Corpus-driven lexical analyses of CorCel: a comparative analysis of preliminary findings of written proficiency in Portuguese as an additional language

Elisa Marchioro Stumpf[1], **Juliana Schoffen**[1], **Luiza Sarmento Divino**[1], **Isadora Dahmer Hanauer**[1], **Amanda Raupp**[1], **Brenda Xavier**[1]

[1] AVALIA research group
Universidade Federal do Rio Grande do Sul
Av. Bento Gonçalves, 9500 - CEP 91501-970
Caixa Postal 15002 - Porto Alegre/RS - Brazil

elisa.stumpf@ufrgs.br

Abstract. *This paper aims to comparatively analyze research on CorCel [Schoffen et al. , forthcoming] - a corpus of written texts produced under exam conditions for the Celpe-Bras exam. It compares studies that examined tasks using Sketch Engine's keywords, wordlist, n-grams and concordance tools. These studies also analyzed text length and lexical richness using lexical diversity indices. The comparative study showed relevant lexical indexes for characterizing the exam proficiency levels, such as text length, and important differences in the use of input material and other linguistic resources among texts rated with different grades. Such analyses offer new possibilities for research in Portuguese as an additional language (PAL) proficiency assessment and teaching.*

1. Introduction

This paper presents preliminary findings from corpus-driven studies analyzing CorCel - a corpus of written texts produced under exam conditions for the Celpe-Bras exam (Certificate of Proficiency in Portuguese for Foreigners), compiled by the AVALIA research group [1] at the Federal University of Rio Grande do Sul [Schoffen et al. , forthcoming]. As the first of its kind, this corpus has been fostering different quantitative and qualitative analyses to identify patterns of language use that distinguish the different proficiency levels. The corpus consists of over 15,000 texts written by examinees in four editions of Celpe-Bras and rated in 6 different levels, which are being digitized, revised and tagged, with future goals of reaching a corpus with 70,000 texts.

Celpe-Bras is the official Brazilian proficiency exam in Portuguese as an Additional Language (PAL) developed and administered by the Brazilian Ministry of Education since 1998, with more than 130 accredited test centers worldwide and over 10,000 examinees each year. The lack of a representative sample of test-takers' scripts had until recently prevented the development of quantitative studies, limiting large-scale description of language usage in each proficiency level with automated tools, such as Corpus Linguistics (CL) ones. These tools have been used consistently in the field of proficiency assessment in the last decades [Cushing 2021]; [Cushing 2017]; [Callies and Götz 2015]

[1]Research group website: https://www.ufrgs.br/grupoavalia/

for other languages, particularly with data collected under exam conditions, as it can show important features of proficiency levels [Wisniewski 2017]; [Banerjee et al. 2007]; [Biber and Gray 2013]; [Paquot 2019]. Analyses of CorCel data can not only provide insight into recurring language features across proficiency levels but also enhance PAL teaching by supporting teachers in designing materials and activities that take into account the particularities of each proficiency level. In addition, the analyses can assist candidates preparing for the exam by providing examples that illustrate the linguistic features and the communicative strategies valued in Celpe-Bras' communicative tasks for each proficiency level. Therefore, the current corpus offers new possibilities for research in the field of PAL proficiency assessment by grounding the assessment framework in empirical evidence from authentic test-takers' performances.

The paper starts by briefly describing the processes of corpus compilation, typing and tagging carried out so far. It then presents and discusses early results of lexical analyses conducted to date, aiming to discriminate texts with different grades in six tasks, and closes by referring to future works CorCel will enable.

2. CorCel

CorCel corpus compiles data from four tasks of four different editions of the Celpe-Bras exam, with approximately five thousand examinees each. From the information provided, it is not possible to identify the examinees, the Test Center where the exam was taken, or the raters [2]. Because of that, there is also no information regarding the examinees' linguistic and educational background, which prevents us from categorizing it as a "learner corpus", a common practice in the field of corpus linguistics. Thus, we opted to classify it as an L2 corpus, considering that the texts were written by candidates who had Portuguese as an additional language when taking the test.

2.1. Celpe-Bras Exam

Celpe-Bras is the Brazilian Portuguese proficiency exam. According to the exam's theoretical construct, being proficient in a given language "means being able to engage in different situations of Portuguese language use in the world, showing adequacy to the demands of several contexts"[3] [INEP 2020, p. 8̃]. Through a single test, Celpe-Bras certifies four proficiency levels: Upper Advanced, Advanced, Upper Intermediate, and Intermediate [4].

The exam consists of a twenty-minute oral interaction and a three-hour written part, comprising four integrated listening-into-writing and reading-into-writing tasks. The written part of the exam, which is where the texts compiled in CorCel come from, presents an audio, video or written input and a prompt that asks the examinees to produce a text accomplishing the task considering the information from the input material and features such as genre, medium, purpose and audience awareness [5].

[2]Texts were received by the AVALIA research group from the National Institute of Educational Studies and Research Anísio Teixeira (Inep), the agency of the Ministry of Education (MEC) responsible for large-scale educational assessment in Brazil. Texts can be used for research purposes only, and will not be made available publicly in their entirety.

[3]Originally: "implica ser capaz de engajar-se em diferentes situações de uso da língua portuguesa no mundo, mostrando adequação às demandas dos vários contextos".

[4]There is no certification below Intermediate level.

[5]More information about Celpe-Bras, including the previous tests and a list of studies about the exam is

Table 1. Number of texts per task and per edition - adapted from [Schoffen et al. 2024]

Score/ Edition/Requested Genre	0	1	2	3	4	5	Total
2015-2							**3,996**
T1 - section of a guide	59	128	200	200	200	193	926
T2 - news article	82	200	200	200	200	200	1082
T3 - letter/e-mail	33	189	200	200	200	138	960
T4 - open letter	28	200	200	200	200	200	1028
2016-1							**4,085**
T1 - personal account	15	200	200	200	200	200	1015
T2 - letter/e-mail	48	200	200	200	200	200	1048
T3 - report	44	200	200	200	200	151	995
T4 - opinion article	93	200	200	200	200	134	1027
2016-2							**3,704**
T1 - news article	21	188	200	200	200	159	968
T2 - letter/e-mail	22	130	200	200	200	73	825
T3 - article	30	143	200	200	200	200	973
T4 - letter to the editor	43	200	200	200	200	95	938
2017-1							**3,530**
T1 - news article	4	54	156	200	200	200	814
T2 - letter/e-mail	19	119	200	200	200	200	938
T3 - letter/e-mail	7	73	200	200	200	135	815
T4 - letter to the editor	8	155	200	200	200	200	963

2.2. CorCel data

CorCel has compiled 15,315 texts (around 3 million words) produced in response to Celpe-Bras written tasks in four editions (2015-2, 2016-1, 2016-2 and 2017-1) [6], divided by task (each edition comprises four tasks) and by grade (each text is rated 0, 1, 2, 3, 4 or 5). The texts were made available to the research group in the form of a digitized manuscript copy. The texts underwent typing and proofreading processes following guidelines developed by the group [Schoffen et al. 2024], [Schoffen et al.]. An important orientation in the typing guidelines, for example, recommends that no grammatical or spelling errors should be corrected, keeping the text exactly as it was originally written. The current corpus is presented in Table 1. Each column displays the number of texts compiled in each grade per task by edition. The rightmost column shows the total number of texts compiled per task and edition.

3. Results and discussion

Early analyses of this corpus have employed both qualitative and quantitative methods. [Kunrath 2019] used Coh-Metrix software to distinguish texts graded 3 and 5 from tasks 1 and 4 of the 2016-1 edition, demonstrating that they differ in the recontextualization of

available at www.ufrgs.br/acervocelpebras/pesquisas.

[6]These tasks, along with all other administered Celpe-Bras exams, can be found at https://www.ufrgs.br/acervocelpebras/acervo/.

Table 2. Summary of results

Editions	Tasks	Score	N_texts	Types	Tokens	Token average	TTR
2015-2	T3	5	100	2,770	23,121	231.2	11.98%
		2	100	2,767	16,507	165.1	16.76%
	T4	5	237	5,663	47,616	200.9	11.91%
		4	477	7,247	88,235	185.0	8.21%
		3	715	8,838	123,033	172.1	7.19%
		2	628	8,537	99,457	158.4	8.58%
		1	211	4,004	30,290	143.6	13.35%
		0	25	826	2,284	91.4	-
2016-2	T3	5	200	4,823	45,848	229.2	10.51%
		2	200	4,083	31,479	157.4	12.97%
	T4	5	50	2,641	12,222	244.4	21.60%
		2	50	1,845	7,891	157.8	23.38%
2017-1	T3	5	100	2,876	23,748	237.5	12.11%
		2	100	2,660	16,593	165.9	16.03%
	T4	5	50	2,626	11,650	233.0	22.54%
		2	50	2,023	8,376	167.5	24.15%

information from the input material and the use of linguistic resources. [Mendel 2019]'s qualitative analysis of texts from tasks 3 and 4 of the 2015-2 edition noted that higher scores corresponded with better use of background knowledge and adherence to genre, demonstrating a greater authorial voice. [Sirianni 2020] qualitatively examined texts rated 1 (uncertified) and 2 (intermediate) from the 2016-2 edition, finding significant differences between these ratings in understanding input material, genre adherence, and linguistic adequacy.

Quantitative corpus-driven studies have been carried out by [Divino 2024], [Divino 2021], [Hanauer 2023], [Sostruznik 2023], and [Raupp 2024], who analyzed linguistic features and lexical patterns in different tasks of CorCel, using Sketch Engine [Kilgariff et al. 2004], the Log-Likelihood (LL) statistical significance test [Rayson 2003], and type-token ratio (TTR) measures. From Sketch Engine, the tools used were: a) Keywords, which creates a list of words that are more used in the focus corpus than in a reference corpus (in these studies, Corpus Brasileiro and Portuguese Trends were used); b) Wordlist, which generates different types of frequency lists with frequency measures; c) Concordance, which retrieves linguistic units (lemmas, words, tags, phrases, etc) and show their immediate right and left contexts and d) N-grams, which generates frequency lists of sequences of tokens.

The value of TTR is calculated by dividing the number of types (the amount of distinct words on the texts) by the number of tokens (total amount of words on the texts) and multiplying this result by 100, to express the result in percentage (Sketch Engine, n.d.). This value expresses the lexical variety of the texts, which tends to be higher in texts written by more proficient writers, considering their lexical knowledge.

Table 2 presents the results of the subcorpora already analyzed quantitatively with corpus linguistics tools. The table shows the number of texts analyzed for each score and

each task, followed by the number of types and tokens, the Token average, and the TTR in each subcorpus. It summarizes the results of [Divino 2024] and [Divino 2021], who analyzed Task 4 from the 2015-2 edition of Celpe-Bras, using the total amount of texts of each grade; [Hanauer 2023], who focused on Tasks 3 from the 2015-2 and 2017-1 editions, using a subcorpora of 100 texts for each task and grade (2 and 5); [Sostruznik 2023], who used 50 texts graded 5 and 50 graded 2 of Tasks 4 from the 2016-2 and 2017-1 editions; and [Raupp 2024], who analyzed 200 texts graded 2 and 200 texts graded 5, of Task 3 from the 2016-2 edition of Celpe-Bras.

3.1. Text length

When it comes to text length, the studies indicate that grade 5 texts are, in general, considerably longer than grade 2 texts in all tasks. [Divino 2021]'s analysis of task 4 from the 2015-2 edition showed a higher average number of words and sentences per text in the corpus composed of grade 5 texts compared to grade 2 texts (10.8 in texts graded 5 and 8.8 in texts graded 2). A second study [Divino 2024] including grade 0, 1, 3, and 4 texts supported her previous findings, suggesting that text length is an important factor regarding proficiency, with more advanced texts generally being longer and containing a higher number of different words from those found in the input material, despite being morphologically similar.

In [Hanauer 2023] (2023), the two analyzed tasks presented very similar values of average text length, in number of words, for the same grades, with a higher number for grade 5 texts compared to grade 2 texts. The average number of sentences was also higher for texts graded 5 in both tasks analyzed (11.51 for grade 5 texts and 8.79 for grade 2 texts from task 3 of 2015-2; 12.56 for grade 5 texts and 9.79 for grade 2 texts from task 3 of 2017-1).

[Raupp 2024] also showed that texts rated 5 are, on average, longer than texts graded 2, with a higher number of types, tokens, and sentences, as well as a higher average number of sentences per text (11.51 in texts graded 5 and 6.93 in texts graded 2). This result also indicates that the higher the level of proficiency of the examinee, the potentially longer their written productions will be.

3.2. Type-Token Ratio (TTR)

The studies conducted so far showed results different from expected on the TTR analysis, as texts with lower grades had larger values of TTR. As stated before, the hypothesis was that texts of higher proficiency levels would have higher lexical diversity. Only in [Divino 2024] and [Divino 2021] was it possible to see a larger value of TTR in grade 5 texts compared to grades 4, 3, and 2. However, the value was still larger in grade 1. The other studies [Hanauer 2023], [Sostruznik 2023], [Raupp 2024] showed greater lexical variety in texts graded 2, compared to texts graded 5.

This result can be explained by a higher number of spelling inadequacies in texts with lower grades, which are identified by the software as different words, no matter how similar they are. Results like this show a limitation of having carried out these studies without normalizing spelling, as addressed by [Granger and Wynne 2000] and discussed in [Divino 2024]. It can also be explained by the difference in corpora length, according to the limitations of TTR referred to in Sketch Engine. This limitation can be evidenced

in the results of Soztrusnik (2023), who analyzed only 50 texts of Task 4 in 2016-2 and 2017-1 editions and found a significantly higher TTR than those in the other studies.

3.3. Use of input material

[Divino 2024]'s, [Divino 2021]'s, [Hanauer 2023]'s, and [Raupp 2024]'s data, agreeing with [Mendel 2019], dispute the idea that only lower-level writers copy excerpts from the input texts. According to [Divino 2024], at least 20% of the texts in the grade 5 subcorpus contain 6-word-length n-grams found in the input material. For instance, "the vitor maria da silva manor house" ("o palacete vitor maria da silva") appears in 22.78% of the compositions, and "more than 50% of the tiles" ("mais de 50% dos azulejos") in 21.94%. [Divino 2024]'s results also showed that the task can be accomplished in different ways, since important terms for fulfilling the task's purposes are not present in 100% of the texts of any grade.

[Hanauer 2023]'s, [Divino 2024]'s, and [Raupp 2024]'s analyses of keywords indicated that grade 5 texts used more words absent from the input text when compared to grade 2 texts. Nevertheless, even proficient writers use many words and excerpts from the input material, but there is still a noticeable relation between these texts and the input material. [Raupp 2024]'s analysis revealed that the keywords in the input material of Task 3 (2016-2 edition) are highly frequent. The high occurrence of these words in the reference corpus[7], suggests that their familiarity might facilitate the paraphrasing process. In this study, grade 5 texts show more paraphrases of the complete necessary information from the input material, while grade 2 texts present more fragmented information, which is not enough to fully accomplish the task.

3.4. Linguistic resources

[Sostruznik 2023]revealed that grade 5 texts use a greater number of conjunctions when compared to grade 2 texts in two tasks in which the target genre is a letter to the editor.

In [Hanauer 2023]'s study, grade 5 texts also presented a higher number of words that are typical of the expected discourse genre (in this case, a letter), such as greetings and letter sign-offs. On one of the tasks analyzed, 63% of the grade 5 texts used greetings, and 68% used sign-offs, while in the grade 2 texts, the percentage was 51% for both. Among the texts graded 5 on the second task analyzed, 66% used greetings and 69% used sign-offs, while on the texts graded 2, the percentage was 33% and 41%, respectively.

Frequency wordlists also demonstrated that more advanced texts addressed more directly the interlocutor proposed by the task more directly, as shown in [Hanauer 2023] and [Divino 2024]. The quantitative analysis so far confirmed the previous qualitative analyses, demonstrating a higher frequency of structures typical of the target genre [Mendel 2019] and terms better suited to the proposed audience [Sirianni 2020].

4. Future developments

In addition to completing the compilation of the full corpus with 70,000 texts, a number of steps and tasks can be envisaged, some of which are already underway. The development of a tagging protocol covering spelling issues and discursive features of the texts is in its

[7]The corpus used as references corpus was Portuguese Trends, available on Sketch Engine.

final phase, together with an interrater reliability study to validate it [Stumpf et al. , in preparation].

We aim to use CorSpell, an AI-trained semiautomatic annotation tool developed by the research group that retrieves the spelling variants and suggests form replacements to normalize spelling. This will enable the reanalysis of the data in order to compare the results and verify how much the orthographic inaccuracies were responsible for the distortion in the TTR calculation.

Once the full corpus is compiled, we intend to use multimodal, state-of-the-art transformer-based large language models, such as BERT, GPT, and Gemini, for corpus classification and analysis. The training of generative AI systems can benefit from the variety of genres in the corpus [Pack et al. 2024]. We also aim to extract linguistic features with natural language processing (NLP) tools, which can improve predictive models for automated essay scoring [Mizumoto and Eguchi 2023]. In turn, this will enhance a more robust and refined description of the linguistic patterns found across different proficiency levels in the corpus. Ultimately, we plan to develop a tool that can automatically assess a text according to Celpe-Bras' construct and give personalized feedback to improve users' writing skills.

5. Final remarks

As the first Brazilian corpus of texts categorized by proficiency levels certified by Celpe-Bras, CorCel will enable analyses that contribute to the validation of the exam, improving the description for each certified proficiency level and allowing for better detailing of the assessment criteria for the written part of the exam. The results of CorCel analyses can also support comparisons and alignments with other studies that describe and examine Celpe-Bras and related documents, such as [Schoffen et al. 2018], which provides a comprehensive description of task characteristics, and [Nagasawa 2019], which examines the textual complexity of input texts for tasks 3 and 4. The findings discussed earlier carry implications for the study of proficiency assessment of PAL, particularly regarding to text length, lexical richness using TTR, use of input material and other linguistic resources.

The compilation of CorCel, given its pioneering nature, offers a substantial contribution to the domains of Portuguese as an additional language, proficiency assessment, and corpus linguistics in Portuguese, with the potential to greatly influence research in these fields. The findings from these studies will enhance the understanding of proficiency levels in Portuguese, providing a more detailed analysis of linguistic features, task performance patterns, and textual quality across bands, supporting the development of more precise rating criteria. The examples of language use across different proficiency levels will also provide valuable insights for teaching, as they will enable teachers to design materials and activities that address learners' actual linguistic challenges in each level. Such targeted instruction can foster learners' ability to produce coherent, context-appropriate texts in Portuguese. Exposure to these corpus-based materials can guide students' practice toward producing higher-quality texts, enabling them to engage more confidently with communicative tasks in Portuguese.

References

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

Biber, D. and Gray, B. (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.

Callies, M. and Götz, S. (2015). Learner corpora in language testing and assessment: Prospects and challenges. *Learner corpora in language testing and assessment*, pages 1–9.

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

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

Divino, L. (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.

Divino, L. S. (2024). Contribuições da linguística de corpus para a descrição dos níveis de proficiência escrita no exame celpe-bras: um estudo sobre léxico. Unpublished masters thesis.

Granger, S. and Wynne, M. (2000). Optimising measures of lexical variation in efl learner corpora. In *Corpora galore*, pages 249–257. Brill.

Hanauer, I. (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.

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

Kilgariff, A., Ryckly, P., Smrz, P., and Tugwell, D. (2004). The sketch engine. i: Williams, g. & s. vessier. In *Proceedings of the Eleventh EURALEX International Congress, Lorient, France July 6–10*, pages 105–114.

Kunrath, S. P. (2019). Os descritores gerais e a progressão dos níveis de proficiência do exame celpe-bras. Unpublished doctoral dissertation.

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

Mizumoto, A. and Eguchi, M. (2023). Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Nagasawa, E. Y. (2019). O conteúdo de insumo em tarefas que integram leitura e escrita no celpe-bras: uma abordagem informada por corpus. Unpublished doctoral dissertation.

Pack, A., Barrett, A., and Escalante, J. (2024). Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second language research*, 35(1):121–145.

Raupp, A. M. (2024). Características lexicais das produções escritas do exame celpe-bras na tarefa 3 de 2016-2: uma pesquisa guiada por corpus. Unpublished undergraduate thesis.

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

Schoffen, J., Schlatter, M., Kunrath, S. P., Nagasawa, E. Y., Sirianni, G. R., Mendel, K., Truyllio, L. R., and Divino, L. S. (2018). Estudo descritivo das tarefas da parte escrita do exame celpe-bras: Edições de 1998 a 2017. Technical report, Porto Alegre.

Schoffen, J., Stumpf, E., Amaral, D., Divino, L., Hanauer, I., Lisboa, I., Raupp, A., and Xavier, B. (2024). Compilation and tagging of a corpus with celpe-bras texts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 627–632.

Schoffen, J., Stumpf, E. M., Divino, L. S., Hanauer, I. D., Amaral, D., Raupp, A., and Xavier, B. Corcel: a brazilian portuguese corpus of celpe-bras exam written texts [in press]. *Revista Brasileira de Linguística Aplicada*.

Sirianni, G. R. (2020). Entre a certificação e a não certificação no celpe-bras: um estudo sobre os níveis de proficiência na parte escrita do exame. Unpublished masters thesis.

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

Stumpf, E. M., Schoffen, J., Divino, L. S., Hanauer, I. D., Amaral, D., Raupp, A., and Xavier, B. Interrater reliability study of a tagging protocol for an l2 corpus: the case of corcel. Manuscript in preparation.

Wisniewski, K. (2017). Empirical learner language and the levels of the common european framework of reference. *Language Learning*, 67(S1):232–253.