Corpus de aprendizes de português da Universidade de Macau e ensino de português L2

University of Macau Portuguese learner corpus and teaching of Portuguese L2

Jing Zhang [*1] **e Mu You** [†1]

[1] *Universidade de Macau, Faculdade de Letras, Departamento de Português, Macau, China.*

Resumo

O presente artigo apresenta um *corpus* de aprendizes chineses de português L2 com a anotação de *PoS* e lema, destacando-se sua potencialidade de análise quantitativa e qualitativa na identificação de padrões linguísticos dos aprendizes, contribuindo, dessa forma, para o ensino de português L2. Este *corpus* (Corpus de Aprendizes de Português da Universidade de Macau), denominado UMPLC, contém, no total, 933 composições produzidas por 122 estudantes de português da Universidade de Macau durante três anos de estudo consecutivos. A anotação de *PoS* e lema realizou-se através do Stanza, anotador automático desenvolvido por Qi *et al.* (2020). A fim de garantir a consistência de anotação, o resultado foi revisado manualmente. Nesta pesquisa, as informações de *PoS* e lema permitem-nos investigar quantitativa e qualitativamente diversos fenômenos existentes no *corpus* relativos ao aspeto lexical e à mudança diacrônica desse aspeto. Dois estudos foram realizados com base em uma abordagem contrastiva, comparando-se o português dos aprendizes do *corpus* com o português nativo. Foram descobertas características de não-natividade linguística desses aprendizes, o que permitirá que os professores de português L2 se concentrem nas áreas em que é necessário um trabalho corretivo.

Palavras-chave: *Corpus* de aprendizes. Aprendizes chineses de português L2. Análises quantitativas e qualitativas. Aplicações pedagógicas.

Abstract

This article presents a corpus of Chinese learners of Portuguese L2 with *PoS* and lemma annotations, highlighting its potential for quantitative and qualitative analysis in identifying linguistic patterns among learners, thus contributing to the teaching of Portuguese L2. This corpus (University of Macau Portuguese Learners Corpus), named UMPLC, contains a total of 933 compositions produced by 122 Portuguese students from University of Macau over three consecutive years of study. PoS and lemma annotation was performed using Stanza, an automatic annotator developed by Qi *et al.* (2020). To ensure annotation consistency, the results were manually reviewed. In this research, the PoS and lemma information enables us to quantitatively and qualitatively investigate various phenomena in the corpus relating to lexical aspects and diachronic changes in this regard. Two studies were conducted based on a contrastive approach, comparing the Portuguese of learners in the corpus with native Portuguese. Non-native linguistic characteristics were discovered, allowing Portuguese L2 teachers to focus on areas requiring corrective work.

Keywords: Learner corpus. Chinese learners of Portuguese L2. Quantitative and qualitative analysis. Pedagogical applications.

Textoline
Linguagem e Tecnologia

DOI: 10.1590/1983-3652.2024.47754

Seção:
Dossiê

Autor Correspondente:
Jing Zhang

Editor de seção:
Daniervelin Pereira
Editor de layout:
Leonado Araújo

Recebido em:
31 de agosto de 2023
Aceito em:
26 de novembro de 2023
Publicado em:
20 de dezembro de 2023

1 Introdução

A ideia de coletar produções linguísticas de aprendizes não é nova, mas a compilação de *corpora* de aprendizes (CA) eletrônicos é um fenômeno relativamente recente, que teve início no final dos anos 1980 e início dos anos 1990 (Nesselhauf, 2004, p. 128)[p. 5]Granger2002bird. Com o auxílio de computadores, esse tipo de *corpus* apresenta duas vantagens significativas (Granger; Gilquin; Meunier, 2015). Primeiramente, devido à sua grande escala, os dados são mais representativos do que aqueles

*Email: jingz@um.edu.mo
†Email: youmuafonso@gmail.com

que envolvem apenas um número limitado de aprendizes. Em segundo lugar, por estarem em formato eletrônico, os dados podem ser analisados com ferramentas de pesquisa que aceleram a observação e permitem um amplo escopo de estudos que não podem ser realizados manualmente ou apenas com um custo considerável em termos de recursos humanos. Essa natureza eletrônica dos CA permite que os resultados de pesquisa sejam compartilhados e verificados de forma eficiente e confiável (Nesselhauf, 2004, p. 130). Além disso, os dados anotados facilitam consideravelmente as análises linguísticas.

Os CA são frequentemente associados à análise da interlíngua dos aprendizes, servindo como base para descobrir padrões gerais da interlíngua (Cobb, 2003; Nesselhauf, 2004; Santos *et al.*, 2016) e revelando as fases do desenvolvimento linguístico e competência transitória (Selinker, 1992), o que pode facilitar a adaptação do ensino de línguas não maternas. Rundell (1996) enfatiza esse significado óbvio dos CA, considerando que esses *corpora* fornecem informações confiáveis sobre o uso da língua pelos aprendizes, indicando suas dificuldades típicas. Os CA, para Granger (1998), desempenham um papel vital no *design* de ferramentas de ensino de inglês, que podem ser aprimoradas com dados de falantes nativos. Esses dados fornecem informações sobre o que é típico em inglês, enquanto os dados dos não-falantes nativos destacam o que é desafiador para os aprendizes em geral e para grupos específicos de aprendizes.

Um *corpus* de aprendizes pode contribuir de diversas maneiras para os objetivos acadêmicos e pedagógicos. Aqui, destacam-se as aplicações dos CA para diversos fins pedagógicos: i) elaboração, com base em CA, de dicionários[1], gramáticas[2] e outros materiais didáticos adequados a diferentes perfis de aprendizes (Granger, 2004; Nesselhauf, 2004; Santos *et al.*, 2016); ii) *design* de programas de ensino que considerem tanto as informações de frequência lexical quanto as dificuldades identificadas por meio da análise de corpora de aprendentes e de falantes nativos (Granger, 2002, p. 22-23); iii) implementação de metodologias de ensino, incluindo a aprendizagem orientada por dados (*data-driven learning*). Granger e Tribble (1998) argumentam que os CA e corpora de falantes nativos proporcionam aos aprendizes oportunidades para explorar diretamente os fatos linguísticos, estimulando seu aprendizado. Ademais, as descobertas da pesquisa com CA podem ser incorporadas às atividades pedagógicas, oferecendo informações cruciais sobre o que deve ser ensinado e como ensinar (Nesselhauf, 2004, p. 139).

O campo de pesquisa dos CA tem experimentado um notável desenvolvimento. O *Learner Corpora Around the World*, organizado pela Universidade Católica de Lovaina, contém 200 CA, dos quais 176 (88%) são *corpora* monolíngues (102 em inglês e 74 em outras línguas), e 23 (11,4%) são bilíngues ou multilíngues. Entre os CA publicados, destacam-se cinco relacionados ao português, sendo dois monolíngues (o *Learner Corpus of Portuguese L2* (COPLE2), da Universidade de Lisboa; o Corpus Oral de Português como Língua Adicional-Brasil, da Universidade de Limerick, em desenvolvimento) e três bilíngues/multilíngues (o *Multlingual Academic Corpus of Assignments - Writing and Speech* (MACAWS) – Português/Russo, da Universidade de Arizona; o *Multilingual Corpus of Second Language Speech* (MuSSeL) – Chinês Mandarim, Francês, Português e Espanhol, da Universidade de Utah; o *Leiden Learner Corpus* – Holandês, Francês, Italiano, Português e Espanhol, da Universidade de Leiden. Além do primeiro *corpus* monolíngue em português anteriormente mencionado, salienta-se ainda o Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2) da Universidade de Coimbra. Ambos compartilham características comuns, consistindo em dados sincrônicos produzidos por falantes de diferentes línguas maternas, incluindo o chinês.

Na China, o primeiro *corpus* de aprendizes chineses, conhecido como o Corpus de Aprendizes Chineses de Inglês (CLEC), foi criado em 1999 pelas Universidades de Estudos Estrangeiros de Guangdong e de Jiaotong de Xangai (Yang, 2001, p. 62). Esse *corpus* é composto por materiais escritos por estudantes de Inglês Profissional, Inglês Universitário e Inglês do Ensino Secundário. Desde então, surgiram outros CA, que predominantemente representam produções em inglês. Simultaneamente, a

[1] Há vários dicionários baseado em CA dedicados à língua inglesa (Granger, 2004, p. 136)[p. 137]nesselhauf2004learner, tais como o *Longman Essential Activator*, o *Longman Dictionary of Contemporary English*, o *Cambridge International Dictionary of English*, o *Cambridge Advanced Learner's Dictionary*, entre outros. No que diz respeito à língua portuguesa, destaca-se o *A frequency dictionary of Portuguese* (Davies; Preto-Bay, 2008), que é baseado em um *corpus* de falantes nativos.

[2] TeleNex e o *Longman dictionary of common errors* (Turton; Heaton, 1996) são dois exemplos.

abordagem metodológica baseada em CA também tem se desenvolvido.

No que diz respeito à língua portuguesa, não há registros de existência de um *corpus* de aprendizes chineses na China. Nossa pesquisa busca preencher essa lacuna por meio da construção de um *corpus* de produções escritas em português L2. Nas seções seguintes, descreveremos esse *corpus* e suas características quantitativas, bem como suas aplicações pedagógicas.

2 Construção de um *corpus* de aprendizes chineses de português L2

Nesta seção, abordamos a contextualização da pesquisa, o *design* do *corpus*, a coleta e documentação de dados, assim como sua estruturação e a anotação.

2.1 Contextualização da pesquisa

Devido ao fortalecimento das relações entre a China e os países de língua portuguesa, estamos testemunhando um crescimento significativo na demanda por profissionais que dominam bem a língua portuguesa. Grosso *et al.* (2021), no Referencial para o Ensino de Português Língua Estrangeira na China, destacam que o português é considerado uma das línguas não comuns[3] mais procuradas na China Interior.

Em Macau, o português, cujo estatuto oficial é garantido pela Lei Básica da Região Administrativa Especial de Macau, continua sendo principalmente utilizado na administração pública e no sistema judicial. No entanto, não é frequentemente utilizado na vida cotidiana dos residentes de Macau. De acordo com dados estatísticos fornecidos pela Direção dos Serviços de Estatística e Censos de Macau[4], em termos de domínio geral de línguas, 86,2% da população fala fluentemente cantonês como meio de comunicação, 45,0% fala mandarim, 22,7% fala inglês e 2,3% fala português.

Para os jovens de Macau, o ensino universitário representa uma das principais oportunidades para aprenderem português. Atualmente, há uma justificativa clara para o aumento de cursos de língua portuguesa, bem como o crescimento do número de estudantes chineses de português, tanto na China Interior quanto em Macau. O referencial mencionado anteriormente lista 52 instituições de ensino superior que oferecem cursos de português na China Interior e seis em Macau.

Com o maior departamento de ensino de português na Ásia e excelentes condições acadêmicas, a Universidade de Macau atrai jovens chineses de Macau e da China Interior para aprenderem português. Isso a coloca em uma posição ideal para desenvolver um projeto de pesquisa sobre a construção e aplicação do *corpus* de aprendizes de português L2. Um dos fatores a serem considerados é a regularidade do aprendizado de estudantes universitários que produzem materiais linguísticos de forma sistemática, o que pode tornar a pesquisa em questão mais produtiva.

Neste contexto, o Departamento de Português e o Laboratório de Processamento de Linguagem Natural e Tradução Automática Português-Chinês da Universidade de Macau iniciaram conjuntamente o projeto Corpus de Aprendizes Chineses de Português L2 (UMPLC) no primeiro semestre do ano letivo 2018/2019, e sua conclusão está prevista para o segundo semestre do ano letivo 2023/2024.

2.2 *Design* do UMPLC

O UMPLC, composto por produções escritas de aprendizes adultos chineses de português L2 matriculados no ensino superior, tem como objetivo criar uma plataforma que ofereça ferramentas de pesquisa sobre o processo e os resultados do ensino e da aquisição de L2. Atualmente, encontra-se na fase final de desenvolvimento.

O projeto envolve 122 alunos matriculados no Curso de Licenciatura ou no Curso de Minor em Estudos Portugueses da Universidade de Macau, os quais são provenientes de Macau e da China Interior e têm o chinês como língua materna. Dentre esses alunos, 47 participaram de todas as onze etapas diferentes de coleta de dados ao longo dos seus primeiros três anos de estudo, nos quais frequentaram disciplinas de compreensão e produção escrita em português L2. Devido à existência de

[3] Conforme mencionado por Yan (2019), o termo 'línguas não comuns' refere-se às línguas estrangeiras excluindo o inglês, russo, japonês, francês, alemão, espanhol e árabe. Esse conceito foi criado para simplificar a organização e a administração do ensino de idiomas.

[4] Disponível em Resultados Globais dos Censos 2021.

um nível de proficiência linguística estabelecido de acordo com a organização curricular dos cursos, que os alunos devem alcançar no final de cada ano de estudo, é possível avaliar as competências linguísticas dos informantes em cada fase da coleta de dados. Assim, trata-se de dados diacrônicos, que podem fornecer evidências das diferentes etapas na evolução da língua (Kübler; Zinsmeister, 2015).

Cada participante é solicitado a preencher um formulário de informações pessoais. Os dados coletados, apresentados na Tabela 1, são posteriormente integrados aos metadados do corpus.

Tabela 1. Perfis dos participantes.

Número total de participantes	122
Instituição de ensino superior	Universidade de Macau
Curso que frequentam	Curso de Licenciatura em Estudos Portugueses Curso de *Minor* em Estudos Portugueses
Contexto de aprendizado	Português L2
Idade	18 a 22 anos
Língua materna	Chinês
Origem	Macau e China Interior

Fonte: Autoria própria.

2.3 Coleta e documentação de dados

Os dados consistem em avaliações textuais (realizadas em testes, exames semestrais e finais) produzidas pelos estudantes dentro da sala de aula, sem acesso a dicionários ou livros de referência e dentro de um limite de tempo. Esses dados se apresentam em duas formas: a maioria são textos manuscritos, enquanto uma pequena parte são textos eletrônicos[5]. Os manuscritos foram inicialmente digitalizados e armazenados em formato PDF e, posteriormente, transcritos manualmente. A tentativa de obtenção automática de arquivos digitais através de OCR não foi bem-sucedida devido à natureza manuscrita dos textos.

A etapa mais demorada do processo é a coleta de textos, que ocorreu ao longo de três anos letivos e seis semestres, de 2018 a 2021. Todos os dados foram digitados manualmente e passaram por uma verificação ortográfica realizada por um grupo de assistentes, composto por mestrandos e doutorandos dos cursos de Linguística Aplicada (Português) e Tradução Português-Chinês. Posteriormente, esses dados foram anotados com informações linguísticas, divididas em duas estruturas: *PoS* (*Part of Speech*) e lema (a forma canônica da palavra). Além disso, os dados originais passaram por correções linguísticas providenciadas por professores de Português, que são falantes nativos da língua.

Dessa forma, este conjunto de dados disponibiliza cinco tipos distintos de informações: dados originais gravados em PDF, dados transcritos digitalmente, que permanecem idênticos aos originais, dados com correções ortográficas, dados anotados linguisticamente e dados com correções linguísticas.

2.4 Estruturação do UMPLC

Os metadados dividem-se em duas partes distintas. A primeira parte está relacionada aos dados pessoais dos participantes, conforme apresentado na Tabela 1. A segunda parte abrange informações sobre os textos. A Tabela 2 revela que o *corpus* abrange um total de 933 textos, contendo 171914 palavras. Esses textos foram coletados em onze momentos diferentes, abrangendo um período de três anos e seis semestres. No primeiro ano, foram registrados cinco momentos, enquanto no segundo e terceiro ano, foram registrados três momentos em cada. Nos dois últimos momentos, foram abordados dois temas distintos em cada ocasião, e o tempo de avaliação dobrou em relação aos períodos anteriores. Diante dessa configuração, os textos com temas diferentes foram agrupados em subconjuntos separados, resultando, assim, em quatro subcorpora: CA10A, CA10B, CA11A e CA11B. Com base

[5] O 8º momento da coleta de dados coincidiu com o período da pandemia da Covid-19, o que levou os estudantes da Universidade de Macau a realizar as avaliações de forma *online*.

nas metas estabelecidas para cada ano de estudo, os estudantes que completarem o segundo ano alcançarão o nível Intermediário no sistema Celpe-Bras. Já os que concluírem o terceiro ano atingirão o nível Intermediário Superior.

Os textos coletados são narrativos, argumentativos e descritivos, abordando uma variedade de tópicos definidos pelos professores que ministram as disciplinas de Compreensão e Produção Escrita em Língua Portuguesa. Esses textos refletem, assim, o aprendizado dos alunos na sala de aula. As informações sobre os textos que compõem o *corpus* são apresentadas na Tabela 2.

Tabela 2. Informações básicas do UMPLC.

Momento	Título de subcorpus	Tipo de texto	Número de texto	Número de palavra	Data de coleta	Ano_Semestre em que estão os participantes	Nível de proficiência - Celpe-Bras
1	CA1	Descritivo	82	12245	06/11/2018	1_1	—
2	CA2	Narrativo	82	15474	12/12/2018	1_1	—
3	CA3	Narrativo	67	9922	27/02/2019	1_2	—
4	CA4	Narrativo	70	10014	14/03/2019	1_2	—
5	CA5	Narrativo	70	10322	09/05/2019	1_2	—
6	CA6	Narrativo	65	11610	10/10/2019	2_1	—
7	CA7	Argumentativo	65	11869	14/12/2019	2_1	—
8	CA8	Argumentativo	61	10821	30/05/2020	2_2	Intermediário
9	CA9	Argumentativo	67	13341	19/11/2020	3_1	—
10	CA10A	Argumentativo	68	18135	06/12/2020	3_1	—
	CA10B	Argumentativo	68	12079			
11	CA11A	Argumentativo	84	24776	25/05/2021	3_2	Intermediário Superior
	CA11B	Narrativo	84	11306			
Total	—	—	933	171914	—	—	—

Fonte: Autoria própria.

2.5 Anotação do UMPLC

A anotação é um processo interpretativo que adiciona informações linguísticas, como morfológicas, semânticas, sintáticas, discursivas, etc., a textos (Garside; Leech; McEnery, 1997, p. 2). Independentemente de ser realizada de forma automática ou não, a anotação fornece o mesmo tipo de informações que as análises linguísticas tradicionais ofereceriam. Em outras palavras, trata-se de uma prática que analisa os textos e fornece os resultados de forma sistemática e acessível (McEnery; Hardie, 2011, p. 13). Isso significa que um *corpus* anotado possui pelo menos duas vantagens: as informações nele armazenadas não apenas facilitam substancialmente investigações linguísticas, evitando a necessidade de repetir o trabalho árduo de examinar os textos palavra por palavra, mas também estabelecem uma base comum a partir da qual diferentes pesquisas podem ser conduzidas com maior consistência.

O nosso *corpus* atual conta com dois tipos de anotação, *PoS* e lema, que foram realizados seguindo o esquema *Universal Dependencies* (UD) (Marneffe *et al.*, 2021). A escolha do UD foi motivada por duas razões principais. Primeiramente, esse esquema é caracterizado por um conjunto de etiquetas e diretrizes que possibilitam uma anotação consistente entre diferentes idiomas, e até mesmo em todos eles, e permite extensões para línguas específicas[6]. O UD já foi adotado por 183 *treebanks* em 104 línguas (Marneffe *et al.*, 2021, p. 256), e tanto os *treebanks* quanto os documentos de anotação acumulados durante a sua compilação representam um recurso valioso. Em segundo lugar, o *tagset* do UD é relativamente simples, o que contribui para a qualidade e a rapidez da anotação manual.

A anotação é realizada em duas etapas. A primeira é feita automaticamente com o uso do Stanza (Qi *et al.*, 2020), uma ferramenta que pode processar textos em 70 idiomas, seguindo o esquema de UD. No entanto, em vez da versão original, optou-se por anotar a versão revisada ortograficamente, uma vez que erros ortográficos geralmente não são relevantes e podem afetar a anotação, resultando em etiquetas incorretas e difíceis de corrigir. Como o Stanza é treinado com dados de falantes nativos,

[6] Para obter informações adicionais, visite o site oficial da UD.

qualquer desvio encontrado nos textos dos aprendizes poderia diminuir a precisão da anotação. Por esse motivo, na segunda etapa, foi realizada uma revisão manual conduzida pelo mesmo grupo de assistentes estudantes mencionado anteriormente. A revisão humana reduziu significativamente os erros de anotação, embora seja inevitável a ocorrência de alguns erros isolados.

Apesar das várias diretrizes relacionadas às possíveis questões linguísticas na anotação de *corpus* em português (Paiva; Real, 2016; Rademaker *et al.*, 2017), durante a fase de revisão, foi identificado um grande número de problemas não abordados anteriormente. O presente artigo não aborda esses detalhes específicos, no entanto, enfatiza uma modificação significativa no *tagset* do UD. Devido à presença de palavras não convencionais no UMPLC, criadas pelos estudantes ou ilegíveis, foi necessário adicionar uma etiqueta ("UNKNOWN") para marcá-las. No total, foram utilizadas 17 etiquetas para a anotação de *PoS* (ver Tabela 3).

Tabela 3. *Tagset* do UMPLC.

Etiqueta	Exemplos
AUX (verbo auxiliar)	ser, estar
ADJ (adjetivo)	alto, comprido
ADP (preposição)	de, em
ADV (advérbio)	não, já
CCONJ (conjunção coordenativa)	e, mas
DET (determinante)	o, um
INTJ (interjeição)	obrigado, olá
NOUN (substantivo)	país, dia
NUM (número)	3, 66
PRON (pronome)	eu, que
PROPN (substantivo próprio)	Sofia, Ronaldo
PUNCT (pontuação)	“, ?
SCONJ (conjunção subordinativa)	porque, quando
SYM (símbolo)	$, ℃
UNKNOWN (palavra inexistente)	aúnueu, attche
VERB (verbo pleno)	querer, fazer
X (palavra estrangeira)	part-time, cosplay

Fonte: Autoria própria.

Um *corpus* de aprendizes, especialmente aqueles com anotações, representa um recurso valioso para pesquisas relacionadas ao ensino e aquisição do português L2. Com base no UMPLC e em *corpora* de referência, que incluem o ptTenTen20 (*corpus* de português L1 com *c.* 12,5 bilhões de palavras) e o zhTenTen17 (*corpus* de chinês L1 com *c.* 13,5 bilhões de palavras), além de três pequenos *corpora* de português L1 desenvolvidos pelos próprios autores, conduziram-se dois estudos. O primeiro concentra-se na análise da variação lexical entre os textos produzidos por falantes nativos e não nativos, enquanto o segundo explora o uso de advérbios, com exemplos de dois deles: "especialmente" e "nomeadamente".

3 Estudo I: variação lexical

Para avaliar o vocabulário utilizado pelos aprendizes no UMPLC, empregamos o índice *Type/Token Ratio* (TTR), que mede a variação lexical (Wolfe-Quintero; Inagaki; Kim, 1998). O TTR é calculado como a relação entre o número de palavras diferentes e o número total de palavras em um texto[7]. Quanto mais elevado o valor, maior a diversidade lexical presente em um determinado texto.

Para evidenciar eventuais disparidades entre os textos produzidos por falantes nativos e não nativos, além do UMPLC, foram introduzidos cinco subcorpora selecionados aleatoriamente do ptTenTen20,

[7] É de notar que a definição de "palavra" não abrange as pontuações neste artigo.

denominados "SUB1", "SUB2", "SUB3", "SUB4" e "SUB5". Além disso, foram utilizados três *corpora* especialmente compilados pelos próprios autores, denominados "CPCOVID", "CPMESSI" e "CPMPOX", para servirem como subcorpora de referência.

O UMPLC está subdividido em 13 subcorpora, designados como "CA1", "CA2", "CA3", "CA4", "CA5", "CA6", "CA7", "CA8", "CA9", "CA10A", "CA10B", "CA11A" e "CA11B", que registram o desempenho dos aprendizes em cada momento de coleta. É importante notar que o valor do TTR é influenciado pela extensão do texto, geralmente diminuindo à medida que os textos se tornam mais longos. Portanto, a escala dos subcorpora de referência foi ajustada para garantir uma comparação justa com os subcorpora do UMPLC. Os corpora CPCOVID, CPMESSI e CPMPOX consistem em notícias recentes coletadas através do *Google News*, com base nas palavras-chave: "surto de COVID", "Messi campeão" e "surto de Mpox em Portugal". Essa seleção foi feita de modo a cada *corpus* ter um tema específico, permitindo uma comparação significativa entre eles e os subcorpora do UMPLC. No total, foram utilizados 21 subcorpora, dos quais oito são (sub)corpora de referência. Os dados pertinentes estão resumidos na Tabela 4 e ilustrados na Figura 1.

Tabela 4. TTRs e dados relevantes dos 21 (sub)corpora.

(Sub)corpus	Ano_Semestre	Número de Tipo	Número de Palavra	TTR(%)
CA1	1_1	416	12245	3,4
CA2	1_1	665	15474	4,3
CA3	1_2	1403	9922	14,14
CA4	1_2	1064	10014	10,63
CA5	1_2	1282	10322	12,42
CA6	2_1	1929	11610	16,61
CA7	2_1	1530	11869	12,89
CA8	2_2	1824	10821	16,86
CA9	3_1	2114	13341	15,85
CA10A	3_1	2370	18135	13,07
CA10B	3_1	1539	12079	12,74
CA11A	3_2	2804	24776	11,32
CA11B	3_2	1199	11306	10,6
SUB1	—	2024	7593	26,66
SUB2	—	2570	11347	22,65
SUB3	—	3605	15647	23,04
SUB4	—	4856	22168	21,91
SUB5	—	5343	25628	20,85
CPCOVID	—	2328	9434	24,68
CPMESSI	—	1687	8220	20,52
CPMPOX	—	1587	10329	15,36

Fonte: Autoria própria.

É possível notar que os TTRs dos CA1 e CA2 são notadamente inferiores em comparação aos outros subcorpora do UMPLC. Por contraste, o CA3, coletado durante o segundo semestre do primeiro ano, exibe um TTR mais elevado (14,14 versus 3,4 e 4,3). Além disso, seu número de tipos de palavras é significativamente maior (1403 versus 416 e 665). Esses resultados sugerem que durante o primeiro semestre do primeiro ano, o vocabulário utilizado permanece bastante restrito, e apenas no segundo semestre os aprendizes começam a incorporar uma variedade maior de palavras em sua escrita.

Após o primeiro semestre do primeiro ano, o TTR segue um padrão oscilante: entre o momento 3 e o momento 8, o TTR flutua, atingindo o valor mais elevado no CA8. Posteriormente ao momento 8, o TTR diminui constantemente, alcançando o menor valor no CA11B. Considerando que é improvável que o conhecimento lexical dos aprendizes mude de maneira tão abrupta, argumenta-se que essa variação no TTR não reflete necessariamente o vocabulário que os estudantes dominam, mas, em vez disso, está associada aos tópicos abordados nas composições.

Em outras palavras, os tópicos podem delimitar, em maior ou menor grau, o conteúdo e, por

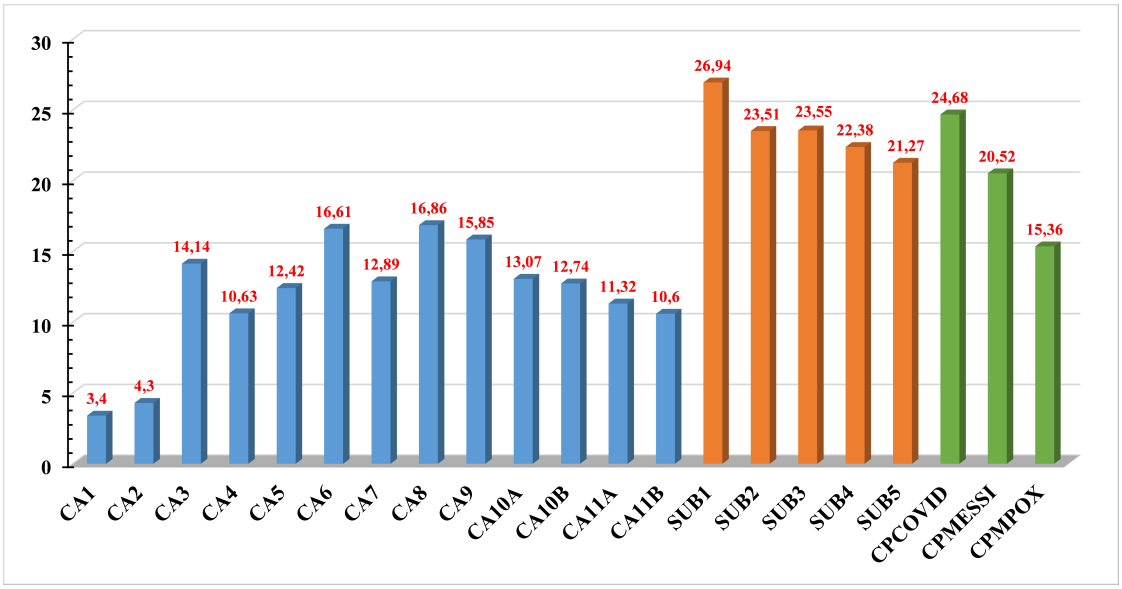

Figura 1. TTRs dos 21 (sub)corpora.
Fonte: Autoria própria.

consequência, o vocabulário nas composições, o que afeta o TTR. Se o conteúdo de um grupo de composições for altamente homogêneo, é natural que o vocabulário não apresente muita variação.

Para testar a hipótese de que a variação do TTR, conforme descrita anteriormente, está relacionada aos tópicos das composições, examinamos os tópicos das composições que compõem o CA8 e o CA11B[8] (ver Tabela 5), ambos com números semelhantes de palavras (10821 versus 11306), mas TTRs significativamente diferentes (16,86 versus 10,6).

Tabela 5. Tópicos dos CA8 e CA11B

O tópico das composições que compõem o CA8	Nos últimos meses tem-se mudado muito a nossa vida quotidiana devido ao surto da pandemia de COVID-19. Com o fechamento de escolas a nível mundial, os professores e alunos tiveram de recorrer a novos meios de ensino e aprendizagem, tudo à distância. Como é que você vê essa mudança que tem experimentado ao longo do semestre? Quais são as vantagens e desvantagens? Apresente a sua opinião e reflexão num texto de 150-200 palavras.
O tópico das composições que compõem o CA11B	Após assistir ao vídeo por três vezes, escreva uma notícia que resuma as suas principais informações. A sua notícia deve ter entre 150 e 300 palavras e conter título. O subtítulo e o lead são opcionais.

Fonte: Autoria própria.

Observa-se que o tópico 8 permite uma maior flexibilidade, permitindo que os estudantes expressem suas próprias opiniões, enquanto no tópico 11B, os alunos são explicitamente instruídos a escrever

[8] O tópico do CA11B utilizou dois vídeos para diferentes turmas: o vídeo"Excursões Subsidiadas" para as turmas 1 e 2, e o vídeo "Andy Wu prevê 40 mil turistas por dia em Maio" para as turmas 3 e 4.

um texto contendo informações específicas. Essa constatação está em conformidade com a nossa hipótese.

Para avaliar a influência do tópico no uso de vocabulário, examinamos as 10 trigramas de palavras mais frequentes (TPMFs) nos CA8 e CA11B, concentrando-nos na frequência das TPMFs e na frequência relativa de documento (FRD), ou seja, a proporção de documentos nos quais um determinado item aparece[9] (ver Tabela 6 e Tabela 7, respectivamente).

Tabela 6. As 10 TPMFs no CA8

Item	Frequência	FRD(%)
ensino a distância	23	34,43
pandemia de covid-19	23	19,67
vantagens e desvantagens	21	32,79
surto da pandemia	19	26,23
da pandemia de	17	24,59
o ensino à	17	14,75
na minha opinião	14	22,95
ensino e aprendizagem	14	19,67
de ensino e	13	18,03
em casa e	12	16,39
	Total: 173	Média: 22,95

Fonte: Autoria própria.

Tabela 7. As 10 TPMF no CA11B

Item	Frequência	FRD(%)
serviços de turismo	93	78,57
dos serviços de	79	71,43
taxa média de	53	41,67
média de ocupação	50	40,48
helena de senna	49	48,81
taxa de ocupação	45	36,90
de senna fernandes	42	41,67
de ocupação hoteleira	41	34,52
a taxa média	37	35,71
a taxa de	36	30,95
	Total: 525	Média: 46,07

Fonte: Autoria própria.

Os dados indicam que no CA11B, as 10 TPMFs são usadas proporcionalmente com maior frequência do que no CA8. Mesmo que a escala dos dois subcorpora seja semelhante, no CA11B, em média, as 10 TPMFs estão presentes em 46,07% das composições, o que é cerca de duas vezes mais do que no CA8 (22,95%). Além disso, a frequência total das 10 TPMFs no CA11B é consideravelmente maior (689 versus 168). Essas descobertas corroboram a hipótese de que um tópico com mais restrições pode resultar em um uso lexical mais limitado e repetitivo.

Uma análise mais detalhada das TPMFs revela que, essencialmente, elas ecoam os tópicos abordados, visto que quase todas essas trigramas podem ser encontradas nos tópicos ou materiais relacionados. Isso sugere que as composições dos estudantes, além de expressar suas próprias ideias, incorporam, em diferentes graus, a linguagem daqueles que elaboram os materiais (Kreyer, 2017).

[9] As TPMFs e as estatísticas são extraídas utilizando o Sketch Engine.

Esse fenômeno, mais uma vez, ilustra a influência dos tópicos nas composições dos alunos.

A partir dessas constatações, pode-se argumentar que a variação do TTR está intrinsecamente associada aos tópicos abordados nas composições. Ao avaliar as composições dos aprendizes, é importante interpretar o TTR com cautela, uma vez que um TTR mais baixo não necessariamente indica um domínio lexical inferior. No entanto, o TTR ainda fornece um ponto de referência para avaliar o conhecimento lexical dos aprendizes, permitindo uma visão geral da evolução diacrônica da variação lexical, especialmente quando o progresso é evidente, como demonstrado na comparação entre o CA1, o CA2 e o CA3.

A seguir, conduzimos análises quantitativas com o objetivo de investigar possíveis diferenças na variação lexical entre os subcorpora do UMPLC e os (sub)corpora de referência (ptTenTen20, CPCOVID, CPMESSI e CPMPOX). Em comparação com os subcorpora do UMPLC, os do ptTenTen20 exibem TTRs significativamente mais elevados: no primeiro grupo, a média do TTR é de 11,91%, enquanto no último, a média é de 23,02%. Em relação ao número de tipos de palavras, o menor subcorpus do ptTenTen20 (SUB1) contém mais tipos do que a maioria dos subcorpora do UMPLC.

Esses dados indicam que as composições apresentam um nível de variação lexical mais baixo do que os textos produzidos por falantes nativos. No entanto, como já mencionado, o TTR representa a diversidade de vocabulário em uso, o qual, por sua vez, pode ser influenciado pelos temas abordados. Portanto, a discrepância do TTR entre esses dois tipos de subcorpora, em vez de refletir uma diferença genuína no conhecimento lexical entre falantes nativos e não nativos, pode ser atribuída ao fato de que, ao contrário das composições que se concentram em um tópico específico, os textos do ptTenTen20 abrangem uma ampla variedade de tópicos.

Para uma análise mais aprofundada das razões subjacentes à diferença na variação lexical mencionada anteriormente, contrastamos os subcorpora do UMPLC com o CPCOVID, o CPMESSI e o CPMPOX. Esses três *corpora* foram construídos em torno de palavras-chave específicas, resultando em conteúdo mais homogêneo em comparação com os subcorpora do ptTenTen20, que consistem em textos aleatórios. Entre os três corpora, o conteúdo do CPCOVID provavelmente é o menos homogêneo, uma vez que as notícias frequentemente abordam tópicos variados, como a reintrodução de máscaras, novas variantes e o cancelamento de políticas relacionadas à pandemia. Em contrapartida, o CPMESSI e o CPMPOX devem possuir um conteúdo mais concentrado, já que o primeiro se concentra principalmente na vitória de Leo Messi na Copa das Ligas, e o segundo aborda os surtos recentes de Mpox em Portugal. Para quantificar o grau de homogeneidade, calculamos a média da Frequência Relativa de Documento (FRD) das 10 TPMFs (ver Tabela 8). Os resultados indicam que essa métrica corresponde bem às avaliações subjetivas mencionadas neste parágrafo.

Tabela 8. Média da FRD das 10 TPMFs dos CPCOVID, CPMESSI e CPMPOX

Corpus	Média da FRD das 10 TPMFs
CPCOVID	29,23
CPMESSI	49,09
CPMPOX	55,83

Fonte: Autoria própria.

A partir da média da FRD das 10 TPMFs, nota-se que o nível de homogeneidade nos CA11B, CPMESSI e CPMPOX é essencialmente equivalente, com o CA11B apresentando uma média da FRD ligeiramente inferior em comparação com os outros dois *corpora* (46,07% versus 49,09% e 55,83%). Da mesma forma, o CA8 e o CPCOVID exibem um nível de homogeneidade semelhante, uma vez que as médias da FRD das 10 TPMFs não diferem significativamente (22,95% *versus* 29,23%).

Com base nesses dados, podemos concluir que o vocabulário dos CPMESSI e CPMPOX é mais diversificado em comparação com o do CA11B, e o vocabulário do CPCOVID é mais variado do que o do CA8. Isso fica claramente evidente quando se analisa o TTR: o TTR nos CPMESSI e CPMPOX

é aproximadamente duas e uma vez e meia maior do que o do CA11B, respectivamente (20,52% e 15,36% versus 10,6%), enquanto o TTR do CPCOVID é cerca de uma vez e meia maior do que o do CA8 (24,68% versus 16,86%). Além disso, o número de tipos de palavras nos CPMESSI, CPMPOX e CPCOVID também é maior, apesar de o CA11B e o CA8 conterem mais palavras no total. Isso sugere que, em condições semelhantes, os aprendizes utilizam um vocabulário menos variado em comparação com os falantes nativos.

Em resumo, a partir das análises apresentadas, podemos chegar a três conclusões. Primeiro, para os estudantes em questão, o primeiro semestre do 1º ano é o período em que o uso de vocabulário permanece mais limitado; somente após esse período eles começam a ser capazes de utilizar palavras mais diversificadas em sua escrita. Segundo, a variação no TTR observada nos subcorpora CA3 a CA11B parece estar relacionada à influência dos tópicos abordados, e não reflete uma mudança real no conhecimento lexical dos estudantes. Terceiro, os dados da comparação entre os subcorpora do UMPLC e os subcorpora de referência revelam que os aprendizes chineses de português L2 não escrevem com o mesmo nível de variação lexical que os falantes nativos.

4 Estudo II: advérbios “especialmente” e “nomeadamente”

Nesta seção, nossa intenção é investigar, tanto quantitativamente quanto qualitativamente, o uso de advérbios que apresentam o sufixo "mente". É importante destacar que a categoria gramatical dos advérbios não costuma ser facilmente confundida com outras categorias nos sistemas de anotação.

Inicialmente, procedemos à análise dos advérbios terminados em "mente" nos subcorpora *Portuguese TLD* (PT) e *Brazilian TLD* (BT) - dois subconjuntos presentes no ptTenTen20 que representam o português europeu e o brasileiro, respectivamente. Essa investigação teve como objetivo identificar padrões e fenômenos intrigantes que merecessem uma investigação mais aprofundada. Verificamos que o advérbio "especialmente" é um dos mais frequentemente empregados entre os advérbios com essa terminação em ambos os subcorpora.

Além disso, notamos que o advérbio "nomeadamente" é o mais comum entre os advérbios terminados em "mente" no PT, mas sua ocorrência é consideravelmente menor no BT. Realizamos um teste qui-quadrado, constatando que "nomeadamente" é significativamente mais frequente no PT em comparação ao BT ($\chi^2 = 1501677$, df $= 1$, p $< 0{,}001$). Os dados relevantes estão apresentados na Tabela 9.

Tabela 9. "Especialmente" e "Nomeadamente" nos PT e BT

	Especialmente		Nomeadamente	
	Frequência	Frequência Relativa†	Frequência	Frequência Relativa
PT	77113	102,32	181802	241,24
BT	809578	120,36	11809	1,76

Fonte: Autoria própria.
Nota: † Frequência por um milhão de palavras.

A seguir, examinamos os advérbios terminados em “mente” no UMPLC. Neste corpus de aprendizes, o advérbio "especialmente" é muito mais frequente do que "nomeadamente" - o primeiro aparece 62 vezes, enquanto o último ocorre apenas 4 vezes. O português europeu é a variante usada em Macau. Levando isso em consideração, nesta seção, o UMPLC será comparado com o PT, em vez do BT.

Para determinar se há uma diferença quantitativa estatisticamente significativa entre os dois advérbios no UMPLC e no PT, conduzimos dois testes estatísticos: um teste qui-quadrado e um teste g. O teste qui-quadrado indica que, em relação ao PT, o "especialmente" é utilizado de forma significativamente mais frequente no UMPLC (χ^2=109,52, df=1, p<0,001). Já o teste g revela que, comparado ao PT, o "nomeadamente" é empregado de maneira significativamente menos frequente no UMPLC (G=56,236, df=1, p<0,001).

Com o objetivo de realizar uma análise detalhada dos dois advérbios, examinamos seus usos e classificamos esses advérbios com base no sentido que eles assumem no contexto. Primeiramente,

apresentamos a análise referente ao advérbio "especialmente".

De acordo com dois dicionários online[10], o advérbio "especialmente" tem cinco significados: A. muito, bastante, B. somente para um certo fim, objetivo ou propósito, C. apenas para uma certa pessoa ou para um certo grupo de pessoas, D. em consideração ou na intenção de, E. de um modo mais especial, acima de tudo, sobretudo, principalmente. Os resultados da análise indicam que no UMPLC, o uso deste advérbio é bastante restrito, sendo utilizado principalmente com dois significados: o significado E é predominante, representando 96,77% dos casos, enquanto o significado A aparece apenas duas vezes. Essas informações são apresentadas na Tabela 10.

Tabela 10. Distribuição do "Especialmente" no UMPLC

Significado	Frequência	Percentagem
A	2	3,32
B	0	0
C	0	0
D	0	0
E	60	96,77

Fonte: Autoria própria.

Para avaliar se essa concentração também é observada na escrita dos falantes nativos, foram aleatoriamente selecionadas 50 frases contendo o advérbio "especialmente" do PT. Essas frases foram classificadas utilizando o mesmo método, e os resultados são apresentados na Tabela 11. Pode-se notar que, nessa amostra, o advérbio também é predominantemente utilizado com o significado E. No entanto, o uso é relativamente mais diversificado em comparação com o UMPLC, abrangendo quatro significados distintos.

Tabela 11. Distribuição do "Especialmente" na Amosta do PT

Significado	Frequência	Percentagem
A	6	12
B	1	2
C	0	0
D	5	10
E	38	76

Fonte: Autoria própria.

Além dessa diferença, observou-se que no UMPLC, o uso do advérbio "especialmente" apresenta uma característica notável: o advérbio frequentemente é colocado no início das frases (9 vezes), enquanto na amostra do PT, ele aparece quase sempre no meio das frases (49 vezes). Para uma melhor compreensão desse fenômeno, investigamos o uso de "especialmente" no PT e descobrimos que, das 77113 ocorrências, o advérbio aparece apenas 3331 vezes no início das frases (4,32%).

Com base nesses dados, realizamos um teste g, cujo resultado ($G = 9{,}8604$, $df = 1$, $p < 0{,}01$) indica que o "especialmente" é colocado significativamente com mais frequência no início das frases no UMPLC em comparação com o PT.

Suscita-se a hipótese de que a tendência dos aprendizes chineses de colocarem o advérbio "especialmente" mais frequentemente no início das frases decorre da interferência negativa da língua chinesa, já que "yóu qí shì" e "tè bié shì" (duas possíveis traduções para "especialmente") aparentam ocorrer com frequência na posição inicial das sentenças em chinês. Para testar essa hipótese, realizamos uma pesquisa nas duas expressões chinesas no *corpus* zhTenTen17 e constatamos que "yóu qí shì" ocorre

[10] Aulete e Infopédia.

1653524 vezes, sendo que 16,93% delas (279915) estão no início das sentenças, enquanto "tè bié shì" ocorre 1293428 vezes, com 25,36% delas (328032) no início das sentenças.

Para avaliar se existe uma diferença quantitativa estatisticamente significativa na ocorrência de "especialmente", "yóu qí shì" e "tè bié shì" no início das frases, conduzimos quatro testes de qui-quadrado. As comparações entre o PT e o corpus zhTenTen17 indicam que tanto "yóu qí shì" (χ^2=8556,4, df=1, p<0,001) quanto "tè bié shì" (χ^2=17575, df=1, p<0,001) ocorrem significativamente mais frequentemente no início das sentenças em comparação com "especialmente" no PT.

Por outro lado, as comparações entre o UMPLC e o *corpus* zhTenTen17 demonstram que "yóu qí shì" (χ^2=0,11368, df=1, p=0,736 > 0,05) e "tè bié shì" (χ^2=3,3007, df=1, p=0,06925 > 0,05) no zhTenTen17 não ocorrem com maior frequência no início das sentenças em comparação com "especialmente" no UMPLC. Em resumo, no zhTenTen17, "yóu qí shì" e "tè bié shì" são mais frequentemente usados no início das sentenças do que "especialmente" no PT, e essas mesmas expressões chinesas apresentam uma taxa similar de ocorrência no início das sentenças em comparação com "especialmente" no UMPLC.

Com base nesses resultados, é possível afirmar que a tendência dos aprendizes de colocar "especialmente" no início das sentenças é resultado da interferência da língua chinesa.

O advérbio "nomeadamente" foi analisado de maneira semelhante. Identificamos sentenças contendo esse advérbio no UMPLC e, em seguida, categorizamos seu uso de acordo com os significados listados nos dicionários mencionados anteriormente: A. dando nome a, indicando o nome de, com ou segundo declinação dos nomes, B. mais exatamente, mais especificamente C. principalmente, mormente. No UMPLC, encontramos apenas quatro ocorrências de "nomeadamente", das quais três apresentam o significado C. e uma inclui um erro inexplicável. Isso sugere que os aprendizes não estão familiarizados com o uso desse advérbio. Com base nos três casos em que o advérbio é corretamente empregado, é possível deduzir que os estudantes simplesmente o consideram um sinônimo de "especialmente", utilizando-o apenas para destacar informações específicas.

Foram aleatoriamente selecionadas 50 sentenças contendo a palavra "nomeadamente" do PT e analisadas (ver Tabela 12). Os resultados evidenciam que, para os falantes nativos, "nomeadamente" é principalmente utilizado para detalhar informações (significado B). Em outras palavras, o uso desse advérbio é diferente no PT em comparação com o UMPLC.

Tabela 12. Distribuição do "Nomeadamente" na Amosta do PT

Significado	Frequência	Percentagem
A	1	2
B	46	92
C	3	6

Fonte: Autoria própria.

Nesta seção, investigamos de forma contrastiva dois advérbios terminados em "mente" - "especialmente" e "nomeadamente". As análises quantitativas revelaram que, em comparação com os textos produzidos por falantes nativos de português, no UMPLC, o advérbio "especialmente" apresenta uma frequência significativamente mais alta. Por outro lado, o "nomeadamente" é extremamente subutilizado.

Além das análises quantitativas, também realizamos análises qualitativas. Ao examinarmos o contexto de uso desses advérbios, constatamos que os aprendizes os utilizam de maneira diferente em comparação com os falantes nativos de português europeu. No UMPLC, o uso de "especialmente" é mais uniforme e, frequentemente, é colocado no início das frases, o que resulta da influência negativa do chinês, língua materna dos aprendizes.

A comparação entre o UMPLC e o PT sugere que os aprendizes provavelmente não estão familiarizados com o uso de "nomeadamente". No UMPLC, esse advérbio não apenas é pouco utilizado, mas também é empregado de maneira diferente em relação ao seu uso pelos falantes nativos de português europeu.

5 Discussões finais

O UMPLC é o primeiro *corpus* exclusivamente focado em falantes chineses de português L2, composto por dados longitudinais gerados em um contexto formal de ensino superior. Todos os dados são anotados com informações linguísticas, divididas em duas estruturas: *PoS* e lema. Atualmente, a pesquisa está na fase final de construção da plataforma na qual o corpus e ferramentas de pesquisa *online* serão disponibilizados.

Neste texto, foram apresentados os resultados de dois estudos preliminares relacionados ao uso lexical de aprendizes chineses de português L2, com base no UMPLC.

O primeiro estudo abordou a variação lexical, que se refere à diversidade de palavras e expressões utilizadas pelos aprendizes. Essa variação reflete não apenas o nível de proficiência, mas também as influências culturais e linguísticas enfrentadas pelos estudantes ao aprender uma língua estrangeira. Os resultados revelam que os aprendizes chineses começaram a apresentar um aumento notável na variação lexical durante o segundo semestre do primeiro ano. Essa descoberta auxiliará os professores na compreensão do desenvolvimento da proficiência linguística dos aprendizes e na expansão de seu vocabulário, levando em consideração essa característica.

Isso ressalta a importância, a partir desta fase de ensino-aprendizagem, de os professores incentivarem ainda mais a leitura e a escrita entre os aprendizes chineses, expondo-os a uma ampla gama de palavras e expressões em português. Além disso, podem motivá-los a escrever mais, praticando o uso correto do vocabulário.

Além disso, observou-se que os tópicos de composição estão intimamente ligados à variação lexical dos aprendizes. Portanto, ao avaliar a competência na produção escrita dos estudantes por meio de suas composições, é essencial definir os tópicos de maneira ampla e genérica para melhor compreender seu domínio vocabular.

A pesquisa revela também que os aprendizes chineses de português L2 não atingem o mesmo nível de variação lexical observado em falantes nativos. Conscientes dessa diferença, os professores podem adaptar seus materiais didáticos, direcionando-se para as áreas onde os estudantes apresentam lacunas no vocabulário. Além disso, podem desenvolver atividades e tarefas específicas para aprimorar a proficiência vocabular dos aprendizes, ao mesmo tempo em que sensibilizam os alunos para as nuances linguísticas presentes em contextos sociais e culturais.

Assim, os resultados da pesquisa sobre a variação lexical no contexto do UMPLC fornecem *insights* valiosos para a avaliação e aprimoramento do ensino de português L2. Os professores podem empregar essas descobertas para adaptar seu ensino de forma personalizada e auxiliar os estudantes a alcançarem um patamar superior de proficiência linguística e competência comunicativa.

Além disso, os resultados do segundo estudo demonstram a transferência negativa da língua materna e o fenômeno de sobreutilização e subutilização no uso de advérbios em português por parte dos aprendizes chineses. Essa análise é valiosa porque identifica padrões de uso inadequado em comparação com os falantes nativos, o que pode oferecer *insights* importantes para o ensino e aprendizagem eficazes. A compreensão de como os aprendizes utilizam advérbios, suas tendências de sobreutilização e subutilização, bem como a comparação com o uso nativo da língua, capacita os professores a desenvolverem estratégias pedagógicas mais direcionadas e a adaptar os materiais didáticos de maneira a atender às necessidades específicas dos aprendizes.

Os exemplos das aplicações pedagógicas demonstram que o UMPLC é de extrema relevância, pois constitui um recurso essencial para uma compreensão mais aprofundada da interlíngua dos aprendizes chineses, servindo como uma base empírica sólida para o desenvolvimento do ensino de português L2.

Por meio desta pesquisa, adquirimos uma compreensão mais sólida da importância da tecnologia, mais especificamente, da tecnologia relacionada a *corpora*, no contexto do ensino de L2. O UMPLC, cuja construção está em fase final, em breve estará disponível para pessoas registradas interessadas em utilizá-lo para fins de pesquisa. Estamos confiantes de que o UMPLC se tornará uma ferramenta amplamente utilizada no campo do ensino e da pesquisa em português L2.

6 Financiamento

O presente estudo foi desenvolvido no âmbito de MYRG2020-00139-FAH e de SRG2020-00021-FAH, projetos financiados pela Universidade de Macau.

Referências

COBB, Tom. Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies. *The Canadian Modern Language Review*, v. 59, n. 3, p. 393–424, 2003. DOI: 10.3138/cmlr.59.3.393. eprint: https://doi.org/10.3138/cmlr.59.3.393. Disponível em: https://doi.org/10.3138/cmlr.59.3.393.

DAVIES, Mark; PRETO-BAY, Ana Maria. *A frequency dictionary of Portuguese*. [*S. l.*]: Routledge, 2008.

GARSIDE, Roger; LEECH, Geoffrey; MCENERY, Tony. *Corpus annotation: linguistic information from computer text corpora*. [*S. l.*]: Routledge, 1997.

GRANGER, Sylviane. The computer learner corpus: a versatile new source of data for SLA research. *In:* GRANGER, Sylviane (ed.). *Learner English on Computer*. [*S. l.*]: Longman, 1998. p. 3–18.

GRANGER, Sylviane. A bird's-eye view of learner corpus research. *In:* GRANGER, Sylviane; HUNG, Joseph; PETCH-TYSON, Stephanie (ed.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. [*S. l.*]: Benjamins, 2002. p. 3–33.

GRANGER, Sylviane. Computer Learner Corpus Research: Current Status and Future Prospects. *Applied Corpus Linguistics*, Brill, p. 123–145, 2004.

GRANGER, Sylviane; GILQUIN, Gaëtanelle; MEUNIER, Fanny. Introduction: learner corpus research – past, present and future. *In: The Cambridge Handbook of Learner Corpus Research*. Edição: Sylviane Granger, Gaëtanelle Gilquin e Fanny Meunier. [*S. l.*]: Cambridge University Press, 2015. p. 1–6. (Cambridge Handbooks in Language and Linguistics). DOI: 10.1017/CBO9781139649414.001.

GRANGER, Sylviane; TRIBBLE, Christopher. Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. *In:* GRANGER, Sylviane (ed.). [*S. l.*]: Addison Wesley Longman, 1998. p. 199–209.

GROSSO, Maria José; ZHANG, Jing; GASPAR, Catarina; TEIXEIRA, Madalena. *Referencial Ensino de Português Língua Estrangeira na China*. [*S. l.*]: Centro Científico e Cultural de Macau Universidade de Macau, 2021.

KREYER, Rolf. 'Multilinguality'in learner corpora: The case of the MILE. *In:* NURMI, Arja; RÜTTEN, Tanja; PAHTA, Päivi (ed.). *Challenging the Myth of Monolingual Corpora*. [*S. l.*]: Brill, 2017. p. 200–219.

KÜBLER, Sandra; ZINSMEISTER, Heike. *Corpus linguistics and linguistically annotated corpora*. [*S. l.*]: Bloomsbury Publishing, 2015.

MARNEFFE, Marie-Catherine de; MANNING, Christopher D.; NIVRE, Joakim; ZEMAN, Daniel. Universal Dependencies. *Computational Linguistics*, MIT Press, Cambridge, MA, v. 47, n. 2, p. 255–308, jun. 2021. DOI: 10.1162/coli_a_00402. Disponível em: https://aclanthology.org/2021.cl-2.11.

MCENERY, Tony; HARDIE, Andrew. *Corpus linguistics: Method, theory and practice*. [*S. l.*]: Cambridge University Press, 2011.

NESSELHAUF, Nadja. Learner corpora and their potential for language teaching. *How to use corpora in language teaching*, v. 12, p. 125–156, 2004.

PAIVA, Valeria de; REAL, Livy. Universal POS tagging for Portuguese: Issues and Opportunities. *Proceedings of LexSem+ Logics 2016*, p. 25, 2016.

QI, Peng; ZHANG, Yuhao; ZHANG, Yuhui; BOLTON, Jason; MANNING, Christopher D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *In:* PROCEEDINGS of the 58th Annual

Meeting of the Association for Computational Linguistics: System Demonstrations. [*S. l.: s. n.*], 2020. Disponível em: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

RADEMAKER, Alexandre; CHALUB, Fabricio; REAL, Livy; FREITAS, Cláudia; BICK, Eckhard; DE PAIVA, Valeria. Universal dependencies for Portuguese. *In:* PROCEEDINGS of the fourth international conference on dependency linguistics (Depling 2017). [*S. l.: s. n.*], 2017. p. 197–206.

RUNDELL, Michael. The corpus of the future, and the future of the corpus. *In:* TALK at a special conference on New Trends in Reference Science at Exeter, UK (a printed hand out). [*S. l.: s. n.*], 1996.

SANTOS, Isabel Almeida; PEREIRA, Isabel; MARTINS, Cristina; LOPES, Ana Cristina Macário; CARAPINHA, Conceição; SILVA, António. Corpus oral de PL2: um novo recurso para a investigação e ensino. *Revista da Associação Portuguesa de Linguística*, n. 1, p. 740–760, 2016.

SELINKER, Larry. *Rediscovering interlanguage*. [*S. l.*]: Addison Wesley Longman, 1992.

TURTON, Nigel D; HEATON, John Brian. *Longman dictionary of common errors*. [*S. l.*]: Longman, 1996.

WOLFE-QUINTERO, Kathryn Elizabeth; INAGAKI, Shunji; KIM, Hae-Young. *Second language development in writing: Measures of fluency, accuracy, & complexity*. [*S. l.*]: Second Language Teaching an Curriculum Center of University of Hawai'i, 1998.

YAN, Qiaorong. O desenvolvimento do ensino de Português na China: história, situação atual e novas tendências. *In:* YAN, Qiaorong; FLEIDE, Daniel Albuquerque (ed.). *O ensino do Português na China: parâmetros e perspectivas*. [*S. l.*]: Edufrn, 2019. p. 24–52.

YANG, Huizhong. *An Introduction to Corpus Linguistics*. [*S. l.*]: Shanghai Foreign Language Education Press, 2001.

Contribuições dos autores

Jing Zhang: Conceituação, Administração de projetos, Investigação, Metodologia, Escrita – rascunho original, Escrita – revisão e edição; **Mu You**: Curadoria de dados, Análise formal, Metodologia, Escrita – rascunho original, Escrita – revisão e edição.