CorSpell: Introducing a Semiautomatic Tool for Spelling Normalization in Brazilian Portuguese

Juliana Roquele Schoffen[1,5], Dennis Giovani Balreira[1,6], Elisa Marchioro Stumpf[1,5], Larissa Goulart[2], Tanara Zingano Kuhn[3], Rafael Oleques Nunes[1,6], Gabriel Ricci Pazinatto[1,6], Isadora Dahmer Hanauer[1,5], José Henrique de Souza Silva[1,6], Luiza Sarmento Divino[1,5], Marine Laísa Matte[4]

[1]Federal University of Rio Grande do Sul; [2]Montclair State University; [3]University of Coimbra, CELGA-ILTEC, FLUC; [4]Federal Institute of Education, Science and Technology Sul-rio-grandense; [5]IL; [6]INF

{julianaschoffen, isadora.dh, luiza.sarmento.divino, elisa.stumpf}@gmail.com, goulartl@montclair.edu, tanarazingano@uc.pt, marinematte@ifsul.edu.br, {dgbalreira, ronunes, grpazinatto, jhssilv}@inf.ufrgs.br

Abstract

With the growing availability of large text collections, efficient tools for corpus annotation and normalization have become increasingly important in linguistic and computational research. This paper presents CorSpell, an open-source semiautomatic tool developed to support the spelling normalization of Brazilian Portuguese texts within the CorCel project—a corpus comprising over 15,000 exam responses from the Celpe-Bras proficiency test. Given the corpus scale, manual normalization is impractical; CorSpell streamlines this process by enabling users to visualize, select, and replace tokens directly through an intuitive web interface. The tool integrates automatic suggestions from PT-BR dictionaries with human validation, providing an interface for users to access and manipulate the texts. CorSpell significantly reduces annotation time, minimizes errors, and facilitates collaborative work, providing a practical and scalable solution for corpus normalization and a foundation for LLM-based modeling of Portuguese proficiency

Keywords: spelling normalization; Brazilian Portuguese; corpus annotation; natural language processing; linguistic proficiency; Portuguese as an additional language

1. Introduction

In recent years, the increasing availability of large text collections has expanded the possibilities for linguistic and computational research, especially in corpus-based studies. However, the effectiveness of such analyses depends heavily on the quality and consistency of the textual data. In second language (L2) or exam-based corpora, such as those derived from proficiency tests, spelling variation is particularly frequent and poses significant challenges for automatic processing tasks like tokenization and part-of-speech (POS) tagging. These inconsistencies can distort linguistic analyses and hinder the development of reliable language models, demanding efficient tools to support spelling normalization to enhance corpus usability and ensure the accuracy of subsequent computational analyses.

This paper presents CorSpell, a semiautomatic tool to aid the process of spelling normalization of texts written in Brazilian Portuguese. More specifically, CorSpell has been developed to facilitate manual annotation of the Brazilian Portuguese Corpus of Celpe-Bras Exam Written Texts (CorCel) (Schoffen et al., 2025, 2024), within the context of a research project that aims to develop artificial intelligence tools based on large-scale language models (LLMs) to describe linguistic proficiency and develop pedagogical resources in Portuguese as an additional language (PAL).

Celpe-Bras stands for Certificate of Proficiency in Portuguese for Foreigners and is the official Brazilian proficiency exam. It ranks proficiency at six levels and comprises two parts: oral and written. The source of CorCel is a set of 15,000 level-assigned, digitized written texts produced under Celpe-Bras exam conditions in four different editions. At present, CorCel consists of the typed and proofread versions of those texts. However, significant spelling variation, which is common in this context of text production, undermines POS-tagging, thus hindering quantitative corpus analysis. Corpus-based researchers would certainly benefit from a spelling-normalised version of CorCel, since the better the tagging, the more accurate the statistical analysis. Moreover, in order to train the LLM models to develop the generative AI-driven tools abovementioned, a diverse sample of texts from CorCel must be manually annotated. As a result, CorSpell was developed to facilitate manual annotation of CorCel at the spelling level.

This paper is structured as follows: In Section 2, we introduce CorCel, a Brazilian Portuguese Cor-

Source code available at https://github.com/jhssilv/corcel-platform.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1659–1667
11–16 May 2026.

pus of Celpe-Bras Exam Written Texts. In Section 3, we review related work in the fields of annotation and automatic normalization. In Section 4, we introduce the tool from the user perspective, together with technical aspects of its development. In Section 5, we examine how the tool has improved the annotation process. In Section 6, we conclude the text by discussing how such a tool can contribute to future works in the fields of natural language processing and linguistics.

2. The CorCel Corpus

As mentioned above, the corpus currently includes around 15,000 texts that were manually typed and proofread from their corresponding images, which is the format in which 70,000 samples were originally provided to researchers. To date, several studies have been carried out using this corpus (Hanauer, 2026; Stumpf et al., 2025; Raupp, 2024; Hanauer, 2023, 2022; Divino, 2021, 2024; Silveira, 2023; Kunrath, 2019; Sirianni, 2020; Mendel, 2019). However, it is possible to gain different insights into linguistic features across different proficiency levels through automated quantitative analysis. For certain quantitative analyses, such as lexical chunks and key-feature analysis, it is essential to normalize word spelling to prevent statistical measurements from being skewed towards a specific word form. Thus, the second step in the corpus compilation process is to create a normalized version, since texts handwritten under exam conditions by candidates who speak Portuguese as an additional language usually contain different forms of the same word (e.g. *casa* vs *caza*, *prezado* vs *presado*).

Our main goals with the normalization process were to enable more accurate measures of lexical diversity and to identify copies of excerpts from the input texts available to candidates in the exam. However, considering the large volume of texts in the CorCel corpus, it would be extremely time-consuming and thus expensive to normalize them manually. In addition, manual normalization is likely to result in inconsistencies across texts. The tool described here has been developed to streamline corpus annotation, so that in the next phase of the project, language models are trained and fine-tuned to perform the spelling normalization according to the guidelines developed by the research team (Section 4). While this tool has a clear purpose in our project, it can also be used as a standalone tool by researchers also dealing with spelling normalization of Brazilian Portuguese.

Our semiautomatic tool allows users to, while viewing a text, select words, replace them, and save this information in a database. Although other corpus annotation tools offer similar features, our tool stands out by suggesting replacement candidates, thus effectively streamlining the annotation process. So far, for token analysis and the suggestion of replacement forms, the tool uses two different dictionaries and a customizable whitelist.

3. Related Work

Spelling normalization is a process usually carried out via annotation tools, which create a new layer of information over a given text span. In this section, we provide an overview of some of these tools and other resources that can be used for general text annotation and spelling normalization.

Many initiatives of learner corpora have fairly comprehensive annotation systems (Rudebeck and Sundberg, 2021; Granger et al., 2022), particularly to annotate different types of errors. The SVALA tool (Wirén et al., 2019) was developed as part of the SweLL project to support the annotation of learner texts written by L2 speakers of Swedish. SVALA offers an integrated environment for pseudonymization, normalization, and correction annotation. Its central feature is a dynamic word-alignment system between the original and corrected versions of the text, allowing annotators to edit learner productions and simultaneously label corrections with detailed taxonomies.

TEITOK (Janssen, 2016) is a web-based platform designed to annotate corpora, which has been used by many projects dealing with learner production and historical texts, which also suffer from spelling variation. It integrates multiple orthographic layers (e.g., original, expanded, normalized) and provides token-level annotation through an inline XML format compatible with TEI. TEITOK supports manual and automatic processes, such as POS tagging and lemmatization, and includes an interactive interface for editing tokens and visualizing changes. It offers modules for error annotation and normalization, including support for stand-off annotation and alignment between original and normalized forms.

When it comes to spelling normalization, VARD 2 is a well-known tool designed specifically to deal with spelling variation in historical texts written in English (Baron and Rayson, 2008), with a focus on the early modern period. Whenever a modern equivalent of a given word is not found in the tool's modern lexicon, a list of candidate modern equivalents ranked by 'confidence' is produced and presented to the user (in its interactive version; there is also a batch processing version). It is also possible to instruct the system to select the top candidate for each variant if its 'confidence' score exceeds a user-defined threshold.

To the best of our knowledge, there is only one resource for spelling normalization in Brazilian Portuguese (Garcia et al., 2023), a pipeline developed

to deal with lexical variation in texts from the agricultural domain. It checks every type against a general lexicon in Portuguese (PortiLexicon-UD (Lopes et al., 2022)) and a specialized lexicon, among other steps, so that a .csv file is generated with unknown words and respective suggestions to replace them.

INCEpTION (Klie et al., 2018) is a machine-assisted, interactive annotation platform with a focus on semantic annotation, including knowledge base population, entity and fact linking. As a highly customizable tool, it is possible to create annotation layers based on different projects' needs. Among other functionalities, INCEpTION offers "recommenders", which assist annotators by recommending annotation suggestions. The recommendation algorithm is based on machine learning and/or knowledge resources and updated or retrained according to the users' activities in the platform.

While both TEITOK and SVALA offer comprehensive environments for annotating learner corpora, our project focuses on the identification and replacement of words in texts written by speakers of PAL, focusing primarily on orthographic features. VARD 2 and the pipeline described above are resources that focus specifically on spelling; however, the former is set to work with English and the latter, while being able to suggest form replacements in Brazilian Portuguese, does not have a user-friendly interface, which would allow its use by researchers without a comprehensive knowledge of programming. A similar shortcoming applies to INCEpTION. Despite its many functionalities supported by advanced technology, for our project purposes and the profile of its annotators - mostly undergraduate research assistants attending the Modern Languages course -, a much simpler but yet powerful tool was required.

Table 1 compares the four tools reviewed above in terms of four key functionalities: the ability to new annotation layers over parts of a text, offer word suggestions in Brazilian Portuguese, export new annotated versions, and provide a user-friendly interface. This analysis led the research team to develop CorSpell, which is described in the next section.

4. CorSpell

This section presents the main components and workflow of CorSpell, outlining how texts are processed, visualized, normalized, and managed within the system. It begins with the pre-processing and data ingestion pipeline, which prepares and structures textual data for use. Next, it introduces the user interface and text visualization modules, detailing how users interact with the corpus and view normalization suggestions. The normalization workflow section explains how users can apply or edit corrections, while progress tracking and export describe features for managing completion status and exporting results. Finally, the system architecture subsection summarizes the underlying technical design that supports all these functionalities. Figures 1 and 2 show a pipeline containing an overview of our methodology.

4.1. Pre-processing and data ingestion

Before being accessible via CorSpell, texts require pre-processing and database insertion. This is accomplished via a pipeline that first tokenizes the text with the spacy-udpipe package[1], then identifies words requiring normalization and generates substitution forms (candidates) for them. This pipeline also employs two distinct dictionary components: a common dictionary[2] and an affix dictionary[3], alongside a self-hosted gemma 3 12B model (Team, 2025). Along with the LLM, both dictionaries detect words to be normalized and generate replacement suggestions. This dual-dictionary strategy contributed to increasing the tagging of non-standard spelling, without significantly increasing computational cost, while the LLM addition substantially improved quality and context awareness of suggestions, though formal benchmarks of different models performance are still pending. Following this stage, all generated information is properly structured and inserted into the database. Users can easily send texts to be processed by uploading them through the tool's interface, which currently supports zipped batches of .docx and .txt files.

4.2. Token whitelisting

A set of "whitelisted" words can be defined, ensuring that context-specific proper nouns previously flagged as incorrect by the dictionaries will not be marked for normalization when visualizing a text. Additionally, the system supports adding suggestions to all the occurrences of a specific token in the corpus. This "global suggestion" feature helps to address cases in which the processing pipeline correctly marked a word for normalization, but failed to propose an adequate substitution form.

4.3. Interface and Corpus Navigation

Upon user authentication, the tool retrieves metadata for all texts from the database and loads the main interface. This interface, as shown in Figure 3, features a text selection utility, enabling users

[1] https://spacy.io/universe/project/spacy-udpipe

[2] https://www.ime.usp.br/~pf/dicios/

[3] https://github.com/wooorm/dictionaries/tree/main/dictionaries/pt

Tool/Resource	TEITOK	SVALA	INCEpTION	VARD 2	Pipeline
Creates a new layer of annotation over one or more tokens	Yes	Yes	Yes	Yes	No
Offers suggestions for token replacement in PT-BR	No	No	Unclear	No	Yes
Exports new versions of annotated texts	Unclear	No	Yes	Yes	No
Has a user-friendly interface	Yes	Yes	Yes (high learning curve)	Yes	No

Table 1: Tools and resources for annotation and spelling normalization.

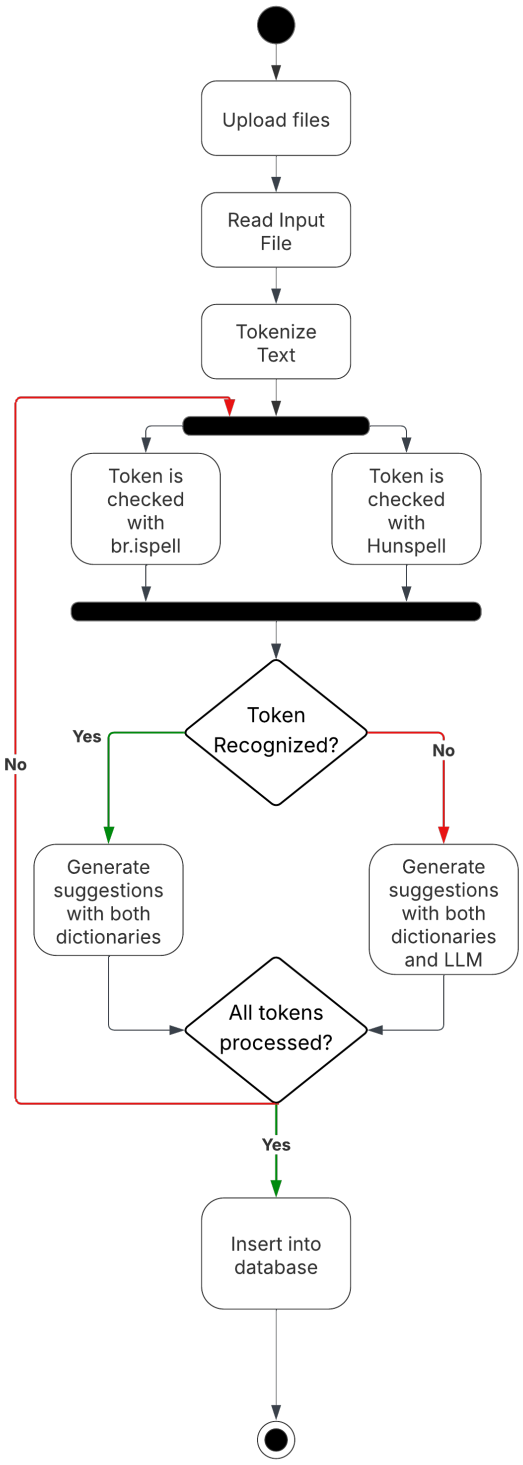

Figure 1: Text pre-processing.

to easily filter and search the corpus based on four criteria: id (original filename), score (as assessed in the Celpe-Bras exam), assigned users, and whether the text is marked as normalized by the current user. Users can then select a text from the filtered list or retrieve it directly by typing the filename.

4.4. Text Rendering and Visualization

Once a text is selected, the tool fetches its corresponding tokens, reconstructing the text with the correct spacing and applying existing normalizations. The text is rendered in a series of clickable spans. These spans are color-coded to visually indicate the status of each token: whether it possesses a replacement suggestion, is already normalized, or is currently selected. We show an example of our text rendering in Figure 4.

4.5. Normalization Workflow

When a token with replacement suggestions is selected, CorSpell displays candidates in a floating panel above the selected word, as seen in Figure 4, allowing users to quickly apply changes by selecting one of them. The tool also allows manual input for any word, accommodating cases when non-standard tokens are not flagged as such or the desired substitution has not been generated. This and more options related to the selected token can be seen in a second floating panel, visible in Figure 3. Submitted changes are instantly stored in the database and applied to the currently selected text, triggering a view refresh. Also, adjacent tokens can be selected and substituted at once.

4.6. Progress Tracking and Exporting

For progress tracking, users can mark a text as "finished" once its normalization is complete. All normalization data is stored on a per-user basis, allowing multiple versions of each text to exist in the database simultaneously. Finally, the application provides an export function that adds the possibility of downloading .zip files containing all currently selected normalized texts. This export includes an option to format the applied changes using XML-like syntax so that these files (in a .txt format) can be used in different tools for corpus analysis, such as AntConc (Anthony, 2024) and Sketch Engine

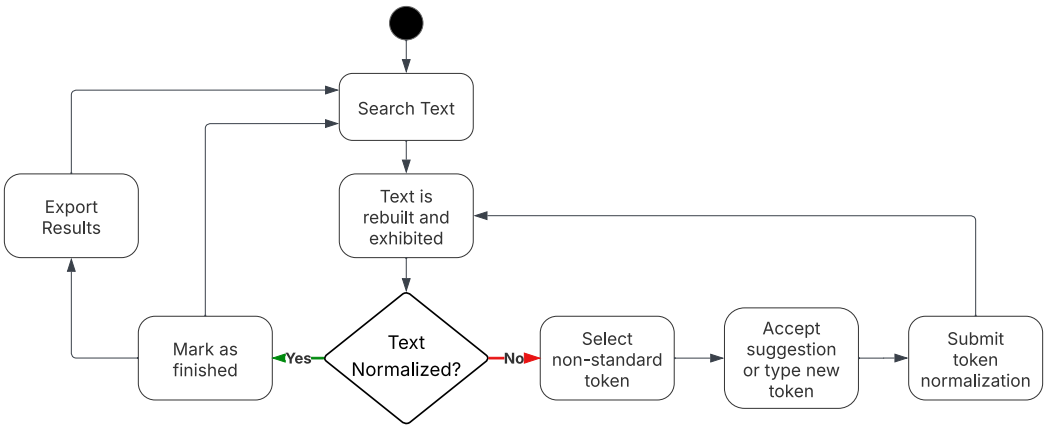

Figure 2: Text normalization process. Upon selecting a specific text from the dataset, the tool renders it (as described in Section 3.3) and allows the user to substitute tokens, marking it as finished or exporting the current results for a given set of texts.

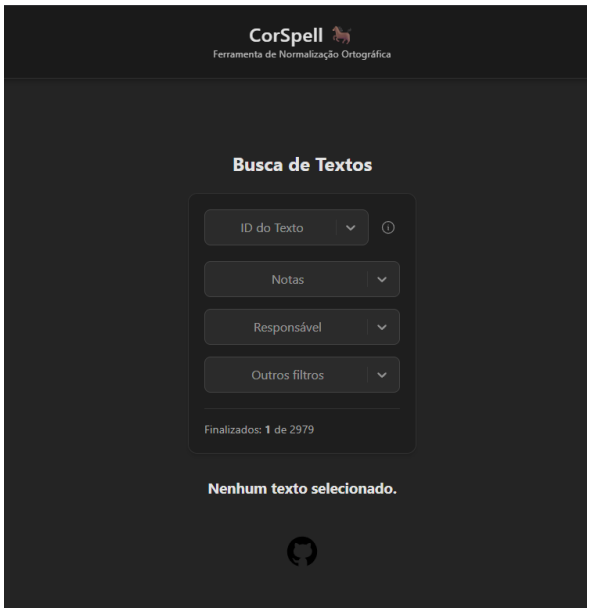

Figure 3: CorSpell's text selection interface.

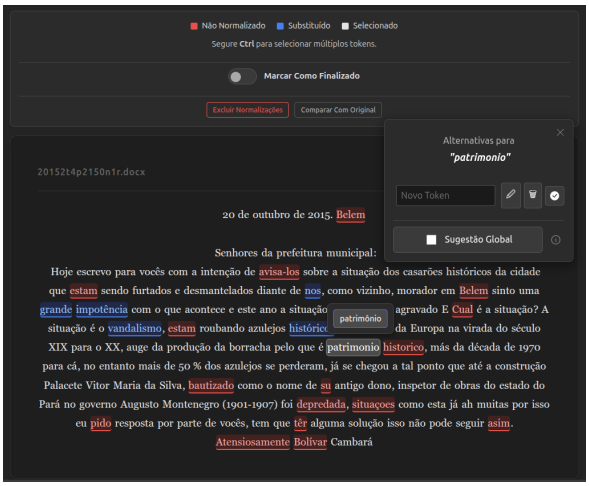

Figure 4: CorSpell's text visualization and editing interface.

(Kilgariff et al., 2004). Users also have the option to generate reports in csv format, containing a summarization of all the performed normalizations in the currently selected texts.

4.7. OCR Module

To accelerate the compilation process, and considering the availability of approximately 70,000 texts, we explored the use of a large language model (LLM) to convert the images into machine-readable text. This initiative led to the development of an OCR module, currently the most recent component of the Corspell platform. The tool also implements an optional OCR page that supports the uploading and processing of scanned documents. On upload, the files are sent to the gemini-flash-latest[4] API for transcription. The returned results can later be accessed in the OCR page, allowing users to easily find and edit texts for eventual mistakes made by the LLM, before processing and inserting them into the corpus database. We show an example of this pipeline in Figure 5.

While this process utilizes a third party API, research into more accessible and privacy-preserving alternatives is ongoing, also taking into consideration the difference in quality of the transcriptions made by larger models and locally-accessible alternatives.

[4]https://deepmind.google/models/gemini/flash/

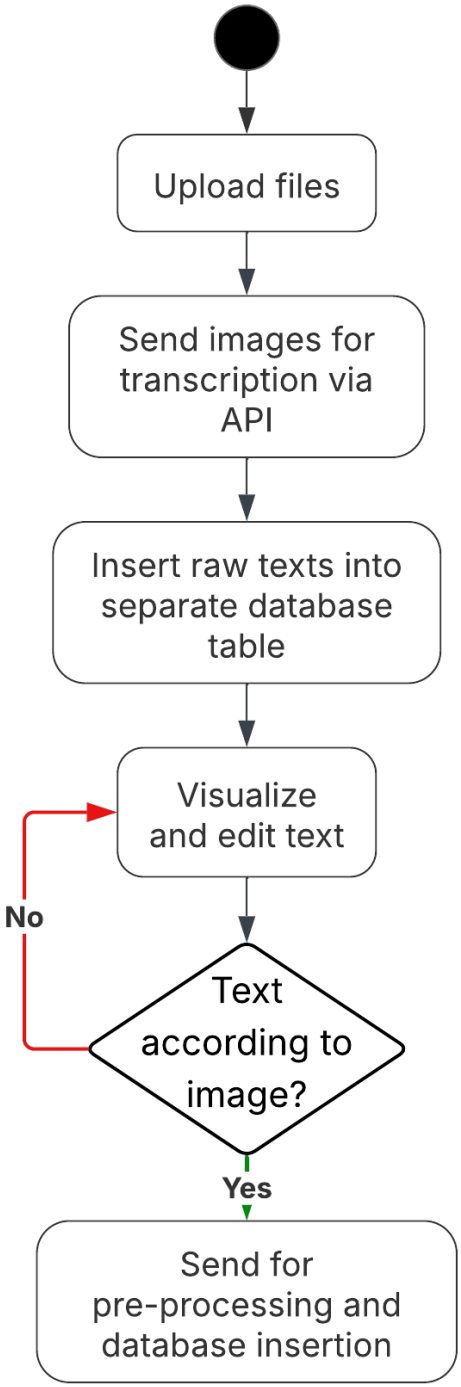

Figure 5: CorSpell's OCR processing pipeline.

4.8. Design Choices and Limitations

By open-sourcing the tool, we allow other researchers to adapt and customize it for their own use cases. To make this process easier, effort is still being put into creating robust documentation for the codebase, while also creating alternatives for non-technical users to easily set up the tool for collaborative use.

Additionally, while scalability was a concern during the development of the platform, its performance-wise capabilities are still to be determined. Currently, our database contains more than 600,000 tokens distributed across almost 3000 processed texts, with hardly any considerable latency or responsiveness issues being noticed.

On the other hand, the user management features were not originally designed to handle a large number of registrations. Currently, administrators must manually create new accounts, after which users set their passwords during initial authentication. Although security measures were implemented to protect the tool and its database from attacks, leaks, and unauthorized access, this registration workflow may lack the robustness required for certain use cases.

4.9. System Architecture

CorSpell was developed following a layered architectural pattern to support the functionalities previously described. Its backend relies on a PostgreSQL database to store tokens, user details, text alterations and metadata. A Flask API intermediates all interactions with the database, utilizing SQLAlchemy for database modeling and connectivity, and Pydantic for request data validation. The frontend, built in React, employs an Axios client for API communication and Zod schemas to validate and handle data exchanges. Asynchronous services, such as the OCR and text processing, have been implemented via Redis and Celery workers.

5. Results

Multiple researchers involved in the project have already tested and used the annotation tool. Compared to the previous workflow, which was carried out manually using XML syntax in Microsoft Word or Google Docs, the new interface speeds up the process. Annotators no longer need to insert tags or handle formatting: the possibility of simply selecting a misspelled word and replacing it with the intended form has reduced the average annotation time by several minutes per text. In addition, less time is spent on training new annotators. Hence, the saved time allows researchers to focus directly on the annotation criteria. So far, only one study has used it extensively (Hanauer, 2026) to normalize 521 texts from the CorCel corpus, following a protocol defined by the research team and adapted to the specific task. Table 2 shows that texts in grades 1-3 (lower levels) undergo the most normalizations, while texts with higher scores require considerably fewer changes, indicating that more pro-

Subcorpus	Number of texts	Number of normalizations	Average number of normalizations per text
0	21	227	10.81
1	100	1802	18.02
2	100	1818	18.18
3	100	1834	18.34
4	100	1287	12.87
5	100	826	8.26
Total	521	7794	14.96

Table 2: Text normalization (Hanauer, 2026)

Subcorpus	Number of texts	Original types	Types after normaliazation	% Reduction
0	21	982	910	7.33%
1	100	2742	2111	23.01%
2	100	2649	2010	24.12%
3	100	2783	2168	22.10%
4	100	2588	2175	15.96%
5	100	2457	2174	11.52%
Total	521	7290	5187	28.85%

Table 3: Difference in types before and after normalization (Hanauer, 2026)

ficient candidates use more standard word forms. Table 3 presents the difference in types before and after normalization, which follows a trend similar to the number of normalizations. This difference affects measures such as the type-token ratio (TTR), thereby providing a more accurate perspective on candidates' lexical competence. The tool features a user-friendly and intuitive interface that makes the annotation process more efficient also by streamlining the management of files and versions. Previously, annotators had to rename files manually or maintain external spreadsheets to track which texts had already been annotated and how many remained. Now, the platform automatically records this information, making it much easier to monitor progress and organize the workflow.

In addition to speeding up the annotation process, the platform also reduces the likelihood of tagging errors. Because the tool automates the tagging process, annotators no longer need to worry about syntax issues such as typos or inconsistent spacing. This ensures more consistent and accurate annotations, especially in collaborative settings where standardization is essential, helping annotators adjust their work as they progress. As a result, annotators can dedicate more attention to the linguistic aspects of the task, rather than spending time on file management or technical formatting.

Another significant advantage is that the tool enables multiple annotators to work on the same dataset without worrying about version control or conflicting file updates, which is particularly beneficial for teamwork and for keeping track of all changes made during the annotation process. This feature is currently being used in an interrater reliability study to test the normalization guidelines developed by the team. The guidelines establish that nonstandard forms found in texts have their spelling normalized, considering the closest word possible in Brazilian Portuguese that makes sense in the context. Besides this general orientation, there are instructions for recurrent cases that were found in the texts, particularly in lower-level ones. This will allow a better understanding of how the different members of the team are working in the annotation process and come up with a workflow to deal with disagreements and other issues.

6. Final Remarks and Future Work

This paper introduced CorSpell, a semiautomatic tool designed to support the spelling normalization of Brazilian Portuguese texts within the CorCel project. By integrating automatic normalization suggestions with a user-friendly annotation interface, CorSpell has considerably improved the efficiency, accuracy, and consistency of the normalization process. The tool not only reduces annotation time and training effort but also facilitates collaborative work and version control, ensuring a more reliable and scalable workflow. Beyond its immediate impact on corpus preparation, CorSpell provides an essential foundation for the development of language models and pedagogical applications aimed at describing and supporting Portuguese as an additional language.

The annotation made with the help of CorSpell will be used, in the next stages of our project, to create an automatic spelling normalization tool with a fine-tuned LLM, thus allowing for further corpus

growth, as well as to support the development of generative-AI tools for proficiency description and pedagogical resource creation. When it comes to the tool's interface, we plan to make the filters to search the corpus easily customizable for different uses in the future.

As stated before, CorSpell can be adapted to other types of spelling normalization, including in other languages. However, it was also built having in mind that Portuguese is a low-resource language, making it an important contribution to the field of natural language processing and can foster different studies using corpora from other exams, such as ENEM[5], for example, and from other contexts, such as social media. From a linguistic perspective, normalization facilitates the creation of cleaner and more reliable corpora by reducing orthographic variation, thereby enabling accurate analyses of lexical, morphological, and syntactic patterns.

There are also broader impacts of CorSpell. We believe that standardized corpora allow other researchers to replicate linguistic findings and improve tools collaboratively. Normalized texts can also be more easily aligned with related ones written in other languages, facilitating cross-lingual embeddings and transfer learning. They can also help the development of applications like text-to-speech and search engines.

Acknowledgements

This work was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and by Petrobras.; by international funding, in the framework of the project https://doi.org/10.54499/UID/PRR2/04887/2025; and by national funds through FCT – Foundation for Science and Technology I.P., in the framework of the Project CELGA-ILTEC (UID/04887/2025).

7. Bibliographical References

Laurence Anthony. 2024. Antconc (version 4.3.1) [computer software]. https://www.laurenceanthony.net/software/AntConc. Tokyo, Japan: Waseda University.

Alistair Baron and Paul Rayson. 2008. Vard2: A tool for dealing with spelling variation in historical corpora.

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

Luiza Sarmento Divino. 2024. Contribuições da linguística de corpus para a definição de níveis de proficiência escrita no exame celpe-bras. Master's thesis, Universidade Federal do Rio Grande do Sul.

Luana Q Garcia, Miguel H Chinellato, Helena de M Caseli, and Leandro HM Oliveira. 2023. Pipeline para identificação de erros lexicais e geração de sugestões de correção. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 357–361. SBC.

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

Isadora Dahmer Hanauer. 2022. Influência das inadequações ortográficas em análise de tarefa escrita do celpe-bras guiada por corpus.

Isadora Dahmer 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. Undergraduate thesis.

Isadora Dahmer Hanauer. 2026. Tarefas de compreensão audiovisual para produção escrita no celpe-bras: análise de produções de examinandos guiada por corpus. Master's thesis, Universidade Federal do Rio Grande do Sul.

Maarten Janssen. 2016. Teitok: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043.

Adam Kilgariff, Pavel Ryckly, Pavel Smrz, and David Tugwell. 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.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

[5]ENEM stands for Exame Nacional do Ensino Médio and is Brazil's national high school exam, used to assess students' performance and as the main gateway to higher education.

Simone Paula Kunrath. 2019. *Os descritores gerais e a progressão dos níveis de proficiência do exame CELPE-BRAS*. Ph.D. thesis, Universidade Federal do Rio Grande do Sul.

Lucelene Lopes, Magali Duran, Paulo Fernandes, and Thiago Pardo. 2022. PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.

Kaiane Mendel. 2019. Proficiência e autoria na avaliação integrada de leitura e escrita do exame celpe-bras. Master's thesis, Universidade Federal do Rio Grande do Sul.

Amanda Michel Raupp. 2024. Características lexicais das produções escritas do exame celpe-bras na tarefa 3 de 2016-2: uma pesquisa guiada por corpus. Undergraduate thesis.

Lisa Rudebeck and Gunlög Sundberg. 2021. Swell correction annotation guidelines.

Juliana Schoffen, Elisa Stumpf, Deise Amaral, Luiza Divino, Isadora Hanauer, Isabel Lisboa, Amanda Raupp, and Brenda Xavier. 2024. Compilation and tagging of a corpus with celpe-bras texts. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 627–632, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Juliana Roquele Schoffen, Elisa Marchioro Stumpf, Luiza Sarmento Divino, Isadora Dahmer Hanauer, Deise Amaral, Amanda Michel Raupp, and Brenda de Souza Xavier. 2025. Corcel: A brazilian portuguese corpus of celpe-bras exam written texts. *Revista Brasileira de Linguística Aplicada*, 25(1):e50034.

Júlia Luiz Sostruznik da Silveira. 2023. O uso de conjunções em produções escritas no exame celpe-bras: um estudo baseado em corpus. Undergraduate thesis.

Gabrielle Rodrigues Sirianni. 2020. Entre a certificação e a não certificação no celpe-bras: um estudo sobre os níveis de proficiência na parte escrita do exame. Master's thesis, Universidade Federal do Rio Grande do Sul.

Elisa Stumpf, Juliana Schoffen, Luiza Divino, Isadora Hanauer, Amanda Raupp, and Brenda Xavier. 2025. Corpus-driven lexical analyses of corcel: a comparative analysis of preliminary findings of written proficiency in portuguese as an additional language. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 673–681, Porto Alegre, RS, Brasil. SBC.

Gemma Team. 2025. Gemma 3.

Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2019. Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora. In *CLARIN Annual Conference, Pisa, Italy, 8-10 October, 2018*, pages 222–234. Linköping University Electronic Press.