Compilation and tagging of a corpus with Celpe-Bras texts

**Juliana Schoffen, Elisa Stumpf, Deise Amaral, Luiza Divino,
Isadora Hanauer, Isabel Lisboa, Amanda Raupp, Brenda Xavier**

Federal University of Rio Grande do Sul
AVALIA Research Group
julianaschoffen@gmail.com

Abstract

This paper presents the compilation and tagging processes of a corpus of written texts produced by test takers of the Celpe-Bras exam - the official Brazilian proficiency exam in Portuguese as an Additional Language (PAL). In order to identify language use patterns that distinguish the different proficiency levels, the main purpose of this corpus is to enable a wider range of quantitative and qualitative analyses. The data consists of approximately 15,000 texts written in four editions of Celpe-Bras, which are in the process of being digitized, de-identified and tagged. According to the guidelines for the typing and proofreading stages, the texts must be typed following the original handwriting and excluding any information that could identify the test taker. The tagging protocol established by the research team includes spelling normalizations to allow the use of automatic analyses besides signaling typical features of the genres required in the exam. Upon completion and availability of this corpus, further analyses will allow for more refined descriptions of each certified proficiency level, enhancing the validation process of Celpe-Bras.

1 Introduction

This article aims to present the process of compiling and tagging the corpus of texts written under exam conditions for the Celpe-Bras exam (Certificate of Proficiency in Portuguese for Foreigners)[1], compiled by the Avalia research group at the Federal University of Rio Grande do Sul (Brazil). Celpe-Bras is the official Brazilian proficiency exam in Portuguese as an Additional Language (PAL). It is currently administered in over 130 accredited test centers since 1998, with around 5,000 test takers in each biannual edition.

Despite the considerable amount of studies already published about the exam, the lack of a more representative corpus of test takers' scripts has limited studies with empirical data, mainly quantitative studies. This limitation has prevented the use of automated methods to describe language usage patterns in each proficiency level, more specifically Corpus Linguistics (CL) tools, which have been used consistently in the field of proficiency assessment in the last decades. Therefore, the compilation and tagging of the current corpus offers new possibilities for research in the field of PAL proficiency assessment.

2 Literature review

Corpus linguistics tools and methodology enable the analysis of features and patterns of language use in texts produced in different proficiency levels, fostering its use in studies attempting to validate exams and refine the description of performance in different proficiency levels (Cushing, 2017, 2021; Gablasova, 2020; Gablasova et al., 2017; Taylor and Barker, 2008) (Taylor and Barker, 2008). Analyses of corpus data can hence "inform decisions about assessment criteria and the development of rating scales" (Taylor and Barker, 2008, p. 246).

Many studies have used Corpus Linguistics tools to describe the language used by test takers in large-scale exams of English[2]. Concerning Portuguese, there are several corpora focusing on the study of language learning, such as the project "Recolha de Dados de Aprendizagem do Português como Língua Estrangeira"[3]; the "Corpus de Aquisição de L2 (CAL2)"[4]; the "Corpus de Produções Escritas de Aprendentes de PL2 (PEAPL2)"[5]; the "Cor-

[1] More information about the exam can be found at Acervo Celpe-Bras.

[2] See Banerjee et al. (2007), Kennedy et al. (2007) Barkaoui (2016) Read and Nation (2002) about IELTS and Cumming et al. (2005), Biber and Gray (2013) and Biber et al. (2004) for TOEFL

[3] Retrieved from: https://tinyurl.com/yrjc6r2v on November 03 2023.

[4] Retrieved from: https://tinyurl.com/bde3m582 on November 03 2023.

[5] Retrieved from: https://tinyurl.com/cms69k3u on Novem-

pus de Português como Língua Estrangeira/Língua Segunda (COPLE2)" (Antunes et al., 2016), including texts by learners as well as by candidates in the proficiency exam of the Portuguese as a Foreign Language Assessment Centre (CAPLE); and the "Corpus Produção Oral em Provas de Português L2 (POPL2)" (Ferreira et al., 2023).

Regarding Celpe-Bras, before the compilation of the corpus described in the following sections, there is only one study that automatically analyzes Celpe-Bras texts (Evers, 2013), but it used only 181 texts to try to identify lexical and cohesive elements that differentiated the levels of texts written by test takers.

3 Data

To expand the possibilities for studies on the exam, this paper reports the ongoing compilation of a corpus of around 15,000 texts produced and assessed in four editions of Celpe-Bras (2015-2, 2016-1, 2016-2 and 2017-1), with up to 200 texts assessed in each grade (0-5, being 0 the lowest and 5 the highest score) for each task (four per edition), estimated to total around 3 million words. This sample was obtained from approximately 70,000 texts in the form of digitized and already de-identified copies, which undergo typing, proofreading and tagging processes.

To compile the corpus, we initially selected texts that received the same score from two different raters, without requiring a third rater to assign a score. Whenever the number of texts was greater than 200, the texts were randomly selected. When we had fewer than 200 texts with two agreeing scores, the number was completed with texts that had been re-assessed, using randomization for the selection. The final corpus is shown in Table 1. Each column displays the number of texts compiled in each grade per task by edition[6]. The rightmost column shows the total number of texts compiled per task and per edition.

The organization of the corpus into different subcorpora takes into account the task and edition of the exam, as well as the grade given to each text, allowing comparisons between all the metadata.

ber 03 2023.

[6]As can be seen in Table 1, in some grades, there are fewer than 200 texts. In these grades, all available texts have been compiled.

4 Metadata

The Celpe-Bras exam consists of a written part and an oral part and certifies, with a single test, four proficiency levels: Upper Advanced, Advanced, Upper Intermediate and Intermediate[7]. The written part of Celpe-Bras is made up of four integrated listening, reading and writing tasks, in which test takers have to produce texts of different discourse genres and purposes.

Since the texts were received without identification, the corpus does not have metadata about the test takers who produced them. There is, however, metadata relating to the tasks that generated these texts and the score assigned. Based on the description by Schoffen et al. (2018), it is possible to identify the task's input material (audio, video or written text), its theme, the sphere of activity in which the requested text is inserted, the purpose(s), the interlocutors, the discourse genre and the medium in which this text would be published[8]. Table 2 shows the expected genre for each task response in each exam edition[9].

As well as the scores assigned for each text, there is information about the scores received by the test taker in each of the other tasks in the edition, the score they received in the oral part of the exam and also their certification level.

5 Data preparation and corpus tagging

The typing process follows guidelines that respect the original writing of the text and excludes any marks that might identify the test takers. This stage is followed by a proofreading process, which aims to ensure that the texts are true to the original. Finally, tagging is done manually in order to standardize the spelling and make it possible to use automated CL tools to describe patterns of language use in the texts of test takers at different levels of proficiency.

The tagging protocol presented in this paper was developed in line with Celpe-Bras' proficiency construct (INEP, 2020) and is based on systems found in the literature (Bick, 2000; Eickhoff, 2023; Granger et al., 2022). We present here the main categories established by the research team, based on a pilot study that tagged and analyzed around

[7]There is no certification below Intermediate level.

[8]All the metadata related to the tasks are available on the research group website.

[9]For a complete description of the tasks in a searchable database, check Grupo Avalia. For a comprehensive analysis of the data, refer to Schoffen et al. (2018)

Table 1: Number of texts per task and per edition

Score/Edition		0	1	2	3	4	5	Total
2015-2	**T1**	5	128	200	200	200	193	926
	T2	82	200	200	200	200	200	1082
	T3	33	189	200	200	200	138	960
	T4	28	200	200	200	200	200	1028
								3.996
2016-1	**T1**	15	200	200	200	200	200	1015
	T2	48	200	200	200	200	200	1048
	T3	44	200	200	200	200	151	995
	T4	93	200	200	200	200	134	1027
								4.085
2016-2	**T1**	21	188	200	200	200	159	968
	T2	22	130	200	200	200	73	825
	T3	30	143	200	200	200	200	973
	T4	43	200	200	200	200	95	938
								3.704
2017-1	**T1**	4	54	156	200	200	200	814
	T2	19	119	200	200	200	200	938
	T3	7	73	200	200	200	135	815
	T4	8	155	200	200	200	200	963
								3.530

50 texts together[10]. The protocol establishes rules for tagging spelling differences so that the same word written with different spellings can be recognized by automated tools and subsequently analyzed, avoiding distortion in the results (Hanauer, 2022).

While many similar corpora employ a system to classify errors across different linguistic levels, our goal was to simply make the texts readable by automatic tools, instead of editing the texts and rewriting them. For the moment, the protocol covers aspects related to lexical and structural features of the texts. POS tagging may be done in future studies. The tagging was done using VBA (Visual Basic for Applications) in Microsoft Word, following Hardie (2014)'s suggestions for using a "Modest XML"[11]. Initially, all the texts are tagged with the identification of the file name [1], followed by the year and edition of the exam, the task, the identification number of the test taker and the score awarded to the text. Each paragraph in the text is also tagged. The spelling normalization is guided by excerpts marked as incorrect by text processors such as Microsoft Office Word or Google Docs [2][12]. Another tag is used to signal words that are incorrectly written as two (or more) separate words, so that it does not interfere with the number of types and tokens of a text, as in example [3].

Considering the high recurrence of discourse genres such as emails and letters in Celpe-Bras and the importance of using certain linguistic resources,

[10]The protocol has not yet been put into practice. We present here a preview of the research group's conclusions based on the literature review and the pilot study mentioned.

[11]By "modest XML", Hardie (2014) refers to a lightweight approach to XML markup that can be implemented by users with little technical expertise and covers most of the needs of corpus linguists. While not standard, using word processing tools (e.g. Microsoft Word) for XML tags makes the files more easily accessible for the research team and allows them to be saved as plain texts while keeping the tags, for future use in other tools.

[12]Since future studies based on this corpus may want to focus on the different forms used to write the same word, the original forms are kept inside the tag.

Table 2: Target genres in each task

	Target genres			
	2015-2	2016-1	2016-2	2017-1
Task 1	section of a guide	personal account	news article	news article
Task 2	news article	letter/e-mail	letter/e-mail	letter/e-mail
Task 3	letter/e-mail	report	article	letter/e-mail
Task 4	open letter	opinion article	letter to the editor	letter to the editor

some of which are relatively standardized, to ensure adequacy for the proposed genre, the protocol marks aspects such as the heading, indication of date and place, title, addressing, greeting and closing. The heading [4] can include an indication of the date and place [5] when it comes to letters, or, more frequently, when it comes to emails, an indication of the subject, sender [15][16][17] and recipient [7][8][13]. The title tag [6] applies to cases where the candidate gives their text a title. As for addressee, we consider any form that shows to whom the text is addressed, which ranges from proper nouns, as in [7], to common nouns in the plural, such as names of groups, companies and institutions, as it can be seen in [8]. Greetings are subdivided into two forms: one that does not include a vocative or an addressee, as in [9] or [10], and another that includes these items, as in [11] and [12]. Closing comprises not only typical farewell forms such as [13], but also passages that signal the author's intention to end their text, as in [14]. As for excerpts that could identify the examinee in the text, there are three different forms of tagging: a) for signature with a proper name or occupation at the end of the text [15]; b) for identification with name or occupation in the middle of the text or in the header [16]; and c) with identification without a proper name in the header or at the end of the text [17].

[1] <texto id='20152t4p3n1'> </texto>
[2] <norm orig='presado'> prezado </norm>
[3] <cn alt='portanto'> por tanto </cn>
[4] <cab> Assunto: Gostaria de Patrocinar o projeto "Favela Orgânica" </cab>
[5] <datloc> Arequipa, 23 de Maio 2017 </datloc>
[6] <tit> Titulo: Projeto Favela orgânica </tit>
[7] <end>Luiza,</end>
[8] Para: <end>Empresas patrocinadoras</end>
[9] <saud>Bom dia</saud>
[10] <saud>Prezados</saud>
[11] <saudend>Prezado don da empresa</saudend>
[12] <saudend>Prezados Senhores: Bom dia</saudend>
[13] <fech>Atenciosamente</fech>
[14] <fech>Qualquer questão não hesitem em contatar-me, estou a disposição dos vossos Exmºs, a qualquer hora. Meus melhores comprimentos.</fech>
[15] Atenciosamente, <IDass> Luan Santana </IDass>
[16] Boa tarde, sou <IDid> Carlos Silva </IDid>
[17] De: <IDsn> gerente de recursos humanos </IDsn>

[13]Fake names were used for illustrative purposes only.

6 Preliminary studies

Preliminary versions of this corpus have already been used in some studies. In an attempt to distinguish the Upper Intermediate and Upper Advanced levels, Kunrath (2019) analyzed, with the help of the Coh-Metrix software, the recontextualization of information and the use of linguistic-discursive resources in 50 texts and proposed a progression of levels based on these aspects. Divino (2021), Hanauer (2023) and Sostruznik (2023) used a version of the corpus without annotations and aimed to list relevant lexical indices of analysis for the characterization of the Intermediate and Upper Advanced levels in different Celpe-Bras tasks. The analyses, carried out with Sketch Engine (Kilgarriff et al., 2014) and the Log-Likelihood (LL) statistical significance test (Rayson, 2003), indicated greater length in Upper Advanced texts. They also corroborated qualitative analyses carried out previously, showing a greater incidence of structures characteristic of the target genre (Mendel, 2019) and terms more suited to the proposed interlocutors' relationship (Sirianni, 2016).

7 Final remarks

Considering that this is the first Brazilian corpus of texts graded according to the proficiency levels certified by Celpe-Bras, this corpus - when finalized and available - will enable analyses that contribute to the validation process of the exam, fostering the development of more robust descriptions for each of the certified proficiency levels and also making it possible to further detail the evaluation parameters of the texts. These results could also help PAL teachers, allowing them to design teaching materials and develop appropriate teaching tasks for the specific needs of each level. The protocol developed also has the potential to support the compilation of other corpora with similar characteristics, such as learner corpora and corpora of texts with spelling differences, since it will allow the texts to be analyzed in CL and natural language processing programs and tools, using the normalized and tagged version, and accessing the original characteristics of the text, such as the different spelling possibilities of each word.

References

Sandra Antunes, Amália Mendes, Anabela Gonçalves, Maarten Janssen, Nélia Alexandre, António Avelar, Adelina Castelo, Inês Duarte, Maria João Freitas, José Pascoal, et al. 2016. Apresentação do corpus de português língua estrangeira/língua segunda–cople2. *Revista da Associação Portuguesa de Linguística*, (1):85–103.

Jayanti Banerjee, Florencia Franceschina, and Anne Margaret Smith. 2007. Documenting features of written language production typical at different ielts band score levels. *IELTS Research Reports*, 7(5):1–69.

Khaled Barkaoui. 2016. What changes and what doesn't? an examination of changes in the linguistic characteristics of ielts repeaters' writing task 2 scripts. Technical report, IELTS Research Reports Online Series.

Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay, and Alfredo Urzua. 2004. *Representing language use in the university: Analysis of the TOEFFL 2000 spoken and written academic language corpus*. Test of English as a Foreign Language.

Douglas Biber and Bethany Gray. 2013. Discourse characteristics of writing and speaking task types on the toefl ibt® test: a lexico-grammatical analysis. *ETS Research Report Series*, 2013(1):i–128.

Eckhard Bick. 2000. *The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework*. Aarhus Universitetsforlag.

Alister Cumming, Robert Kantor, Kyoko Baba, Usman Erdosy, Keanre Eouanzoui, and Mark James. 2005. Differences in written discourse in independent and integrated prototype tasks for next generation toefl. *Assessing Writing*, 10(1):5–43.

Sara T. Cushing. 2017. Corpus linguistics in language testing research. *Language Testing*, 34(4):441–449.

Sara T. Cushing. 2021. Corpus linguistics and language testing. In *The Routledge Handbook of Language Testing*, pages 545–560. Routledge.

Luiza Divino. 2021. Índices lexicais de análise para a caracterização dos níveis intermediário e avançado superior no exame celpe-bras: uma pesquisa guiada por corpus. Unpublished undergraduate thesis.

Seda Acıkara Eickhoff. 2023. Ptc error correction protocol [unpublished manuscript]. Unpublished Manuscript.

Aline Evers. 2013. Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame celpe-bras. Unpublished masters thesis.

Tânia Ferreira, Isabel Santos, Conceição Carapinha, Cristina Martins, Isabel Pereira, Graça Rio-Torto, Liliana Inverno, Rui Pereira, Carla Ferreira, Sara Sousa, et al. 2023. Construção do corpus" produção oral em provas de português l2"(popl2). *Études romanes de Brno*, 44(1):245–261.

Dana Gablasova. 2020. Corpora for second language assessments. In *The Routledge handbook of second language acquisition and language testing*, pages 45–53. Routledge.

Dana Gablasova, Vaclav Brezina, and Tony McEnery. 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 67(S1):130–154.

Sylviane Granger, Helen Swallow, and Jennifer Thewissen. 2022. The louvain error tagging manual. version 2.0.

Isadora Hanauer. 2022. Influência das inadequações ortográficas em análise de tarefa escrita do celpe-bras guiada por corpus [conference presentation abstract].

Isadora Hanauer. 2023. Caracterização dos níveis intermediário e avançado superior do exame celpe-bras em produções escritas de examinandos no gênero carta/e-mail: contribuições de uma análise guiada por corpus. Unpublished undergraduate thesis.

Andrew Hardie. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal*, 38(1):73–103.

INEP. 2020. *Documento base do exame Celpe-Bras*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

Christopher Kennedy, Dilys Thorp, L Taylor, and P Falvey. 2007. A corpus-based investigation of linguistic responses to an ielts academic writing task. In *IELTS collected papers: Research in speaking and writing assessment. Studies in language testing*.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Simone Paula Kunrath. 2019. Os descritores gerais e a progressão dos níveis de proficiência do exame celpe-bras. Unpublished doctoral dissertation.

Kaiane Mendel. 2019. Proficiência e autoria na avaliação integrada de leitura e escrita do exame celpe-bras. Unpublished masters thesis.

Paul Edward Rayson. 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Lancaster University (United Kingdom).

John Read and Paul Nation. 2002. An investigation of the lexical dimension of the ielts speaking test. Technical report, IELTS Research Reports.

Juliana Schoffen, Margarete Schlatter, Simone Paula Kunrath, Ellen Yurika Nagasawa, Gabrielle Rodrigues Sirianni, Kaiane Mendel, Luana Ramos Truyllio, and Luiza Sarmento Divino. 2018. Estudo descritivo das tarefas da parte escrita do exame celpe-bras: Edições de 1998 a 2017. Technical report, Porto Alegre.

Gabrielle Rodrigues Sirianni. 2016. Descrição dos níveis de proficiência em tarefa de leitura e escrita a partir de produções textuais de alunos do curso preparatório celpe-bras. Unpublished undergraduate thesis.

Júlia Sostruznik. 2023. O uso de conjunções em produções escritas no exame celpe-bras: um estudo baseado em corpus. Unpublished undergraduate thesis.

Lynda Taylor and Fiona Barker. 2008. Using corpora for language assessment. *Encyclopedia of language and education*, 7:241–254.