Explorando a Inteligência Artificial no ensino de línguas: Criação e classificação automática de corpus de narrativas a partir de provérbios para o ensino de PLE com Large Language Models

Exploring Artificial Intelligence in language teaching: Automatic creation and classification of narrative corpora based on proverbs for teaching PFL with Large Language Models

Jorge Baptista[1,2], *Sónia Reis*[1,2]

[1] Universidade do Algarve – Faculdade de Ciências Humanas e Sociais
[2] INESC-ID Lisboa – Human Language Technology Lab

Abstract

Artificial Intelligence (AI), particularly Large Language Models (LLMs), has been transforming language education. This study explores the application of LLMs in the teaching of Portuguese as a Foreign Language (PFL), focusing on the automatic creation, classification, and validation of a *corpus* of short narratives based on Portuguese proverbs. The objectives are: (i) to automatically generate short narratives based on proverbs, suitable for different proficiency levels as defined by the CEFR; (ii) to automatically classify the linguistic proficiency level of these narratives using traditional machine learning techniques and foundational models (LLMs), followed by validation by human evaluators; (iii) to assess the adequacy of the narratives in relation to the original proverbs, justifying their potential didactic use in the context of PFL teaching. The methodology involves generating narratives with LLMs, which are then validated using automatic tools and human experts, as well as analysing the correspondence between the proverb and the generated narrative with a view to pedagogical application.

Keywords: automatic generation of narrative *corpora*, Portuguese as a Foreign Language (PFL), Artificial Intelligence (AI), Large Language Models (LLMs), Automatic classification of proverbs.

Resumo

A Inteligência Artificial (IA), em particular os Modelos de Linguagem de Grande Escala (LLMs), tem vindo a transformar o ensino de línguas. Este estudo explora a aplicação de LLMs no ensino de Português como Língua Estrangeira (PLE), centrando-se na criação, classificação e validação automática de um *corpus* de narrativas curtas baseadas em provérbios portugueses. Os objetivos centram-se em: (i) gerar automaticamente narrativas curtas baseadas em provérbios, adequadas a diferentes níveis de proficiência definidos pelo QECR; (ii) classificar automaticamente o nível de proficiência linguística dessas narrativas, utilizando técnicas de *machine learning* tradicionais e modelos fundacionais (LLMs), com posterior validação por avaliadores humanos; (iii) avaliar a adequação das narrativas aos provérbios originais, justificando o seu potencial uso didático no contexto do ensino de PLE. A metodologia envolve a geração de narrativas com LLMs, validadas por ferramentas automáticas e por especialistas humanos, bem como a análise da correspondência entre provérbio e narrativa gerada, tendo em vista a sua aplicação pedagógica.

Palavras-chave: geração automática de *corpus* de narrativas, Português como Língua Estrangeira (PLE), inteligência artificial (IA), modelos de linguagem de grande escala (*Large Language Models*), classificação automática de provérbios.

1. Introdução

Nos últimos anos, a inteligência artificial, em particular os modelos de linguagem de grande escala (LLMs), tem vindo a transformar profundamente diversas áreas do conhecimento – e a área da educação não é exceção. Estes modelos, treinados com grandes quantidades de dados linguísticos, são capazes de responder adequadamente a instruções complexas dadas em linguagem natural, gerar e traduzir texto com um grau de complexidade que se aproxima, em muitos casos, do texto produzido por um ser humano, o que os torna ferramentas versáteis para uso pedagógico.

Por *complexidade*, entende-se aqui os aspetos linguísticos do texto cuja variação está ligada à sua maior ou menor legibilidade/interpretabilidade, tais como o comprimento e a estrutura do texto, o vocabulário, a estrutura gramatical, referências culturais e processos estilísticos e, ainda, aspetos relacionados com a coesão e coerência (Monteiro et al., 2023). Como aproximação ao conceito de complexidade, adotamos a perspetiva de Ribeiro et al. (2024), segundo a qual, embora os níveis do *Quadro Europeu Comum de Referência para as Línguas* (QECR) (Conselho da Europa, 2001; Council of Europe, 2018) constituam uma ferramenta amplamente reconhecida para classificar a proficiência em línguas estrangeiras, podem também ser utilizados para identificar o nível de proficiência linguística necessário à compreensão de um determinado texto (tradução nossa) e, nesse sentido, servirem de indicadores do nível de complexidade do texto.

No contexto educativo, os LLMs têm sido progressivamente integrados em ambientes de aprendizagem, quer como apoio à docência, na criação de materiais e atividades, quer como ferramentas ao dispor dos alunos para treino autónomo e feedback imediato (Bonner et al., 2023, Dong et al., 2024, Kasneci et al., 2023). Em particular no ensino de línguas, têm demonstrado potencial para adaptar conteúdos a diferentes níveis de proficiência, sugerir vocabulário contextualizado, simular diálogos realistas e gerar textos idênticos aos produzidos por humanos com objetivos pedagógicos diversos (Benedetto et al., 2025).

Este estudo centra-se na aplicação destas tecnologias ao ensino de Português como Língua Estrangeira (PLE), com especial destaque para a geração automática de narrativas baseadas em provérbios. Os provérbios, enquanto expressões condensadas de sabedoria popular, são recursos linguísticos e culturais particularmente ricos (Reis, 2020). Para além do seu valor lexical e sintático, transportam referências culturais que contribuem para o desenvolvimento da competência pragmática dos alunos (Reis & Baptista, 2017, Reis et al., 2021). Neste sentido, a exploração de provérbios em contexto de criação de texto narrativo, com o apoio da inteligência artificial, revela-se uma estratégia pedagógica inovadora, com potencial para promover não só a aprendizagem da língua, mas também a compreensão intercultural.

Pretende-se, assim, explorar a utilidade dos modelos de linguagem na criação de materiais didáticos para o ensino de Português como Língua Estrangeira. Concretamente, parte-se dos provérbios para gerar automaticamente, com recurso a LLMs, narrativas adaptadas a diferentes níveis de proficiência de PLE, seguindo as orientações do QECR. Para tal, foram definidos três objetivos principais.

O primeiro consiste na constituição de um *corpus* de narrativas curtas, geradas automaticamente com base em provérbios da tradição portuguesa. Estas narrativas serão elaboradas com instruções específicas, fornecidas ao modelo de linguagem, de modo a assegurar a adequação destas a diferentes níveis de proficiência linguística, desde o A1 ao C1. A diversidade temática e a riqueza cultural inerente aos provérbios visam estimular o envolvimento dos alunos e facilitar a aprendizagem de estruturas e vocabulário em contexto.

O segundo objetivo prende-se com o desenvolvimento de um método de classificação automática do nível de proficiência das narrativas assim geradas. Para tal, serão testadas duas abordagens complementares: por um lado, o uso de técnicas tradicionais de aprendizagem automática (ing. *machine learning*) com base em traços (ing. *features*) linguísticos; por outro, a utilização de LLMs enquanto classificadores. A precisão desta

classificação será posteriormente avaliada com recurso a um conjunto de avaliadores humanos, especialistas em ensino de PLE, permitindo aferir a validade das soluções automatizadas propostas.

Por fim, o terceiro objetivo centra-se na avaliação da adequação pedagógica do conteúdo das narrativas face aos provérbios que lhes deram origem. Esta análise terá em conta a coerência temática, a clareza da mensagem e o potencial didático dos textos produzidos, nomeadamente no que diz respeito à sua integração em práticas de ensino-aprendizagem do PLE. Pretende-se, com isto, assegurar que as narrativas não só mantêm o valor cultural e expressivo do provérbio, como também se revelam úteis para serem aplicadas em contextos reais de sala de aula.

A investigação desenvolveu-se, pois, em torno de três questões principais:

i. Até que ponto as narrativas geradas automaticamente são adequadas ao significado e contexto de uso dos provérbios?

ii. Estão essas narrativas alinhadas com os descritores dos níveis de proficiência definidos pelo Quadro Europeu Comum de Referência para as Línguas (QECR)?

iii. Qual é a precisão da classificação automática dos níveis de proficiência, em comparação com a avaliação realizada por especialistas humanos?

Nas secções seguintes, apresentam-se a metodologia adotada, os procedimentos de análise e os principais resultados obtidos, seguidos da respetiva discussão.

2. Metodologia

A metodologia deste estudo desenvolve-se em três etapas principais, correspondentes às fases de geração, classificação e validação de narrativas, geradas automaticamente, com base em provérbios portugueses.

Na primeira etapa, recorremos ao modelo ChatGPT-4 (OpenAI, 2023) para gerar narrativas curtas inspiradas nos provérbios do *Mínimo Paremiólogico do Português Europeu* (MP; Reis & Baptista, 2020). Esta lista de 318 provérbios resulta da análise de um conjunto de mais de 114 mil provérbios recolhidos a partir de dicionários de referência. A seleção dos provérbios baseou-se na sua disponibilidade lexical, estimada com base na frequência de ocorrência em diferentes *corpora*, incluindo manuais escolares e textos jornalísticos. Adicionalmente, dois especialistas contribuíram para a seleção dos provérbios mais representativos, validando as escolhas com recurso a motores de pesquisa e a dois questionários aplicados a um número considerável de falantes. Com base nesta lista final de provérbios muito usuais e bem conhecidos da generalidade dos falantes, o ChatGPT-4 foi instruído para gerar múltiplas narrativas para cada provérbio, adaptadas a diferentes níveis de proficiência linguística, de acordo com os descritores do QECR (Conselho da Europa, 2001; Council of Europe, 2018).

Na segunda etapa, procedemos à classificação automática do nível de proficiência das narrativas geradas. Para isso, foram utilizados diferentes instrumentos: o próprio ChatGPT-4, as ferramentas de classificação Clavis[1] (Curto, 2014; Curto et al., 2015), e Lx-Proficiency[2] (Branco et al., 2014; Santos et al., 2021), ambas baseadas na análise da complexidade de aspetos linguísticos; e um modelo de rede neuronal, treinado especificamente para o português europeu (Ribeiro et al., 2024).

A terceira e última etapa consistiu na validação manual das classificações obtidas, realizada por especialistas na área do ensino de PLE. Esta validação teve como objetivo verificar a adequação linguística, pedagógica e pragmática das narrativas aos níveis atribuídos, bem como garantir a correspondência entre o conteúdo da narrativa e o significado e valor pragmático do provérbio que a originou.

[1] https://string.hlt.inesc-id.pt/demo/classification.pl
[2] https://portulanclarin.net/workbench/lx-proficiency

A seguir, descrevemos cada uma destas etapas e os critérios subjacentes à sua implementação.

2.1. **Geração de narrativas**

A geração de narrativas constituiu a primeira etapa do estudo, envolvendo a definição de instruções específicas para orientar o modelo de linguagem. As instruções aos modelos (ing. *prompts*) foram cuidadosamente construídas de forma a tentar garantir que as narrativas assim geradas fossem o mais adequadas possível aos diferentes níveis de proficiência linguística do QECR, do A1 ao C1 (ver Anexo).

Para cada provérbio, o modelo ChatGPT-4 recebeu instruções explícitas: deveria produzir uma narrativa curta, com um máximo de 100 palavras, em português europeu, evitando o uso direto das palavras-chave associadas ao provérbio. Esta última restrição visava incentivar uma interpretação mais profunda e contextualizada do conteúdo. Para cada provérbio, fornecia-se, então, essa lista de palavras-chave. Quanto ao nível de proficiência, não foi fornecida qualquer instrução adicional ao modelo, uma vez que o objetivo era precisamente testar a capacidade deste para gerar, de forma autónoma, narrativas ajustadas aos diferentes níveis do QECR. Os provérbios foram apresentados ao modelo em blocos de 15, de modo a evitar *prompts* excessivamente longos, o que facilitou o processo de geração das narrativas. A mesma instrução de base foi submetida cinco vezes, uma para cada nível de proficiência (A1, A2, B1, B2 e C1).

Para gerar a lista de palavras-chave associadas a cada provérbio, recorremos ao mesmo modelo de linguagem que foi utilizado na criação das narrativas. As sugestões automáticas foram posteriormente revistas e ajustadas manualmente. Em média, cada provérbio foi associado a três palavras-chave, embora este número possa variar consoante, sobretudo, a extensão da expressão. Por exemplo, o provérbio *Com o fogo não se brinca* foi associado a duas palavras-chave: *fogo* e *brincar*. As restantes palavras foram excluídas por se tratar de elementos gramaticais (ou funcionais) sem carga semântica relevante. Já o provérbio *A laranja de manhã é ouro, à tarde é prata e à noite mata* foi relacionado com sete palavras-chave: *laranja*, *manhã*, *ouro*, *tarde*, *prata*, *noite* e *matar*.[3] A Tabela 1 apresenta a distribuição dos provérbios segundo o número de palavras-chave associadas.

Tabela 1

Distribuição dos Provérbios por Número de Palavras-chave

Palavras-chave	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Provérbios	1	107	113	80	11	5	1	**318**

Verifica-se que a maioria apresenta entre 2 a 4 palavras-chave. Existem, no entanto, alguns casos fora deste padrão, em que o número de palavras-chave varia entre 1 e 7. No total, foram analisados os 318 provérbios do Mínimo Paremiológico.

Com base nestes provérbios e nas respetivas palavras-chave, foram geradas narrativas adaptadas a diferentes níveis de proficiência. Abaixo, apresentamos um exemplo de uma narrativa automaticamente gerada pelo modelo para o nível A1, a partir do provérbio *Em casa de ferreiro, espeto de pau*. As palavras-chave associadas ao provérbio – *casa*, *ferreiro*, *espeto* e *pau* – foram excluídas, para não se revelar explicitamente o provérbio subjacente.

> O Ricardo é mecânico e cuida muito bem dos carros dos seus clientes. No entanto, o seu próprio carro tem alguns problemas que ele nunca tem tempo de resolver. Os amigos acham engraçado, mas sabem que é porque ele está sempre ocupado a ajudar os outros.

[3] Ignorou-se, assim, o valor adverbial das expressões multipalavra *de manhã*, *à tarde* e *à noite*, retendo apenas os elementos lexicais principais (os nomes).

Tendo por base os 318 provérbios e os cinco níveis de proficiência definidos, foram geradas automaticamente 1 590 narrativas.[4] A Tabela 2 apresenta algumas estatísticas desses textos, nomeadamente, o comprimento (médio, mínimo e máximo) em número de caracteres e o respetivo desvio-padrão, bem como o total de palavras e a média de palavras por texto e por nível.

Tabela 2

Estatísticas do Corpus

Estatísticas	A1	A2	B1	B2	C1
Comp. médio (caracteres)	452	537	574	615	637
Comp. máximo (car.)	604	630	755	832	808
Comp. mínimo (car.)	223	406	465	514	566
Desvio padrão (car.)	67	37	35	38	36
N.º total palavras	**26 184**	**29 687**	**30 801**	**31 495**	**31 581**
N.º médio palavras	**82**	**93**	**97**	**99**	**99**

Verifica-se uma tendência de aumento gradual tanto no comprimento médio das narrativas (em número médio de caracteres), como na contagem média de palavras, à medida que se progride nos níveis de proficiência. Esta tendência confirma o aumento expectável da complexidade textual à medida que se avança nos níveis de proficiência linguística. Adicionalmente, o desvio-padrão do comprimento em número de caracteres mantém-se relativamente estável entre os níveis, exceto para os textos do nível A1, o que parece indicar alguma consistência na extensão dos textos gerados para cada nível. A diferença no comprimento dos textos do nível A1 prende-se com a geração neste caso de vários textos muito curtos, como se vê no comprimento mínimo em caracteres observado para este nível.

2.2. **Classificação automática com base no nível de proficiência**

Para proceder à classificação automática das narrativas geradas, de acordo com o nível de proficiência linguística, tomou-se como referência (ing. *ground truth*) o nível de proficiência (A1-C1) utilizado para a geração dos textos, ainda que essa classificação pudesse não estar inteiramente correta (v. §3.4.).

Foram adotadas quatro abordagens distintas. A primeira consistiu na utilização do ChatGPT-4 numa abordagem *zero-shot*, ou seja, sem qualquer treino específico para a tarefa de classificação. A segunda abordagem recorreu ao *Clavis* (Curto, 2014; Curto et al., 2015), um classificador baseado em algoritmos clássicos de *machine learning*. A terceira envolveu o uso de uma rede neuronal, assente num modelo fundacional, desenvolvido especificamente para a classificação do nível de complexidade de textos em português europeu (Ribeiro et al., 2024). Por fim, para um subconjunto das narrativas, foi também utilizada a ferramenta *Lx-Proficiency* (Branco et al., 2014; Santos et al., 2021), uma plataforma concebida para a classificação automática de textos com base no nível de proficiência linguística.

Apresentam-se, a seguir, estas quatro abordagens com mais pormenor, de forma a clarificar os processos subjacentes a cada uma.

[4] O *corpus* de textos gerados automaticamente (1590_Narrativas_ChatGPT_Proverbios_QECR) está disponível em https://doi.org/10.13140/RG.2.2.22435.16166

2.2.1. Classificação com o ChatGPT

Para esta etapa, recorremos ao modelo ChatGPT, acedido através de uma API (Interface de Programação de Aplicações). Esta interface permite a comunicação direta com o modelo, enviando instruções e recebendo respostas. A principal vantagem da utilização da API reside na sua maior flexibilidade e controlo, permitindo personalizar os pedidos, ajustar diversos parâmetros e realizar múltiplos pedidos em simultâneo – o que não seria viável na interface convencional. Além disso, a API mostra-se particularmente eficaz no tratamento de grandes volumes de dados, como exigido no nosso estudo, que implicava a classificação automática de um número elevado de textos.

Adotámos uma abordagem *zero-shot*, ou seja, o modelo não foi previamente afinado (ing. *fine-tuned*) para a tarefa específica de classificação de textos segundo os níveis do QECR, nomeadamente, não foram fornecidos exemplos de textos anotados quanto ao nível. Solicitámos apenas que atribuísse um nível de proficiência entre A1 e C1 a cada texto, sendo cada instância processada individualmente, sem acesso à informação das classificações anteriores nem a qualquer outro tipo de contextualização adicional. A Tabela 3 apresenta a matriz de confusão dos resultados obtidos com a classificação automática realizada através do ChatGPT. Nas colunas estão os níveis dos textos produzidos pelo modelo na etapa anterior, e nas linhas o resultado da classificação. Assim, por exemplo, para os 318 textos do nível A1, o modelo considerou apenas 2 como sendo desse nível, 157 do nível A2, 159 do nível B1, e nenhum dos níveis B2 e C1.

Tabela 3

Matriz de Confusão dos Resultados daCclassificação com o ChatGPT

	A1	**A2**	**B1**	**B2**	**C1**	**#**
A1	**2**	0	0	0	0	2
A2	157	**34**	3	3	0	197
B1	159	278	**308**	308	181	1 234
B2	0	6	7	**7**	136	156
C1	0	0	0	0	**1**	1
#	318	318	318	318	318	1 590

Tabela 4

Resultados da Classificação do ChatGPT

	A1	**A2**	**B1**	**B2**	**C1**	**Total**
Precisão	0,01	0,11	0,97	0,02	0,00	**0,22**
Precisão adjacente	0,50	0,98	1,00	0,99	0,43	**0,78**

Traduzindo estes resultados em termos de *precisão* (número de classificações corretas pelo total de classificações produzidas pelo modelo) e de *precisão adjacente* (considerando também os casos classificados um nível acima ou abaixo do nível de referência), verifica-se (Tabela 4) que (i) a classificação automática com

o ChatGPT revelou uma precisão global bastante baixa (0,22),[5] com exceção dos textos do nível B1, ou seja o modelo tem uma tendência para classificar a maioria dos textos como B1, o nível médio/central. Como é natural, a precisão adjacente é mais elevada (0,78), o que significa que o desvio entre o nível de referência e o nível atribuído não é muito grande. Assim, podemos concluir que ChatGPT ainda apresenta muitas limitações na tarefa de classificação, não sendo capaz de estabelecer uma distinção clara entre os diferentes níveis do QECR.

2.2.2. Classificação com o Clavis

Nesta secção, analisamos a utilização do classificador Clavis (Curto, 2014; Curto et al., 2015), uma ferramenta treinada com um *corpus* de 237 textos retirados dos exames do Instituto Camões e previamente classificados quanto ao nível de proficiência. O Clavis extrai 52 características linguísticas de cada texto e aplica um algoritmo de aprendizagem automática para proceder à respetiva classificação, em 3 ou em 5 níveis.

Abaixo, apresenta-se a matriz de confusão relativa à classificação automática realizada com a ferramenta Clavis. A diagonal principal da matriz representa os casos corretamente classificados, enquanto os restantes valores correspondem a classificações incorretas. Por exemplo, observa-se que, no nível B1, o modelo conseguiu classificar corretamente 294 textos. Na Tabela 6, encontra-se um resumo com os principais indicadores de desempenho: o modelo acertou 340 casos (verdadeiros positivos) e cometeu 1.250 erros de classificação (falsos positivos). A precisão global foi de 21,4%, e a precisão adjacente – que, como se disse, considera aceitáveis as classificações atribuídas a níveis imediatamente próximos do correto – foi de 46,4%. Este último valor sugere que, embora o desempenho global seja limitado, o modelo revelou alguma capacidade de aproximação ao nível correto em muitos casos.

Tabela 5

Matriz de Confusão da Classificação com o Clavis (Curto et al., 2015)

Clavis	A1	A2	B1	B2	C1	#
A1	**46**	37	24	213	224	544
A2	0	**0**	0	24	55	79
B1	272	281	**294**	79	37	963
B2	0	0	0	**2**	2	4
C1	0	0	0	0	**0**	0
#	318	318	318	318	318	1 590

[5] Na avaliação quantitativa do grau de concordância, usou-se apenas a *concordância bruta* (n.º de casos de acordo/total de casos). Já para a interpretação qualitativa, recorreu-se à escala de Landis e Koch (1977), desenvolvida para os valores de Cohen Kappa (Cohen, 1960), métrica que não foi usada nesta situação já que leva em conta a classificação que seria obtida por mero acaso.

Tabela 6

Resultados da Classificação do Clavis

	A1	A2	B1	B2	C1	Total
Precisão	0,14	0,00	0,92	0,01	0,00	**0,22**
Precisão adjacente	0,14	1,00	0,92	0,25	0,01	**0,47**

Os resultados apresentados na Tabela 6 revelam um desempenho semelhante ao observado na abordagem anterior, com valores baixos de precisão global (22%) e valores moderados de precisão adjacente (47%), talvez porque o *corpus* de treino fosse pouco adequado à tarefa de classificação. Efetivamente este *corpus* não se encontra balanceado (ing. *balanced*; McEnery et al., 2006, pp. 16–19) por níveis, o que afeta a robustez do modelo. Além disso, os dados de anotação manual incluídos no treino evidenciaram baixa concordância entre anotadores, o que fragiliza ainda mais a fiabilidade das classificações produzidas.

2.2.3. Classificação com modelo de rede neuronal

Neste ponto, comparamos os resultados obtidos com o Clavis com os de um novo classificador desenvolvido por Ribeiro et al. (2024). Este classificador foi treinado sobre uma versão ampliada do mesmo *corpus* Camões utilizado pelo Clavis, agora com 598 textos – mais do que o dobro do conjunto de dados original. Além da ampliação do *corpus*, esta abordagem integra uma tecnologia mais recente, baseada numa rede neuronal profunda, e sustentada pelo modelo fundacional *Albertina PT-PT*, desenvolvido por Rodrigues et al. (2023). Trata-se de um modelo especificamente concebido para o processamento do português europeu, o que o torna particularmente adequado para tarefas de classificação linguística em contextos educativos. O modelo foi afinado (ing. *fine-tuned*) com a versão alargada do *corpus* acima referida.

Como se pode observar na Tabela 7, o número de acertos obtido por este classificador é superior ao verificado com o Clavis, totalizando 385 verdadeiros positivos. No entanto, as principais dificuldades do modelo concentram-se nos níveis mais avançados: nenhum texto foi corretamente classificado como B2 ou C1, o que evidencia limitações na diferenciação das produções mais complexas.

Tabela 7

Matriz de Confusão da Classificação com a Rede Neuronal (Ribeiro et al., 2024) & o Modelo Fundacional Albertina (Rodrigues et al., 2023)

RN	A1	A2	B1	B2	C1	#
A1	102	46	37	16	5	206
A2	57	64	62	75	79	337
B1	159	208	219	227	234	1 047
B2	0	0	0	0	0	0
C1	0	0	0	0	0	0
#	318	318	318	318	318	1 590

Tabela 8

Resultados da Classificação da Rede Neuronal

	A1	A2	B1	B2	C1	Total
Precisão	0,32	0,20	0,69	0,00	0,00	**0,24**
Precisão adjacente	0,50	1,00	0,88	0,71	0,00	**0,62**

A precisão global do modelo foi de 24,2%, enquanto a precisão adjacente atingiu 61,9%, revelando uma capacidade considerável de aproximação ao nível correto, mesmo quando a classificação exata não foi atingida.

Comparação de resultados: De um modo geral, em termos de precisão, os resultados dos dois classificadores utilizados são bastante fracos (aprendizagem automática clássica & Clavis: 0,22; rede neuronal + Albertina-PT: 0,24), ainda que o segundo tenha utilizado um *corpus* de treino com praticamente o dobro do tamanho. Mais expressiva foi a diferença observada na precisão adjacente. Neste caso, o Clavis registou 47%, ao passo que o novo classificador alcançou 62% – um ganho significativo, que demonstra maior sensibilidade desta tecnologia às subtilezas linguísticas entre níveis contíguos. Ainda assim, ambos os modelos evidenciam fragilidades na identificação de produções de níveis mais avançados: o Clavis classificou corretamente apenas quatro textos nos níveis B2 e C1 (dois em cada), e o novo classificador não atribuiu corretamente nenhum texto a esses níveis. Estes resultados sugerem que a distinção entre os níveis superiores continua a representar um desafio para estes sistemas de classificação automática.

2.2.4. Classificação automática com o Lx-Proficiency

Nesta secção, e a título meramente ilustrativo, exploramos o desempenho da Lx-Proficiency na tarefa de atribuição automática de níveis de proficiência a um subconjunto de textos do nosso *corpus*.

Este sistema foi inicialmente desenvolvido (Branco et al., 2014) de um modo semelhante ao do sistema Clavis, como um sistema de aprendizagem supervisionada com base em quatro traços (ing. *features*) selecionados (índice de Flesch, proporção de substantivos, comprimento médio de palavras em número de sílabas e comprimento médio de frases em número de palavras). O serviço é disponibilizado numa aplicação *on-line*.[6] Posteriormente, Santos et al. (2021) desenvolveram um modelo do ChatGPT2 afinado (ing. *fine-tuned*) com uma versão maior do *corpus* do Instituto Camões, com 500 textos, e que passou a ser usado como base daquela aplicação.

Dadas as limitações na utilização da ferramenta Lx-Proficiency, cuja interface não permite processar em bloco o conjunto dos 1 590 textos gerados, esta foi aplicada apenas aos 75 textos de 15 provérbios, selecionados aleatoriamente (esta seleção de provérbios será apresentada mais adiante). Veja-se, na Tabela 9, a distribuição das classificações atribuídas.

[6] https://portulanclarin.net/workbench/lx-proficiency

Tabela 9

Matriz de Confusão da Classificação com o Lx-Proficiency (Branco et al., 2014; Santos et al., 2021)

Lx-Proficiency	A1	A2	B1	B2	C1	#
A1	**8**	0	7	0	0	15
A2	5	**0**	10	0	0	15
B1	4	0	**11**	0	0	15
B2	2	0	13	**0**	0	15
C1	2	0	12	0	**1**	15
#	21	0	53	0	1	75

A precisão global da Lx-Proficiency foi de 27%, ligeiramente superior à registada pelos classificadores previamente analisados. A precisão adjacente, por sua vez, situou-se nos 64%, o que revela uma performance moderada na identificação dos níveis de proficiência, com alguma capacidade de aproximação ao nível correto.

Tabela 10

Resultados daCclassificação do Lx-Proficiency

	A1	A2	B1	B2	C1	Total
Precisão	0,38	-	0,21	-	1,00	**0,27**
Precisão adjacente	0,62	-	0,64	-	1,00	**0,64**

Em síntese, e de um modo geral, a precisão global foi baixa em todos os classificadores aqui testados, situando-se entre os 22% do Clavis e os 27% do Lx-Proficiency, o que demonstra as dificuldades persistentes na classificação precisa de textos por nível segundo o QECR. No entanto, os valores de precisão adjacente, especialmente no caso do ChatGPT (78%) e do classificador neuronal (62%), revelam já uma capacidade de acerto, o que pode ser particularmente útil em contextos pedagógicos onde a gradação por níveis contíguos seja razoavelmente aceitável. Outro dado comum às diferentes abordagens diz respeito à dificuldade na distinção dos níveis mais avançados, nomeadamente B2 e C1. Nenhum dos modelos foi capaz de classificar corretamente um número significativo de textos nesses níveis, o que sugere que a complexidade lexical e estrutural destas produções ainda não é plenamente capturada pelos sistemas automáticos, mesmo quando suportados por modelos de língua relativamente recentes e especificamente desenvolvidos para o português europeu.

Apesar destas limitações, os resultados obtidos oferecem pistas para futuras melhorias nos processos de classificação automática de textos. A utilização de *corpora* mais extensos, adequadamente balanceados, e a diversificação/complexificação das instruções aos modelos (ing. *prompts*) poderão contribuir para o desenvolvimento de ferramentas mais eficazes e sensíveis às nuances dos diferentes níveis de proficiência linguística em PLE.

3. Classificação manual

Avançamos agora para a fase de classificação manual, que visava validar e complementar os resultados das abordagens automáticas previamente descritas. A classificação manual foi realizada com base nos mesmos 15 provérbios utilizados na fase anterior, selecionados aleatoriamente do Mínimo Paremiológico e totalizando

75 textos. Os textos foram avaliados e classificados manualmente por dois anotadores, com conhecimento aprofundado do QECR e do ensino de PLE. Os 15 provérbios utilizados nesta fase do estudo são os seguintes:

Tabela 11

Lista dos 15 Provérbios da Seleção Aleatória

ID	Provérbios
ID_102	*Em boca fechada não entra mosca.*
ID_108	*Enquanto o pau vai e vem folgam as costas.*
ID_111	*Errar é humano.*
ID_112	*Faz o que eu digo e não faças o que eu faço.*
ID_135	*Manda quem pode.*
ID_176	*O prometido é devido.*
ID_199	*Os fins justificam os meios.*
ID_212	*Para bom entendedor, meia palavra basta.*
ID_215	*Para quem é, bacalhau basta.*
ID_240	*Quem corre por gosto não cansa.*
ID_249	*Quem está mal que se mude.*
ID_291	*Recordar é viver.*
ID_305	*Tristezas não pagam dívidas.*
ID_306	*Tudo está bem quando acaba bem.*
ID_314	*Vão-se os anéis, ficam os dedos.*

Para a avaliação da complexidade linguística das narrativas geradas automaticamente, foram consideradas cinco categorias principais, adaptadas da definição dos níveis de complexidade do projeto iRead4Skills[7] (Monteiro et al., 2023) e articuladas com as orientações do QECR. As categorias em análise foram:

i. Comprimento e estrutura do texto – Varia de textos curtos, até 50 palavras, com estruturas simples e lineares (nível 1), até textos longos e complexos, com estruturas intricadas e variação significativa de tópicos, exigindo leitura atenta e conhecimento prévio (nível 5).

ii. Vocabulário – Vai de palavras simples e frequentes, relacionadas com o quotidiano e objetos concretos, sem uso de siglas ou estrangeirismos (nível 1), até vocabulário mais complexo, com palavras raras, jargões específicos e uso intensivo de estrangeirismos e siglas não explicadas (nível 5).

iii. Complexidade gramatical – Vai de frases curtas e simples, com estruturas de coordenação básicas, uso predominante da voz ativa e tempos verbais limitados ao presente e aos pretéritos simples (nível 1), a frases muito complexas, com coordenação e subordinação, construções gramaticais mais complexas como a voz passiva impessoal e o uso de tempos verbais compostos, incluindo o pretérito mais-que-perfeito (nível 5).

[7] https://zenodo.org/records/10459090

iv. Referências culturais e estilísticas – Vai de textos sem qualquer figura de estilo ou referência cultural, caracterizados por uma linguagem denotativa, clara e direta (nível 1), a textos densamente marcados por figuras de estilo mais complexas (como metáforas ou analogias sofisticadas) e referências culturais ou históricas específicas, cuja interpretação requer conhecimento prévio e elevado grau de inferência (nível 5).

v. Coesão e coerência – Vai de textos com cadeia referencial completa, sem omissões, e com repetição de termos para garantir clareza, organizados segundo uma sequência temporal linear (nível 1), a textos com coesão e coerência altamente complexas, marcados por uso extensivo de omissões, pronomes e sequências temporais intricadas, que requerem uma atenção acrescida por parte do leitor (nível 5).

Cada categoria foi avaliada numa escala de Likert de 1 a 5, em que 1 corresponde a "muito fácil" e 5 a "muito difícil". Esta grelha permitiu uma análise sistemática da complexidade de cada narrativa, contribuindo para a aferição da sua adequação ao nível de proficiência-alvo. A atribuição do nível global ao texto[8] resultou da soma das pontuações atribuídas a cada uma das cinco categorias anteriormente descritas, segundo o critério seguinte: A1 ={5,8}, A2={9,12} B1={13,16} B2={17,20} e C1={21,25}.

A seguir, apresenta-se a matriz de confusão relativa à concordância entre os dois avaliadores na classificação manual das narrativas segundo os níveis de proficiência. Esta matriz permite observar o alinhamento (ou divergência) entre os juízos atribuídos por cada anotador.

Tabela 12

Acordo entre Dois Anotadores (nas Colunas o Anotador 1 e nas Linhas o Anotador 2)

	A1	**A2**	**B1**	**B2**	**C1**
A1	**4**	6	3	2	
A2	0	**11**	4		
B1	1	3	**11**	0	
B2		2	6	**7**	0
C1		2	4	9	**0**

A taxa de concordância exata foi relativamente baixa, situando-se em torno dos 44% (ver Tabela 13), o que reflete a complexidade da tarefa e a subjetividade envolvida na avaliação de produções linguísticas com base em múltiplos critérios.

[8] Adotou-se, assim, uma abordagem analítica e indutiva, partindo do princípio de que a identificação de cada uma das categorias individuais é, por si só, uma tarefa complexa, mas que oferece maior sustentação à classificação final. Considerou-se, ainda, que a atribuição direta de um grau de dificuldade global seria mais suscetível a enviesamentos.

Tabela 13

Taxa de Acordo entre Anotadores

	A1	A2	B1	B2	C1	Total
Precisão	0,19	0,00	0,21	0,00	0,00	**0,44**
Precisão adjacente	0,19	0,00	0,40	0,00	0,00	**0,81**

Apesar dessa dificuldade, a precisão adjacente foi significativamente mais elevada, atingindo os 81%. Este valor indica que, mesmo quando os avaliadores não coincidiram no nível exato a atribuir ao texto, a grande maioria das discrepâncias ocorreu entre níveis contíguos, o que evidencia um grau considerável de alinhamento e consistência interpretativa.

3.2 **Acordo entre anotador e referência**

Com o intuito de validar as avaliações manuais, esta secção apresenta os dados comparativos entre a classificação de cada anotador e o nível de referência definido para a geração de cada texto. Nas tabelas seguintes, apresentam-se os dados relativos à comparação entre cada anotador e a referência, com base nas classificações manuais atribuídas às 75 narrativas.

Tabela 14

Acordo Anotadores-Referência (à Esquerda Anotador1 e à Direita o Anotador 2)

Anot 1	A1	A2	B1	B2	C1	#	Anot 2	A1	A2	B1	B2	C1	#
A1	**9**	1		3	2	15	**A1**	**5**	10				15
A2	3	**10**	2			15	**A2**		**12**	2			14
B1	3	1	**10**	1		15	**B1**		3	**13**	9	3	28
B2		1	1	**8**	5	15	**B2**				**6**	12	18
C1		2	2	3	**8**	15	**C1**					**0**	0
#	15	15	15	15	15	75	#	5	25	15	15	15	75

Tabela 15

Matriz de Confusão das Anotações (à Esquerda, o Anotador1 e, à Direita, o Anotador 2)

Anot1	A1	A2	B1	B2	C1	Total	Anot2	A1	A2	B1	B2	C1	Total
Precisão	0,60	0,67	0,67	0,53	0,53	**0,60**	**Precisão**	1,00	0,48	0,87	0,40	0,00	**0,48**
PAdj.	0,80	0,80	0,87	0,80	0,87	**0,83**	**PAdj.**	1,00	1,00	1,00	1,00	0,80	**0,96**

À esquerda, observa-se a matriz de confusão do Anotador 1 em relação à referência, com uma taxa de concordância exata de 60% e uma taxa de concordância adjacente de 83%. Estes valores indicam um acordo moderado entre as classificações atribuídas pelo anotador e os níveis definidos como referência, mas elevado no que diz respeito à distinção entre níveis contíguos.

À direita, é apresentada a comparação entre o Anotador 2 e a referência, com uma taxa de precisão também moderada, mas inferior à do Anotador 1 (48%), e uma taxa de precisão adjacente bastante mais elevada, quase perfeita (96%). Apesar da menor coincidência no nível exato, os dados sugerem que as classificações do Anotador 2 se mantiveram muito próximas das da referência, com a maioria das discrepâncias situadas entre níveis vizinhos, o que evidencia uma boa sensibilidade ao grau de dificuldade relativo das narrativas.

3.3 **Classificação manual texto-provérbio**

A última tarefa realizada neste estudo consistiu em atribuir manualmente a cada texto 1 dos 15 provérbios do mesmo conjunto anteriormente referido. Os anotadores foram os mesmos que realizaram a tarefa anterior. Trata-se de uma tarefa de estabelecimento de correspondência, pois os anotadores conheciam a lista dos provérbios a partir dos quais foram gerados os textos. O objetivo foi avaliar a adequação do texto ao provérbio. Utilizou-se uma escala de Likert de 0 a 2, de acordo com as diretrizes que se apresentam abaixo:

0: O texto não corresponde a nenhum provérbio da lista
1: O texto corresponde, mas de forma inadequada a um dos provérbios da lista, ou pode aplicar-se-lhe mais do que um provérbio
2: O texto corresponde perfeitamente a um dos provérbios da lista

Apresentamos a seguir os resultados da correspondência entre os textos e os provérbios, feita pelo Anotador 1. Nas colunas indica-se o n.º de identificação de referência do provérbio que serviu de base para a geração do texto (ver Tabela 16) e nas linhas o n.º de identificação atribuído pelo anotador. Para 3 textos, o Anotador 1 não conseguiu determinar um provérbio da lista a que pudesse fazer corresponder o texto gerado (coluna '0').

Cada célula da matriz mostra o número de provérbios atribuídos correta ou incorretamente a um determinado texto.

Tabela 16

Verificação da Correspondência entre o Texto e o Provérbio, Feita pelo Anotador 1

Anot1	0	102	108	111	112	135	176	199	212	215	240	249	291	305	306	314	
102		**5**															5
108	**1**		**3**												**1**		5
111				**5**													5
112					**5**												5
135	**1**					**3**									**1**		5
176							**5**										5
199							**1**	**2**							**2**		5
212									**5**								5
215										**0**					**5**		5
240	**1**										**4**						5
249												**5**					5
291													**5**				5
305														**5**			5
306															**5**		5
314																**5**	5
Total	3	5	3	5	5	3	6	2	5	0	4	5	5	5	14	5	

A diagonal mostra os casos de correspondência, onde as células a **verde** indicam que o anotador acertou todas as cinco vezes, atribuindo corretamente aos cinco textos o provérbio correspondente. Já as células a **amarelo** indicam que houve acertos, mas não em todos os textos. As células fora da diagonal revelam os erros de atribuição, indicados a **vermelho**. O total de casos corretamente assinalados nesta avaliação foi de 62, resultando numa precisão de 83%, o que pode ser considerado um acordo quase perfeito.

A Tabela 17 mostra os resultados da atribuição de correspondência entre os textos e os provérbios, realizada pelo Anotador 2.

Tabela 17

Verificação da Correspondência entre o Texto e o Provérbio, Feita pelo Anotador 2

Anot2	0	102	108	111	112	135	176	199	212	215	240	249	291	305	306	314	
102		**5**															5
108			**5**														5
111				**5**													5
112					**5**												5
135	**1**					**4**											5
176							**5**										5
199								**4**			**1**						5
212									**5**								5
215										**5**							5
240											**5**						5
249												**5**					5
291													**5**				5
305														**5**			5
306															**5**		5
314																**5**	5
Total	1	5	5	5	5	4	5	4	5	5	6	5	5	5	5	5	

Com um total de 73 verdadeiros-positivos, a precisão da classificação foi de 97%, o que demonstra uma precisão quase perfeita na atribuição dos provérbios aos textos, com apenas dois casos em que os provérbios foram incorretamente identificados.

A elevada qualidade da classificação por parte dos dois anotadores pode ser explicada pelo facto de se trabalhar com um conjunto limitado de 15 provérbios e previamente conhecidos pelos avaliadores, o que terá facilitado o processo de correspondência entre textos e provérbios.

Comparando as classificações de ambos os anotadores, como se pode observar na Tabela 18, foram registados 63 casos de concordância, o que se traduz numa taxa de acordo de 0,84 – valor que representa um nível bastante elevado de consistência entre os dois anotadores.

Tabela 18

Concordância entre os Dois Anotadores

Anot1/2	0	102	108	111	112	135	176	199	212	215	240	249	291	305	306	314
0	1		1								1					
102		5														
108			3													
111				5												
112					5											
135						3										
176							5	1								
199								2								
212									5							
215										0						
240											4					
249												5				
291													5			
305														5		
306			1			1		1		5	1				5	
314																5
Total	1	5	5	5	5	4	5	4	5	5	6	5	5	5	5	5

3.4 Narrativas vs. Referencial Camões: análise linguística

Uma outra tarefa desenvolvida consistiu em verificar em pormenor se os fenómenos linguísticos presentes nos textos gerados estavam adequados ao nível de proficiência atribuído. Como meros exemplo, apresentam-se a seguir alguns fenómenos gramaticais identificados nos textos dos níveis A1 e A2, que foram gerados com base no provérbio *Em boca fechada não entra mosca*.

> [ID_102, **A1**] O Tiago estava com os amigos a conversar sobre um assunto delicado. Ele tinha vontade de dar a sua opinião, mas percebeu que poderia causar confusão. Decidiu ouvir os outros em vez de falar, e assim evitou uma discussão desnecessária.
>
> [ID_102, **A2**] Durante uma reunião, o João ouviu uma opinião com a qual não concordava. Quis responder imediatamente, mas lembrou-se de que era melhor refletir antes de falar. Esperou até o fim da reunião para dar a sua opinião de forma ponderada e construtiva.

No texto de nível A1, observa-se o uso do pretérito perfeito do indicativo (*evitou*), tempo verbal associado ao nível A2 conforme o inventário de gramática[9] do Referencial Camões PLE (Direção de Serviços de Língua e Cultura, 2017). Identificam-se também estruturas como o complexo (ou perífrase) verbal "estar a + infinitivo",

[9] https://www.instituto-camoes.pt/activity/centro-virtual/referencial-camoes-ple

mas com o verbo *estar* no pretérito imperfeito; a preposição *sobre* com função de "expressão de noção situacional (sobre)"; e a locução conjuncional *em vez de,* a introduzir uma oração (*em vez de falar*) – todos elementos que, no Referencial, são associados ao nível A2.

No texto classificado como A2, surgem fenómenos de maior complexidade, como o uso de um verbo auxiliar modal no pretérito perfeito + infinitivo (*quis responder*); o pronome relativo *a qual*, que o Referencial apresenta associado ao nível B2; e a preposição *durante*, empregue para fazer a situação de um evento (a oração principal) no tempo de outro evento (*a reunião*) – um fenómeno explicitamente descrito no Referencial como próprio do nível B1.

Seria fastidioso estar aqui a alistar todos os fenómenos detetados em que se verificou uma não correspondência entre as estruturas observadas nos textos gerados artificialmente e as orientações do Referencial. Esse trabalho de anotação prossegue ainda, a fim de se poderem tirar conclusões mais sólidas sobre este aspeto.

4. Conclusões e trabalho futuro

Podemos concluir que, de um modo geral, o ChatGPT se revela capaz de produzir textos que são, na sua maioria, adequados aos diferentes níveis de proficiência do QECR para que foram solicitados, mesmo numa abordagem *zero-shot*. Contudo, identificámos algumas situações em que se observa a introdução de fenómenos linguísticos que, segundo o Referencial Camões, seria mais apropriado explorar em níveis superiores.

O modelo consegue igualmente produzir textos apropriados aos provérbios, mantendo, na maioria dos casos, uma relação coerente entre o conteúdo gerado e o sentido do provérbio.

Verificámos ainda que os humanos não são particularmente eficazes na classificação do nível de proficiência dos textos, mas que os modelos de linguagem também não o são, independentemente dos dados em que foram treinados ou da abordagem adotada – seja baseada em redes neuronais ou em métodos mais tradicionais de aprendizagem automática.

Do ponto de vista pedagógico, consideramos que estas ferramentas têm elevado potencial para poderem ser utilizadas, mas a sua aplicação requer validação humana, especialmente quando os textos são integrados em contextos de aprendizagem reais.

Para trabalho futuro, pretendemos envolver mais anotadores nas tarefas manuais aqui apresentadas, com o objetivo de (a) contribuir para uma avaliação mais robusta dos níveis de proficiência associados e (b) confirmar a adequação das narrativas aos provérbios que lhes deram origem, mas sem os revelar previamente a esses anotadores.

Agradecimentos

Os autores agradecem igualmente ao Professor Eugénio Ribeiro a utilização do seu modelo de redes neuronais (Ribeiro et al., 2024) para a classificação automática das narrativas inspiradas em provérbios analisadas neste estudo.

Financiamento

Este trabalho foi parcialmente financiado por fundos nacionais através da FCT (Referência: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) e pela Comissão Europeia (Projeto: iRead4Skills, Número da subvenção: 1010094837, Tópico:HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

Referências

Benedetto, L., Gaudeau, G., Caines, A., & Buttery, P. (2025). Assessing how accurately large language models encode and apply the Common European Framework of Reference for Languages. *Computers and Education: Artificial Intelligence*, *8*, 100353. https://doi.org/10.1016/j.caeai.2024.100353

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, *1*(23), 23–41.

Branco, A., Rodrigues, J., Silva, J., Costa, F., & Vaz, R. (2014). Assessing automatic text classification for interactive language learning. In *Proceedings International Conference on Information Society (i-Society 2014)* (pp. 70–78). IEEE. https://doi.org/10.1109/i-Society.2014.7009014

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Conselho da Europa. (2001). *Quadro Europeu Comum de Referência para as Línguas – Aprendizagem, ensino, avaliação*. Edições ASA.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment, companion volume with new descriptors*. Council of Europe. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Curto, P. (2014). *Classificador de textos para o ensino de português como segunda língua* [Dissertação de mestrado não publicada]. Instituto Superior Técnico-Universidade de Lisboa.

Curto, P., Mamede, N., & Baptista., J. (2015). Automatic text difficulty classifier. In *CSEDU: Proceedings of the International Conference on Computer Supported Educatio*n (pp. 36–44). SCITEPRESS. https://doi.org/10.5220/0005428300360044

Direção de Serviços de Língua e Cultura. (2017). *Referencial Camões PLE – Português Língua Estrangeira*. Camões, Instituto da Cooperação e da Língua, I.P. https://www.instituto-camoes.pt/images/REFERENCIAL_ebook.pdf

Dong, B., Bai, J., Xu, T., & Zhou, Y. (2024). Large language models in education: A systematic review. In *Proceedings 2024 6th International Conference on Computer Science and Technologies in Education (CSTE)* (pp. 131–134). IEEE. https://doi.org/10.1109/CSTE62025.2024.00005

Ghosh, S. & Srivastava, S. (2021). *EPiC: Employing proverbs in context as a benchmark for abstract language understanding*. ArXiv preprint.

Grosso, M., Soares, A., Sousa F. & Pascoal, J. (2011). *QuaREPE – Quadro de Referência para o Ensino Português no Estrangeiro – Documento Orientador*. Ministério da Educação.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllemeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, A., 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

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochem Med*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Monteiro, R., Amaro, R., Correia, S., Pintard, A., Gauchola, R., Moutinho, M., & Blanco Escoda, X. (2023). *iRead4Skills – Complexity Levels* (V1.0). Zenodo. https://doi.org/10.5281/zenodo.10459090

OpenAI. (2020). *GPT-3 via API* [Modelo de linguagem de grande escala]. https://platform.openai.com/

OpenAI. (2023). *ChatGPT* (versão GPT-4) [Modelo de linguagem de grande escala]. https://chat.openai.com/

Reis, S., & Baptista, J. (2017). Os provérbios em manuais de ensino de português língua não materna. In V. Pinheiro & G. H. Paetzold (Orgs.), *Anais do XI simpósio brasileiro de tecnologia da informação e da linguagem humana* (pp. 247–255). SBC.

Reis, S., & Baptista, J. (2020). Determinação de um mínimo paremiológico do português europeu. *Acta Scientiarum. Language and Culture*, *2*(42), e52114.

Reis, S. (2020). *Expressões proverbiais do português-usos, variação formal e identificação automática* [Tese de doutoramento não publicada]. Universidade do Algarve.

Reis, S., Baptista, J., & Mamede, N. (2021). Provérbios portugueses usuais: Distribuição em *corpora*. In E. Ruiz, T. Torrent, J. Souza, M. Madruga, R. Rodrigues, C. Barros & D. Claro (Orgs.), *Anais do XIII simpósio brasileiro de tecnologia da informação e da linguagem humana* (pp. 325–334). SBC.

Ribeiro, E., Mamede, N., & Baptista, J. (2024). Automatic Text Readability Assessment in European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese* (Vol. 1, pp. 97–107). Association for Computational Linguistics.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H., & Osório, T. (2023). Advancing neural encoding of Portuguese with transformer Albertina PT-*. In N. Moniz, Z. Vale, J. Cascalho, C. Silva & R. Sebastião (Eds.), *Progress in Artificial Intelligence. EPIA 2023* (part I, pp. 441–453). Springer. https://doi.org/10.1007/978-3-031-49008-8_35

Santos, R., Rodrigues, J., Branco, A., & Vaz, R. (2021). Neural text categorization with transformers for learning Portuguese as a second language. In G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso & L. Reis (Eds.), *Progress in Artificial Intelligence. EPIA 2021* (pp. 715–726). Springer. https://doi.org/10.1007/978-3-030-86230-5_56

Anexo

Este anexo apresenta um exemplo de *prompt* elaborado para orientar a geração automática de narrativas com recurso a um modelo de linguagem de grande escala (LLM).

Prompt:

Considera que és um especialista em Ensino de Português como Língua Estrangeira e que tens a tarefa de construir textos adequados ao nível de proficiência esperado para estudantes desse nível.

Vais receber abaixo uma linha de exemplo com um provérbio português, precedido do respetivo ID e uma lista de palavras-chave que *NÃO* devem ser usadas no texto a produzir. Aqui está um exemplo de um provérbio (input):

ID_075 Casa roubada, trancas à porta. casa,roubar,tranca,porta

O total de provérbios é 318. No entanto, vou apresentar-te somente 15 provérbios de cada vez.

Instrução:

Constrói uma pequena narrativa, em português europeu (respeitando o Novo Acordo Ortográfico), com um máximo de 100 palavras, que conte uma história adequada ao provérbio indicado e que o texto seja adequado a leitores com um nível de proficiência *A1* do Quadro Comum Europeu de Referência para as Línguas (QECR). Nesse texto, *NÃO* deves usar as palavras-chave indicadas.

Os textos do Nível A1 (Iniciante) podem ser descritos assim: Textos simples e curtos, frases simples, vocabulário básico, temas quotidianos (saudações, informações pessoais, compras).

Aqui estão os provérbios a considerar e a lista de palavras-chave que *NÃO* devem ser usadas no texto a produzir:

ID_001 A beleza está nos olhos de quem a vê. beleza,olho,ver
ID_002 A carne é fraca. carne,fraco
ID_003 A cavalo dado não se olha ao dente. cavalo,dar,olhar,dente
ID_004 A culpa morreu solteira. culpa,solteiro
ID_005 A curiosidade matou o gato. curiosidade,matar,gato
ID_006 A descer todos os santos ajudam. descer,santos,ajudar
ID_007 A esperança é sempre a última a morrer. esperança,sempre,último,morrer
ID_008 A experiência é a mãe da ciência. experiência,mãe,ciência
ID_009 A falar é que a gente se entende. falar,gente,entender
ID_010 A fé move montanhas. fé,mover,montanha
ID_011 A fome é o melhor tempero. fome,melhor,tempero
ID_012 A galinha da vizinha é sempre melhor que a minha. galinha,vizinho,melhor,minha
ID_013 A idade não perdoa. idade,perdoar
ID_014 A intenção é que conta. intenção,contar
ID_015 A justiça tarda mas não falta. justiça,tardar,faltar

Atenção:

(i) o texto deverá estar em PORTUGUÊS EUROPEU: não te esqueças dos artigos antes dos nomes próprios (e.g. *o* João) e do pronome possessivo (*o* seu amigo), usa apenas vocabulário do português europeu.

(ii) Não deves inserir nenhum provérbio e não podes usar as palavras que formam o provérbio que te foi indicado.

(iii) Os textos não podem ter nenhum newline (parágrafo).