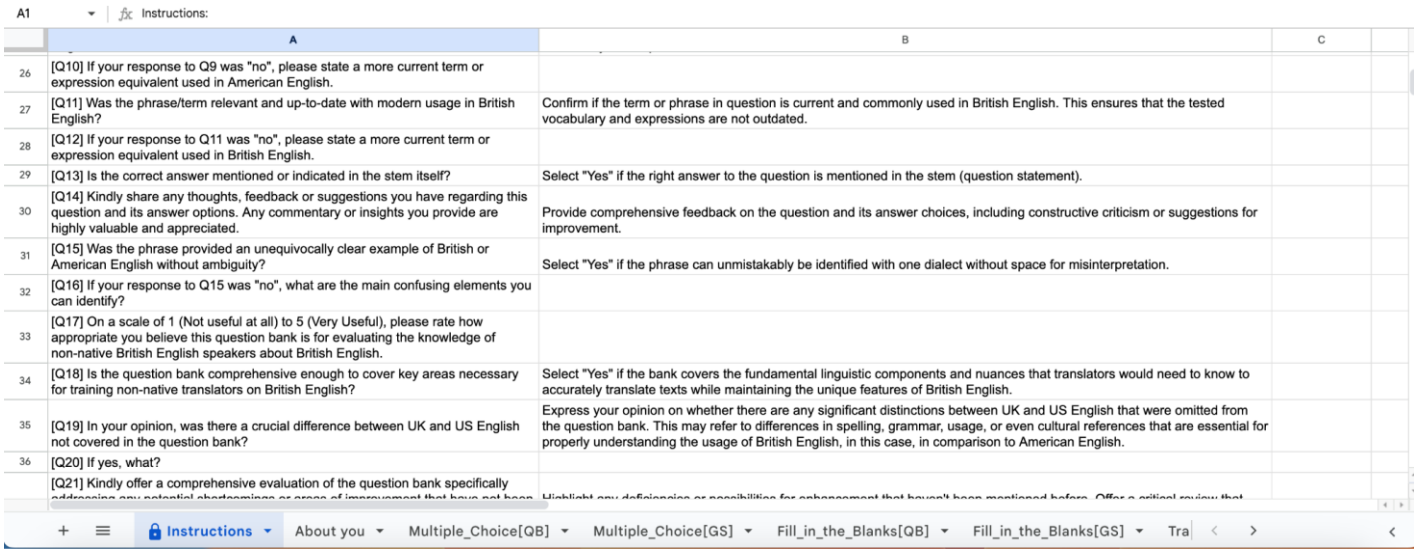
	A	B	C
1	**Instructions:**		
2			
3	Please take a moment to fill out your details in the "About You" tab to help us better understand your background.		
4	Review each of the 5 question banks in this spreadsheet using the provided scorecard. They are identified by the QB prefix. Each question bank has different types of questions with their corresponding gold standards. Evaluate both individual questions and the overall test. The table below gives more details about each question.		
5	When evaluating the question banks, the 'Does it need a Gold Standard?' column will automatically be filled with either 'Yes' or 'No.' Subsequently, the Gold Standard tabs (prefixed with 'GS') will also be automatically populated with the question stem and answers that require a corrected version, thereby generating a clear question stem and its corresponding answer. These gold standards serve as benchmarks for quality and accuracy in each question type and must be created with precision and correctness.		
6	If you need more information, feel free to reach out to Marina Sanchez via the invitation email. Thank you!		
7			
8			
9			
10			
11			

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Tra



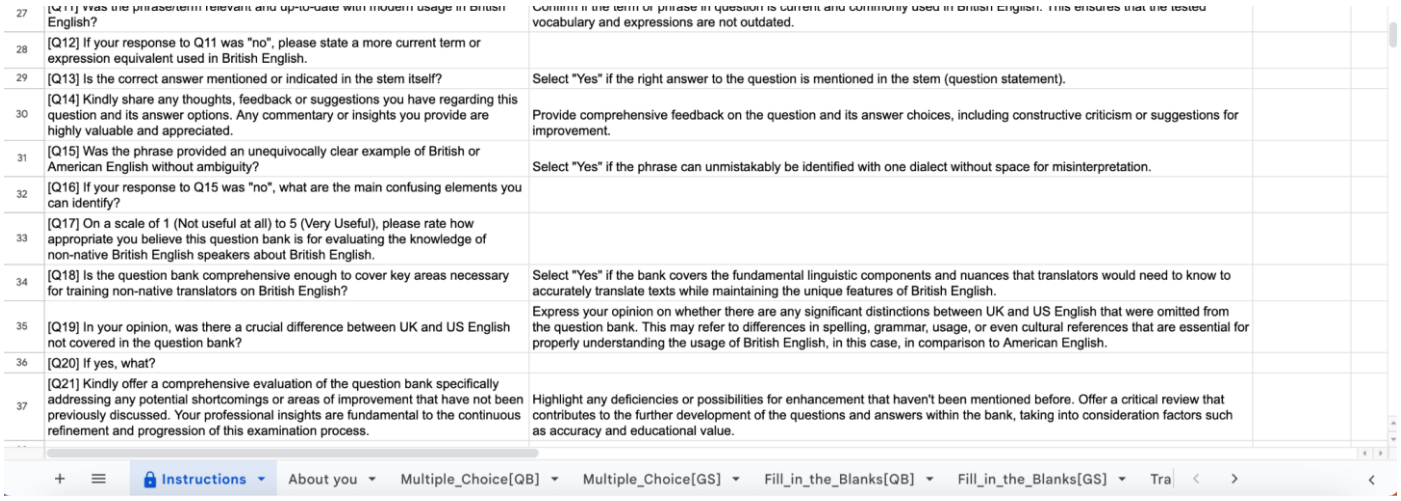
	A	B	C
14	**Question:**	**Explanation:**	
15	[Q1] Please choose the option that best applies:	For each stem and answer set, choose the statement that best applies. If none fits, choose "Other". In all cases, please provide a brief explanation for your choice in [Q2] .	
16	[Q2] Please explain your answer on Q1		
17	[Q3] Does it need a Gold Standard?	You don't need to fill this column, it will fill out automatically depending on your answer to the previous questions. Gold Standard, means a correct unequivocal stem and response that can be used as a benchmark for quality and correctness in the question bank.	
18	[Q4.1] Is the suggested right answer actually the most frequent?	Select "Yes" if the suggested answer is the most commonly used term in British English.	
19	[Q6.1] Is there more than one frequently used answer?	Select "Yes" if the suggested answer is the most commonly used term in British English.	
20	[Q4] Is the suggested right answer correct?	Verify that the right answer indicated is truly correct considering the question posed. Select "Yes" if it is correct.	
21	[Q5] If your response to Q4 was "no", please specify what the correct answer should be.		
22	[Q6] Is there more than one correct answer?	Determine if the question allows for multiple correct answers. Select "Yes" if there is more than one correct answer.	
23	[Q7] If your response to Q6 was "yes", please specify what are the other correct answers.		
24	[Q8] Are the distractors close in meaning to the correct answer?	A well-crafted multiple-choice question should feature convincing incorrect answers, known as distractors. Take for example the question, "Which is a primary color?" Suitable choices could be "red," "blue," "yellow," and "green." This selection presents one non-primary color ("green") in conjunction with the actual primary colors, ensuring the question is appropriately challenging. In contrast, if the options included "red," "blue," "yellow," and a non-relevant option like "car," it would indicate poor question design since "car" is clearly not a color. Please respond with "Yes" if you believe the provided choices are compelling and relevant.	
25	[Q9] Was the phrase/term relevant and up-to-date with modern usage in American English?	Confirm if the term or phrase in question is current and commonly used in American English. This ensures that the tested vocabulary and expressions are not outdated.	
	[Q10] If your response to Q9 was "no", please state a more current term or		

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Tra



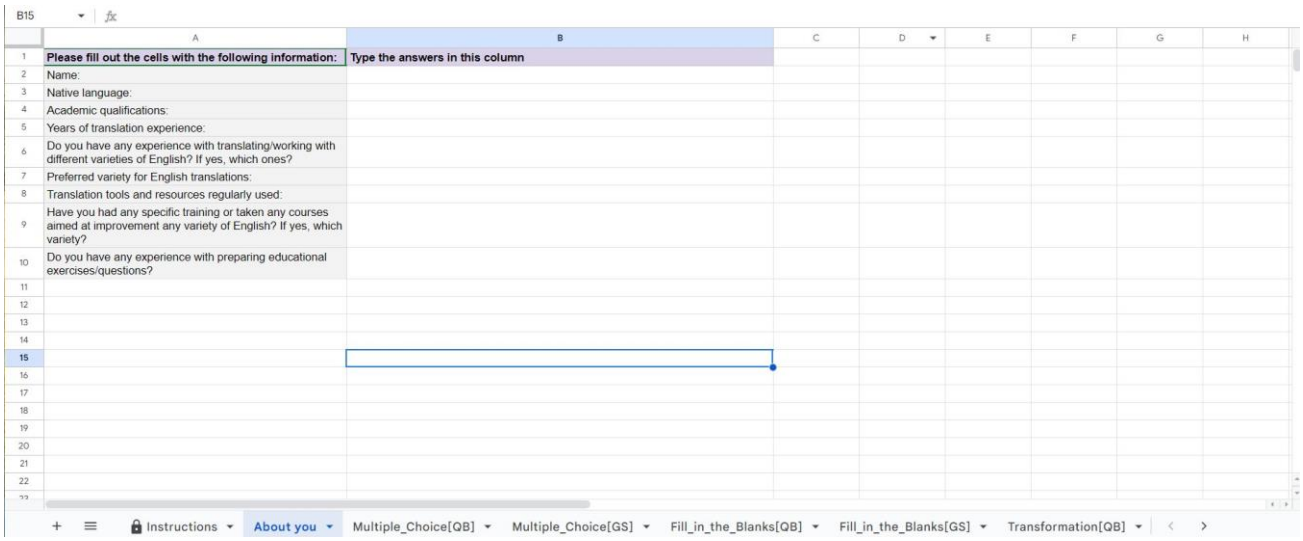
	A	B	C
26	[Q10] If your response to Q9 was "no", please state a more current term or expression equivalent used in American English.		
27	[Q11] Was the phrase/term relevant and up-to-date with modern usage in British English?	Confirm if the term or phrase in question is current and commonly used in British English. This ensures that the tested vocabulary and expressions are not outdated.	
28	[Q12] If your response to Q11 was "no", please state a more current term or expression equivalent used in British English.		
29	[Q13] Is the correct answer mentioned or indicated in the stem itself?	Select "Yes" if the right answer to the question is mentioned in the stem (question statement).	
30	[Q14] Kindly share any thoughts, feedback or suggestions you have regarding this question and its answer options. Any commentary or insights you provide are highly valuable and appreciated.	Provide comprehensive feedback on the question and its answer choices, including constructive criticism or suggestions for improvement.	
31	[Q15] Was the phrase provided an unequivocally clear example of British or American English without ambiguity?	Select "Yes" if the phrase can unmistakably be identified with one dialect without space for misinterpretation.	
32	[Q16] If your response to Q15 was "no", what are the main confusing elements you can identify?		
33	[Q17] On a scale of 1 (Not useful at all) to 5 (Very Useful), please rate how appropriate you believe this question bank is for evaluating the knowledge of non-native British English speakers about British English.		
34	[Q18] Is the question bank comprehensive enough to cover key areas necessary for training non-native translators on British English?	Select "Yes" if the bank covers the fundamental linguistic components and nuances that translators would need to know to accurately translate texts while maintaining the unique features of British English.	
35	[Q19] In your opinion, was there a crucial difference between UK and US English not covered in the question bank?	Express your opinion on whether there are any significant distinctions between UK and US English that were omitted from the question bank. This may refer to differences in spelling, grammar, usage, or even cultural references that are essential for properly understanding the usage of British English, in this case, in comparison to American English.	
36	[Q20] If yes, what?		
	[Q21] Kindly offer a comprehensive evaluation of the question bank specifically addressing any potential shortcomings or areas of improvement that have not been	Highlight any deficiencies or possibilities for enhancement that haven't been mentioned before. Offer a critical review that	

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Tra



27	English?	vocabulary and expressions are not outdated.	
28	[Q12] If your response to Q11 was "no", please state a more current term or expression equivalent used in British English.		
29	[Q13] Is the correct answer mentioned or indicated in the stem itself?	Select "Yes" if the right answer to the question is mentioned in the stem (question statement).	
30	[Q14] Kindly share any thoughts, feedback or suggestions you have regarding this question and its answer options. Any commentary or insights you provide are highly valuable and appreciated.	Provide comprehensive feedback on the question and its answer choices, including constructive criticism or suggestions for improvement.	
31	[Q15] Was the phrase provided an unequivocally clear example of British or American English without ambiguity?	Select "Yes" if the phrase can unmistakably be identified with one dialect without space for misinterpretation.	
32	[Q16] If your response to Q15 was "no", what are the main confusing elements you can identify?		
33	[Q17] On a scale of 1 (Not useful at all) to 5 (Very Useful), please rate how appropriate you believe this question bank is for evaluating the knowledge of non-native British English speakers about British English.		
34	[Q18] Is the question bank comprehensive enough to cover key areas necessary for training non-native translators on British English?	Select "Yes" if the bank covers the fundamental linguistic components and nuances that translators would need to know to accurately translate texts while maintaining the unique features of British English.	
35	[Q19] In your opinion, was there a crucial difference between UK and US English not covered in the question bank?	Express your opinion on whether there are any significant distinctions between UK and US English that were omitted from the question bank. This may refer to differences in spelling, grammar, usage, or even cultural references that are essential for properly understanding the usage of British English, in this case, in comparison to American English.	
36	[Q20] If yes, what?		
37	[Q21] Kindly offer a comprehensive evaluation of the question bank specifically addressing any potential shortcomings or areas of improvement that have not been previously discussed. Your professional insights are fundamental to the continuous refinement and progression of this examination process.	Highlight any deficiencies or possibilities for enhancement that haven't been mentioned before. Offer a critical review that contributes to the further development of the questions and answers within the bank, taking into consideration factors such as accuracy and educational value.	

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Tra

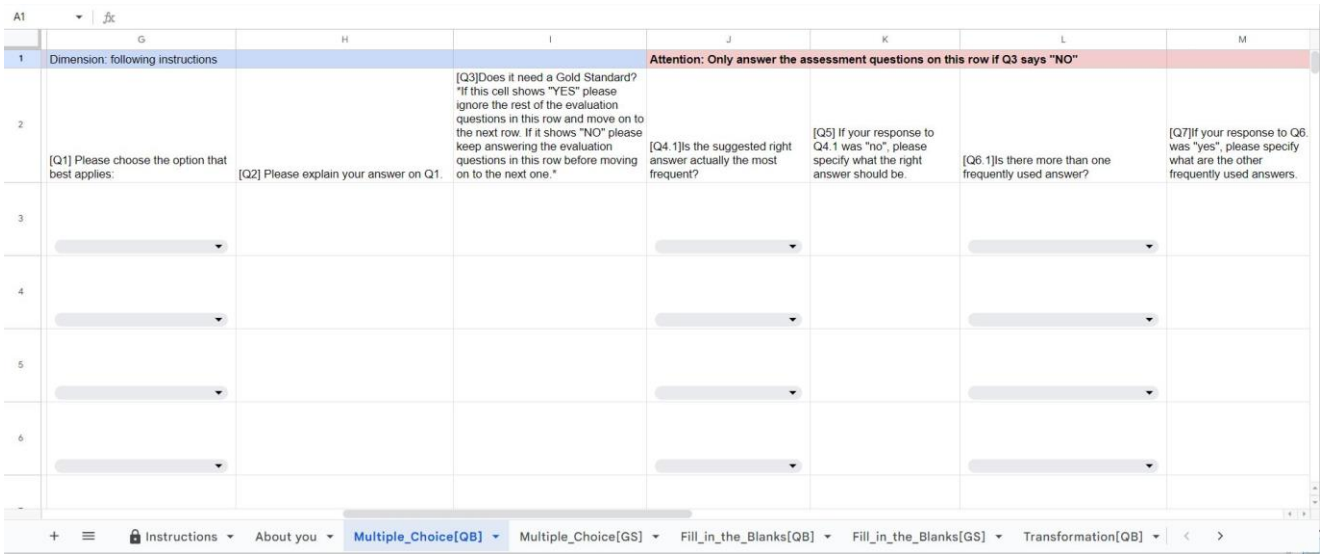


	A	B
1	**Please fill out the cells with the following information:**	**Type the answers in this column**
2	Name:	
3	Native language:	
4	Academic qualifications:	
5	Years of translation experience:	
6	Do you have any experience with translating/working with different varieties of English? If yes, which ones?	
7	Preferred variety for English translations:	
8	Translation tools and resources regularly used:	
9	Have you had any specific training or taken any courses aimed at improvement any variety of English? If yes, which variety?	
10	Do you have any experience with preparing educational exercises/questions?	

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Transformation[QB]

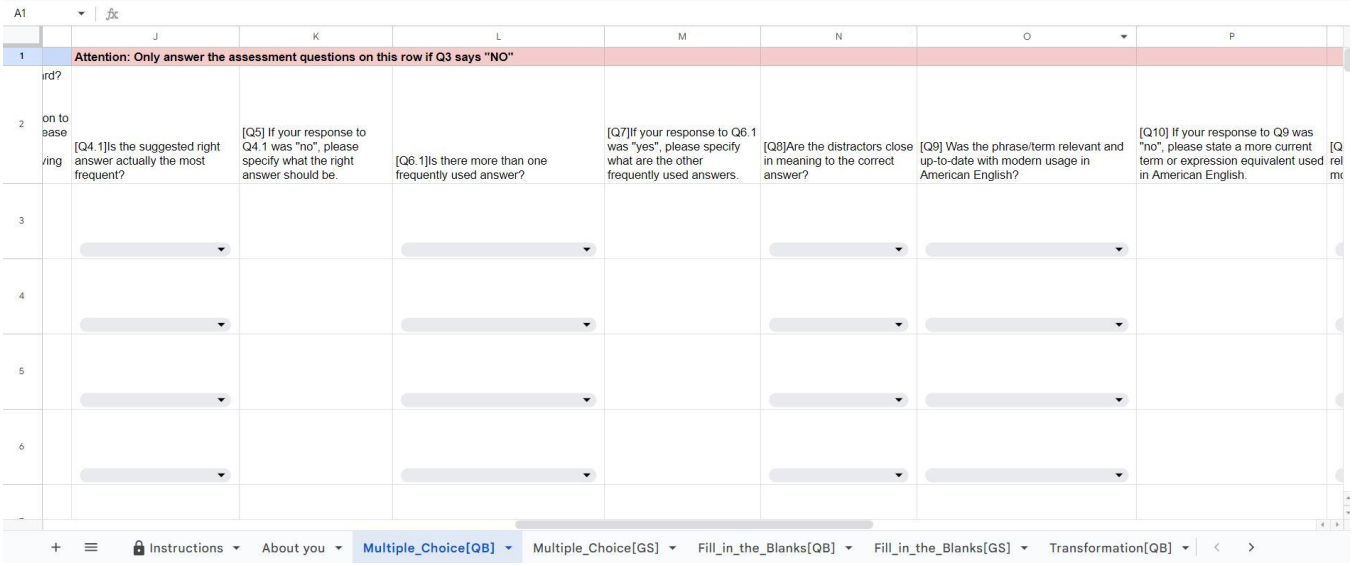
	A	B	C	D	E	F	G
1							Dimension: following instructions
2	**Stem (question statement)**	Answer_A	Answer_B	Answer_C	Answer_D	Suggested right answer	[Q1] Please choose the option tha best applies:
3	Which of the following British English expressions is most frequently used as the equivalent to the American English expression "I can't find my keys anywhere."?	I can not find my keys anywhere.	I mustn't find my keys anywhere.	I haven't found my keys anywhere.	I cannot find my keys anywhere.	D	
4	Which of the following British English expressions is most frequently used as the equivalent to the American English expression "You don't have to wear a tie."?	You haven't to wear a tie.	You aren't wearing a tie.	You needn't wear a tie.	You mustn't wear a tie.	C	
5	Which of the following British English expressions is most frequently used as the equivalent to the American English expression "The project was going on for months."?	The project is going on for months.	The project has been going on for months.	The project was been going on for months.	The project had been going on for months.	B	
6	Which of the following British English expressions is most frequently used as the equivalent to the American English expression "She certainly has finished the report."?	She has certainly finished the report.	She certainly finished the report.	She finished the report certainly.	She has finished the report certainly.	A	
7	Which of the following British English expressions is most frequently used as						

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Transformation[QB]



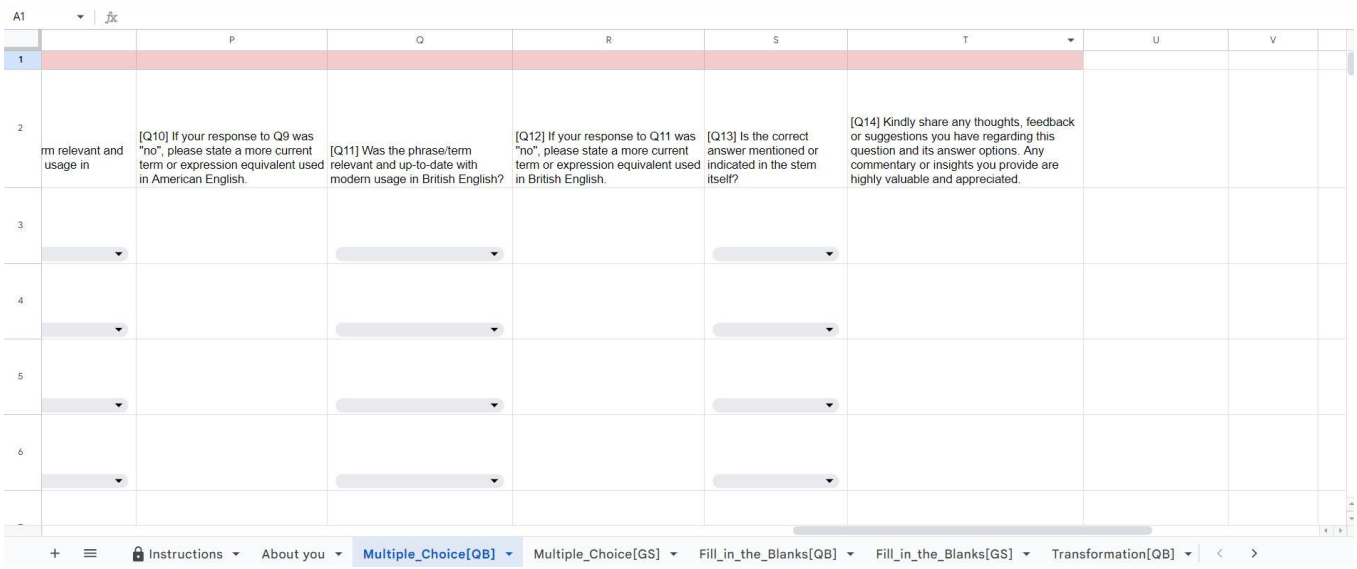
	G	H	I	J	K	L	M
1	Dimension: following instructions			Attention: Only answer the assessment questions on this row if Q3 says "NO"			
2	[Q1] Please choose the option that best applies:	[Q2] Please explain your answer on Q1.	[Q3]Does it need a Gold Standard? "If this cell shows "YES" please ignore the rest of the evaluation questions in this row and move on to the next row. If it shows "NO" please keep answering the evaluation questions in this row before moving on to the next one."	[Q4.1]Is the suggested right answer actually the most frequent?	[Q5] If your response to Q4.1 was "no", please specify what the right answer should be.	[Q6.1]Is there more than one frequently used answer?	[Q7]If your response to Q6. was "yes", please specify what are the other frequently used answers.
3							
4							
5							
6							

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Transformation[QB]



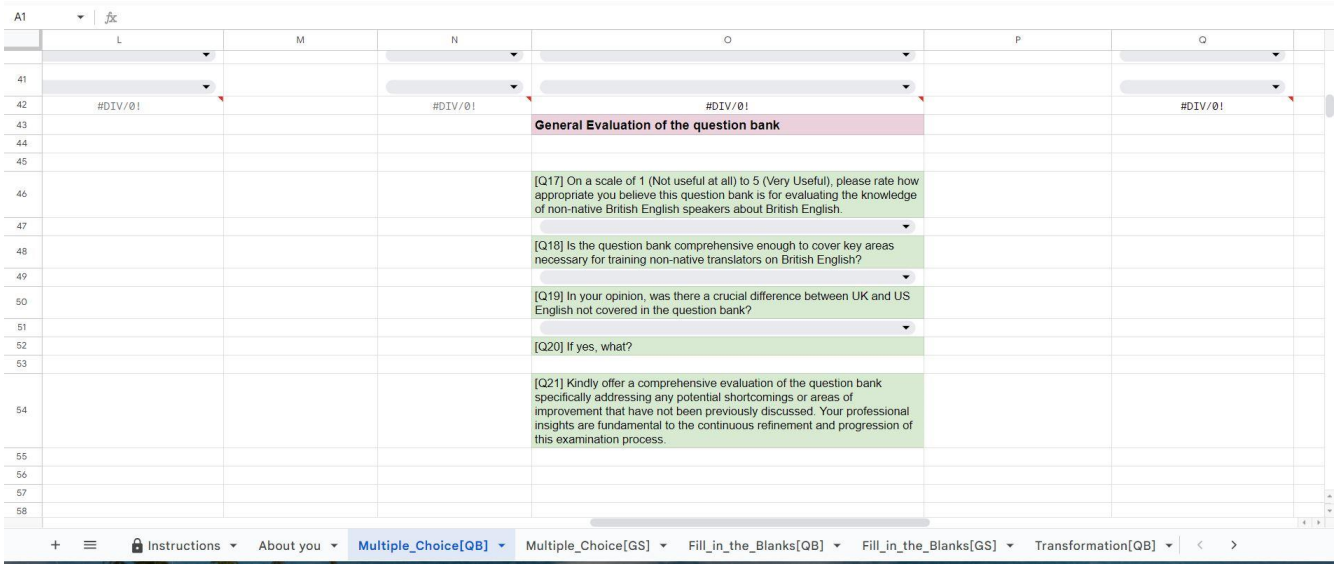
	J	K	L	M	N	O	P
1	Attention: Only answer the assessment questions on this row if Q3 says "NO"						
2	[Q4.1]Is the suggested right answer actually the most frequent?	[Q5] If your response to Q4.1 was "no", please specify what the right answer should be.	[Q6.1]Is there more than one frequently used answer?	[Q7]If your response to Q6.1 was "yes", please specify what are the other frequently used answers.	[Q8]Are the distractors close in meaning to the correct answer?	[Q9] Was the phrase/term relevant and up-to-date with modern usage in American English?	[Q10] If your response to Q9 was "no", please state a more current term or expression equivalent used in American English.
3							
4							
5							
6							

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Transformation[QB]



	P	Q	R	S	T	U	V
1							
2	rm relevant and usage in	[Q10] If your response to Q9 was "no", please state a more current term or expression equivalent used in American English.	[Q11] Was the phrase/term relevant and up-to-date with modern usage in British English?	[Q12] If your response to Q11 was "no", please state a more current term or expression equivalent used in British English.	[Q13] Is the correct answer mentioned or indicated in the stem itself?	[Q14] Kindly share any thoughts, feedback or suggestions you have regarding this question and its answer options. Any commentary or insights you provide are highly valuable and appreciated.	
3							
4							
5							
6							

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Transformation[QB]



	L	M	N	O	P	Q
41						
42	#DIV/0!		#DIV/0!	#DIV/0!		#DIV/0!
43				**General Evaluation of the question bank**		
44						
45						
46				[Q17] On a scale of 1 (Not useful at all) to 5 (Very Useful), please rate how appropriate you believe this question bank is for evaluating the knowledge of non-native British English speakers about British English.		
47						
48				[Q18] Is the question bank comprehensive enough to cover key areas necessary for training non-native translators on British English?		
49						
50				[Q19] In your opinion, was there a crucial difference between UK and US English not covered in the question bank?		
51						
52				[Q20] If yes, what?		
53						
54				[Q21] Kindly offer a comprehensive evaluation of the question bank specifically addressing any potential shortcomings or areas of improvement that have not been previously discussed. Your professional insights are fundamental to the continuous refinement and progression of this examination process.		
55						
56						
57						
58						

Instructions | About you | Multiple_Choice[QB] | Multiple_Choice[GS] | Fill_in_the_Blanks[QB] | Fill_in_the_Blanks[GS] | Transformation[QB]