

INTRODUCING CROPS: A SPECIALIZED CORPUS OF ENGLISH-LANGUAGE
AGRARIAN SCIENCES POSTGRADUATE THESES AND DISSERTATIONS FROM
BRAZILIAN UNIVERSITIES

*Apresentando o CROPS: um Corpus Especializado de Teses e Dissertações em Língua Inglesa
das Ciências Agrárias de Universidades Brasileiras*

DOI: 10.14393/LL63-v41S-2026-10

Deise Prina Dutra¹

Danilo Duarte Costa²

Carolina Godoi de Faria Marques³

Gustavo Leal Teixeira⁴

Ana Eliza Pereira Bocorny⁵

Jhonatan Henrique Lopes Alves⁶

¹ PhD in Linguistics. Universidade Federal de Minas Gerais, Brazil. ORCID: 0000-0001-5799-5174. E-mail: deisepdutra(AT)gmail.com

² PhD in Linguistic Studies. Universidade Federal dos Vales do Jequitinhonha e Mucuri. ORCID: 0000-0003-2619-0175. E-mail: danilo.costa(AT)ufvjm.edu.br

³ MA in Linguistic Studies. Universidade Federal de Minas Gerais. ORCID: 0009-0004-3927-4428. E-mail: carol.godoi350(AT)gmail.com

⁴ PhD in Linguistics. Universidade Federal de Minas Gerais. ORCID: 0000-0001-7293-0790. E-mail: lealtex(AT)gmail.com

⁵ PhD in Language Studies. Universidade Federal do Rio Grande do Sul. ORCID: 0000-0002-0515-9630. Email: ana.bocorny(AT)gmail.com

⁶ BA in English Language and Literature. Universidade Federal de Minas Gerais. ORCID: 0009-0008-4914-0262. E-mail: jhlopesalves(AT)gmail.com

ABSTRACT: In this paper, we detail the design and organization of the Corpus of Agrarian Sciences Postgraduate Studies (CROPS), a specialized corpus composed of English-language master's theses and doctoral dissertations from Brazilian postgraduate programs in Agrarian Sciences, comprising approximately 2.4 million words, distributed across five subject areas: Animal Science, Agricultural Engineering, Agronomy, Food Engineering, and Forestry. CROPS is segmented into multiple subcorpora, organized according to disciplinary subareas, text types, text formatting, and text sections. The methodological procedures adopted in compiling the corpus are grounded in Corpus Linguistics principles, particularly the concept of representativeness (Biber, 1993), as well as rigorous methodological parameters of corpus compilation (McEnery; Brookes, 2022; Reppen, 2022; Ädel, 2020; Collins, 2019). Finally, we provide a quantitative overview of the corpus. CROPS constitutes a reliable resource for linguistic research, representing contemporary academic writing practices of Agrarian Sciences graduate students in Brazil.

KEYWORDS: Corpus compilation. Specialized corpus. English for Academic Purposes. Academic writing. Agrarian Sciences.

RESUMO: Neste artigo, detalhamos a construção e a organização do CROPS (*Corpus of Agrarian Sciences Postgraduate Studies*), um corpus especializado composto de dissertações e teses escritas em inglês provenientes de programas de pós-graduação brasileiros na área das Ciências Agrárias, totalizando aproximadamente 2,4 milhões de palavras, distribuídas em cinco subáreas: Zootecnia, Engenharia Agrícola, Agronomia, Engenharia de Alimentos e Engenharia Florestal. O CROPS está segmentado em múltiplos subcorpora, organizados de acordo com as subáreas das Ciências Agrárias, tipos, formato e seções dos textos. Os procedimentos metodológicos adotados na compilação do corpus estão fundamentados nos princípios da Linguística de Corpus, como a representatividade (Biber, 1993), além de rigorosos parâmetros metodológicos de compilação de corpora (McEnery; Brookes, 2022; Reppen, 2022; Ädel, 2020; Collins, 2019). Por fim, apresentamos um panorama quantitativo do corpus. O CROPS constitui um recurso confiável para pesquisas em linguística, sendo representativo das práticas contemporâneas de escrita acadêmica de estudantes de pós-graduação em Ciências Agrárias no Brasil.

PALAVRAS-CHAVE: Compilação de corpus. Corpus especializado. Inglês para Fins Acadêmicos. Escrita acadêmica. Ciências Agrárias.

1 Introduction

In Brazil, the vast majority of PhD dissertations and master's theses are written in Portuguese. However, over the past two decades, there has been an increase in the number of these works written in English. In the year 2000, they accounted for 0.6% of all dissertations and theses in the country, whereas by 2023 this figure had risen to 3.4% (CAPES, 2024, 2025a).

Indeed, English has become the global language of academia, serving as the main medium of communication and publication in leading international journals. Unsurprisingly,

non-Anglophone scholars often report being pressured to publish in English (Hyland, 2016; Lillis; Curry, 2010). For early-career researchers and graduate students who use English as an additional language, international publishing represents an even greater challenge (Flowerdew; Habibie, 2022; Shamsi; Osam, 2022).

Academic writing displays specific conventions and linguistic features that distinguish it from other genres/registers. These distinctive characteristics operate not only at the lexical and syntactic levels but also in the rhetorical and discourse organization of texts (see, e.g., Swales; Feak, 2012; Biber; Gray, 2010; Hyland; Bondi, 2006; Hyland, 2004).

In the field of Corpus Linguistics, numerous studies have sought to identify the linguistic characteristics of academic writing. These include detailed linguistic descriptions of academic writing in general (e.g., Biber, 2006; Biber; Gray, 2010), as well as important findings regarding variations and patterns at the lexical, syntactic, and discursive levels across different academic disciplines (e.g., Hyland; Bondi, 2006; Swales; Feak, 2012; Parkinson, 2013; Gray, 2015). As Hyland (2012) points out, corpus-based research provides valuable insights into how language is used and varies across disciplines, thereby contributing to a more informed understanding and teaching of academic discourse.

However, for the field of Agrarian Sciences specifically, Shi and Wannaruk (2014) report that studies on the characteristics of academic writing are scarce — a reality that, to the best of our knowledge, remains unchanged. In response, we address this gap by compiling a specialized corpus of English-language academic texts produced by Brazilian postgraduate students in Agrarian Sciences: the Corpus of Agrarian Sciences Postgraduate Studies (CROPS).

CROPS is a novel, specialized corpus consisting of 166 master's theses and doctoral dissertations written in English by students from Agrarian Sciences postgraduate programs at Brazilian universities, comprising approximately 2.4 million words. We will use this corpus to develop instructional materials that build disciplinary academic writing in the Brazilian Agrarian Sciences community. In line with Nassi-Calò (2016), the goal goes beyond general English proficiency to understanding and producing the field's discourse, including its genres, rhetorical moves, and phraseology, which can strengthen the visibility and impact of research from non-Anglophone contexts.

The present paper outlines the methodology employed in the development of CROPS, from its initial phases to completion, detailing the challenges and methodological procedures adopted throughout its compilation.

2 The compilation of CROPS

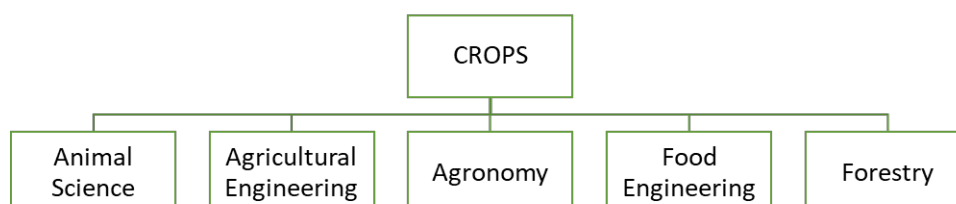
2.1 Corpus design

The construction of CROPS was preceded by a planning phase that involved decisions about its scope and structure. First, we addressed an essential step in corpus building: to have a well-defined research question (Reppen, 2022). Here, this research question is best understood as a research goal: to identify and examine linguistic patterns in the Agrarian Sciences English academic writing for the further development of instructional materials.

Our second strategic decision involved determining which texts would be included in the corpus - an elementary step in corpus design (McEnery; Brookes, 2022), as it underlies the concept of representativeness proposed by Biber (1993). The need for a careful selection of texts is also stressed in Ädel (2020, p. 4), who asserts that samples should “maximally represent the target discourse or population.” In this regard, we initially opted for including three types of academic written work: undergraduate theses, master’s theses, and doctoral dissertations.

Given that the Agrarian Sciences comprise a broad field of inquiry, it was determined that the corpus should comprise the following subject areas: Animal Science, Agricultural Engineering, Agronomy, Food Engineering, and Forestry. This choice was based upon the existing subject areas at the *Instituto de Ciências Agrárias* (ICA) from the *Universidade Federal de Minas Gerais* (UFMG, Brazil). These subject areas make up one of CROPS’s subcorpora, as illustrated in Figure 1.

Figure 1 - Agrarian Sciences subject areas that compose CROPS.



Source: the authors.

2.2 Search and selection of texts

From the outset of the corpus compilation, we ensured that the collection process adhered to legal and ethical standards. Accordingly, texts were retrieved from online institutional repositories of Brazilian universities and research centers, as it is standard practice for these institutions to archive and provide open access to students' academic work. By obtaining the texts directly from these sources, we guaranteed compliance with copyright regulations.

For the inclusion criteria, it was established that the texts should (i) be written in the form of undergraduate theses, master's theses, or doctoral dissertations within the specified Agrarian Sciences subject areas; (ii) be written in English; and (iii) have been published from the year 2000 onward. With regard to the temporal criterion (iii), the year 2000 was adopted as the cut-off point to ensure that the data would represent the language used in contemporary academic writing in Agrarian Sciences, as specified in Collins (2019).

No undergraduate theses written in English were found - a limitation that was partly anticipated, given that undergraduate theses in Brazilian universities are predominantly written in Portuguese. As a result, CROPS's final composition comprises exclusively master's theses and doctoral dissertations.

The process of searching for texts in institutional repositories was long and detailed; it involved several collaborators, including undergraduate research students under our supervision. By resorting to data from the Plataforma Sucupira⁷ (CAPES, 2025b), we compiled a list of all universities and research institutions in Brazil that offer graduate programs (Master's and PhD) in Agrarian Sciences within the five subject areas previously defined. We aimed to collect samples representing academic writing nationwide, rather than concentrating on a particular university or region, thereby enhancing the diversity of research topics encompassed in the corpus.

A total of 85 institutions were identified across the five geographical regions of Brazil: South, Southeast, Central-West, North, and Northeast (see Appendix 1). In CROPS, 22 of these

⁷ Plataforma Sucupira is an online environment that offers content related to the National Graduate Education System (SNPG) in Brazil. <<https://sucupira-legado.capes.gov.br/sucupira/public/consultas/coleta/programa/listaPrograma.jsf>>.

institutions are represented, mainly concentrated in the South and Southeast, as shown in Table 1.

Table 1 - Number of institutions in Brazil offering graduate programs in Agrarian Sciences in comparison to the number of institutions represented in CROPS.

Geographical Region	N. of institutions in Brazil	N. of institutions represented in CROPS	Percentage of institutions represented in CROPS
Central-West	12	1	8,3%
Northeast	24	4	16,7%
North	10	2	20,0%
Southeast	24	8	33,3%
South	15	7	46,7%
Total	85	22	25,9%

Source: the authors.

As can be seen from Table 1, CROPS comprises texts from approximately one quarter (25.9%) of the Brazilian institutions that offer graduate programs in Agrarian Sciences within the five defined subject areas. This coverage can be considered significant, given the specific challenges of compiling a corpus of English-language theses and dissertations in a country where they are predominantly written in Portuguese. Data from CAPES (2025a) show that in the year 2023, for instance, only 3.4% of all theses and dissertations produced in Brazil were written in English. It is also worth mentioning that our methodology of sourcing texts exclusively from institutional repositories presented additional constraints, such as the absence of filters to select texts by language and the inaccessibility of repository websites, which at times were offline or incomplete.

2.3 Corpus organization

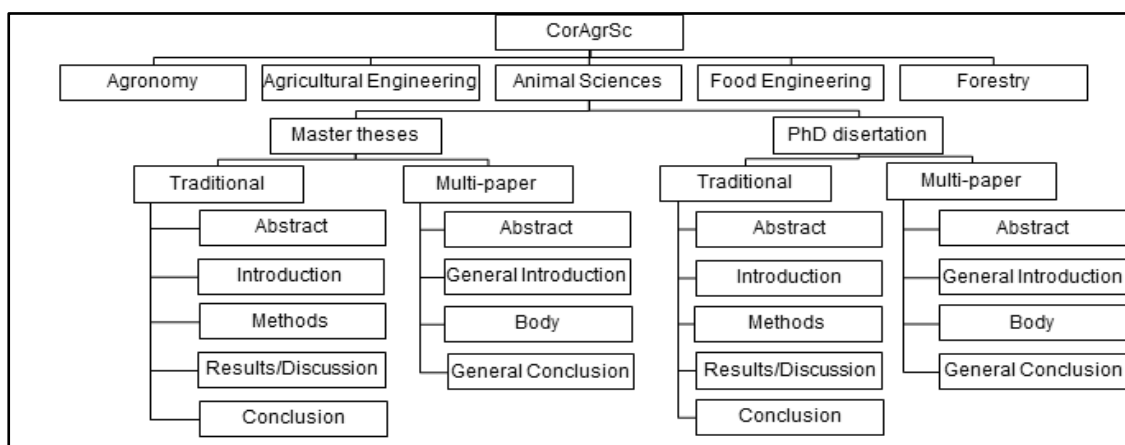
CROPS is organized into multiple subcorpora. For each subcorpus, texts were grouped based on their textual properties and formatting styles, as follows: 1) Subject area: Animal Science, Agricultural Engineering, Agronomy, Food Engineering, and Forestry; 2) Type: master's theses or doctoral dissertations; 3) Format style: monograph or multi-paper; 4) Subsections: abstract, introduction, methods, results/discussion, conclusion, general introduction, body, and general conclusion.

As regards item 3 (format style), upon collecting texts, we identified that the selected master's theses and doctoral dissertations followed two distinct formats: monograph or multi-paper. The first type consists of a single text written as a monograph divided into separate chapters/sections that usually include, but are not limited to, the following: abstract, introduction, methods, results, discussion, and conclusion. The latter consists of a text that includes multiple research articles (ranging from two to eight) developed by the author themselves, often including a general introduction chapter that contextualizes and presents the papers, and a general conclusion that synthesizes and integrates the main findings of the papers.

Additionally, incorporating subcorpora based on individual sections/chapters (item 4) allows for a more comprehensive investigation of the data, since different sections of academic texts exhibit distinct linguistic features (see, e.g., Parkinson, 2013; Swales; Feak, 2012; Biber; Finegan, 2001). We also opted to keep the texts of each section in their entirety in order to avoid the loss of linguistic data, as suggested in Sinclair (2005).

To do so, all texts were manually segmented and the resulting files were individually revised to ensure accuracy. Texts of the monograph format were segmented according to their sections, whereas multi-paper texts were subdivided into abstract, general introduction, body, and general conclusion. In this subcorpus, the 'body' section consists of the whole collection of research papers within the theses and dissertations. Figure 2 provides an overview of our corpus structure, using the Animal Science subcorpus as an example.

Figure 2 - CROPS's structure exemplified with the Animal Science subcorpus.



Source: the authors.

2.4 Naming conventions of text files

For a neat organization of the documents, patterned naming conventions were established for each text file. Following Reppen (2022) and Dutra *et al.* (2022), abbreviations were adopted to keep file names concise while still conveying relevant details efficiently. These abbreviations and their corresponding meanings are shown in Figure 3.

Figure 3 - Abbreviations adopted for file names in CROPS.

Subject Areas	AgE	Agricultural Engineering
	Agr	Agronomy
	AnS	Animal Science
	FEg	Food Engineering
	For	Forestry
Types of texts	MTh	Master's thesis (monograph format)
	PhD	Doctoral dissertation (monograph format)
	AMTh	Master's thesis (multi-paper format)
	APhD	Doctoral dissertation (multi-paper format)
Text Sections	Abs	Abstract
	Int	Introduction or General Introduction
	Met	Methods
	RD	Results and Discussion
	Con	Conclusion or General Conclusion
	Body	Collection of research papers comprised in the multi-paper theses and dissertations

Source: the authors.

Each text file in CROPS contains up to six identification codes, as illustrated in Figure 4. The bracketed numbers refer to the following identification codes: [1] corpus name; [2] subject area (e.g. AgE for Agricultural Engineering); [3] type of text (e.g. PhD for doctoral dissertation); [4] text section (e.g. Abs for Abstract); [5] date of publication; and [6] a random sort number.

Figure 4 - Illustration of file naming conventions adopted in CROPS.

CorAgrSc.AgE.PhD.Abs.2016.022					
[1]	[2]	[3]	[4]	[5]	[6]

Source: the authors.

2.5 Corpus cleaning

Corpus cleaning refers to the task of removing extraneous material from text data, such as headers, footers, special characters, line breaks, and other non-linguistic elements that do not contribute to the actual linguistic content (Weisser, 2016). It is a fundamental step in the corpus compilation process because these unwanted items “may adversely affect the accuracy of the analytical procedures we intend to carry out” (McEnery; Brooks, 2022, p. 43).

When performed manually, corpus cleaning is a complex and time-consuming task, as it demands strict consistency and careful attention to detail. To address this challenge, our research team developed *CorpusAid*⁸, formerly *CorpusScript* (Alves, 2024), a modular Python program designed to automate corpus-cleaning tasks. The tool features a customizable processing pipeline and a user-friendly graphical interface built with PySide6. Its core processing relies on *re* for regular expressions, *spaCy* for linguistic tasks, *BeautifulSoup4* for HTML parsing, and *unicodedata* and *charset-normalizer* for robust character encoding and normalisation.

The pipeline processes both PDF and .txt files, featuring an integrated PDF-to-text conversion as its initial step. The subsequent cleaning modules apply a series of discrete transformations that users can select and customize according to their preferences, including: *i*) header, footer, and page marker removal: filters are included to remove standalone Arabic page numbers, Roman numerals, inline page indicators (e.g. Page xvi), and converter-specific delimiters (e.g. -- Page X --); *ii*) whitespace and line break handling: corrects common conversion artefacts by normalizing inconsistent spacing, removing unwanted line breaks, and fixing words split by end-of-line hyphens; *iii*) character and encoding normalization: ensures text consistency through Unicode normalization (NFKC) and the removal of diacritics (e.g., á becomes a); *iv*) specialized character set removal: provides modules to filter specific scripts and symbols, including Greek and Cyrillic characters, as well as superscript and subscript; *v*) content and markup stripping: removes HTML markup from web-sourced texts and filters scholarly citations based on common patterns; *vi*) linguistic processing: uses *spaCy*'s 'en_core_web_sm' model to perform case conversion (lowercasing), word tokenization, and optional removal of

⁸ The complete source code, documentation, and release notes for *CorpusAid* (version 0.8) are available at the official repository: <<https://github.com/jhlopesalves/CorpusAid>>. The software is distributed under the MIT License.

English stop words; and *vii*) Custom Pattern Matching: offers advanced control by allowing users to remove any user-defined sequence of characters or apply custom regular expressions for complex cases.

Notably, *CorpusAid* features robust error handling; all cleaning actions are logged, and the interface supports a full undo/redo history for each document. However, highly complex PDF layouts may cause conversion issues, requiring a final manual inspection.

3 CROPS in numbers

The corpus comprises a total of 166 texts - in the form of master's theses and doctoral dissertations - in both monograph and multi-paper formats, comprising approximately 2.4 million words. It is divided into six main subcorpora, namely Animal Science, Agricultural Engineering, Agronomy, Food Engineering, and Forestry. Table 2 shows the number of texts, words, and the average word count per text for each subcorpus (subject area) in CROPS.

Table 2 - Number of texts, words, and average word count per text for each subcorpus (subject area) in CROPS.

Subject area	Texts	Words	Average n. of words
Agricultural engineering (AgE)	16	266,635	16,664
Agronomy (Agr)	59	925,524	15,686
Animal Science (AnS)	56	730,215	13,039
Food Engineering (FEg)	27	327,081	12,114
Forestry (For)	8	148,698	18,587
TOTAL	166	2,398,153	14,447

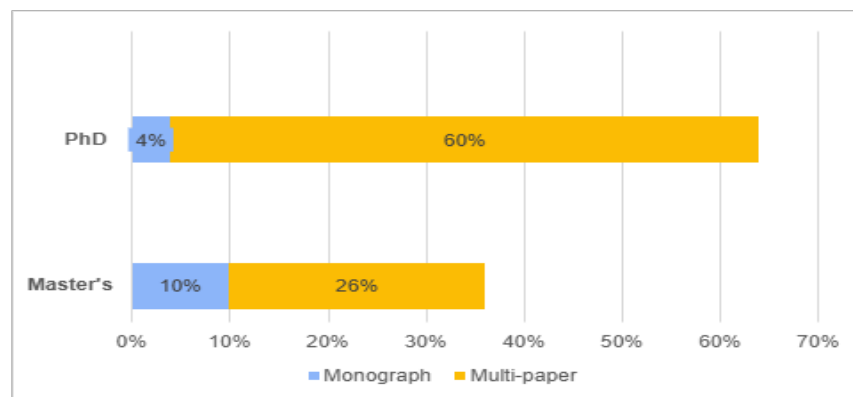
Source: the authors.

As shown in Table 2, the most productive subject areas in CROPS are Agronomy and Animal Science, which together account for nearly 70% of the texts and words in CROPS, while the Forestry subcorpus presents only 5% of the number of texts. Despite this difference, the average word count per text is fairly consistent across all subcorpora, suggesting uniformity across subject areas.

Regarding text types, CROPS includes a higher proportion of doctoral dissertations (64%) than master's theses (36%). In terms of format styles, it contains a larger number of multi-paper (86%) than monograph texts (14%), indicating a preference for the former format

in Brazilian Agrarian Sciences postgraduate programs. When total counts for text type and format style are combined, monograph doctoral dissertations are the least frequent, while multi-paper doctoral dissertations predominate, as shown in Figure 5.

Figure 5 - Distribution of texts by type and format in CROPS.



Source: the authors.

As mentioned in 2.3, the organization of texts into sections (or chapters) varies according to their formatting style. The total number of words and texts for each subject area and section is shown in Table 3 (monograph format) and Table 4 (multi-paper format).

Table 3 - Number of texts and words by subject area and section of monograph theses and dissertations in CROPS.

Section	AgE		Agr		AnS		Feg		For		Total	
	words	texts	words	texts	words	texts	words	texts	words	texts	words	texts
Abstract	795	3	2784	8	2411	6	2654	8	444	1	9088	26
Introduction	5490	3	28290	8	15283	6	31565	7	5053	1	85681	25
Methods	3626	3	8455	8	6945	6	14954	7	4674	1	38654	25
Results/Discuss.	9640	3	34496	8	15413	6	34384	7	4456	1	98389	25
Conclusion	387	2	1340	8	638	6	3130	7	643	1	6138	24
Total	19938	14	75365	40	40690	30	86687	36	15270	5	237950	125

Source: the authors.

Table 4 - Number of texts and words by subject area and section of multi-paper theses and dissertations in CROPS.

Section	AgE		Agr		AnS		Feg		For		Total	
	words	texts	words	texts	words	texts	words	texts	words	texts	words	texts
Abstract	6377	13	21416	55	23182	50	8393	19	3064	9	62432	146
Gen. Introd	13559	10	86918	51	40822	50	10974	13	12649	9	164922	133
Body	222774	13	722535	54	612519	55	217078	17	113058	9	1887964	148
Gen. Conc.	3987	5	19290	43	13002	40	3949	12	4657	8	44885	108
Total	246697	41	850159	203	689525	195	240394	61	133428	35	2160203	535

Source: the authors.

4 Final remarks

This work was carried out within the *Grupo de Estudos em Corpora Especializados e de Aprendizizes (GECEA)*, as part of the project “*Escrita acadêmica em língua inglesa nas Ciências Agrárias: necessidades e insumos para aplicações pedagógicas*”. The project aims to describe and analyze academic writing in English within the field of Agrarian Sciences, with the goal of developing academic writing instructional materials. Therefore, CROPS results from a collaborative effort that involved numerous participants, including professors, researchers, undergraduate, and postgraduate students from different universities.

In constructing this specialized corpus, several challenges related to its design and compilation were addressed by relying on rigorous theoretical and methodological criteria. Technical issues in corpus cleaning were resolved by automating the process through the implementation of a new computer program developed by our team.

CROPS (Corpus of Agrarian Sciences Postgraduate Studies) constitutes a valuable resource for linguistic research, broadly reflecting the written practices of postgraduate students in Agrarian Sciences in Brazil. In the near future, we intend to make CROPS freely available for use by other researchers.

Lastly, the compilation process faced practical constraints, mainly related to the availability of texts written in English. In some cases, institutional repositories were incomplete or even temporarily offline, which hindered access to additional theses and dissertations. Due

to time constraints, it was not possible to revisit those databases or to double-check the availability of further texts after the initial collection. For future work, we intend to expand CROPS by incorporating more texts from other repositories, such as the CAPES *Catálogo de Teses e Dissertações*.

Acknowledgements

The authors would like to thank FAPEMIG for funding this project (n. APQ-01173-22), as well as CAPES for the research grant n. 88887.939578/2024-00.

References

ÄDEL, Annelie. Corpus compilation. *In*: PAQUOT, Magali; GRIES, Stefan (Ed.). **A practical handbook of corpus linguistics**. Springer Nature, 2020. p. 3-24.

ALVES, Jhonatan Henrique Lopes, et al. CorpuScript: an automated text-cleaning tool for corpus linguistics. *In*: XVI Encontro de Linguística de Corpus e XIII Escola Brasileira de Linguística Computacional, 2024, Brasília. **Anais eletrônicos** [...]. Brasília: UnB. October, 2024. Available at: <https://www.elc-ebralc.net.br/anais>.

BIBER, Douglas. Representativeness in corpus design. **Literary and linguistic computing**, v. 8, n. 4, p. 243-257, 1993.

BIBER, Douglas. **University language**. Amsterdam: John Benjamins, 2006.

BIBER, Douglas; FINEGAN, Edward. Intra-textual variation within medical research articles. *In*: CONRAD, Susan BIBER, Douglas (ed.). **Variation in English: Multi-dimensional Studies**. London: Longman, 2001. p 108-123.

BIBER, Douglas; GRAY, Bethany. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. **Journal of English for Academic Purposes**, v. 9, n. 1, p. 2-20, 2010.

CAPES. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Catálogo de Teses e Dissertações - Brasil [1987-2012]. *In*: **Dados Abertos CAPES**. 2024. Available at: <https://dadosabertos.capes.gov.br/dataset/1987-a-2012-catalogo-de-teses-e-dissertacoes-brasil>. Retrieved 2 Oct, 2025.

CAPES. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. **Catálogo de Teses e Dissertações - Brasil [2021 a 2024]**. *In*: Dados Abertos CAPES. 2025a. Available at: <https://dadosabertos.capes.gov.br/dataset/2021-a-2024-catalogo-de-teses-e-dissertacoes-brasil>. Retrieved 2 Oct, 2025.

CAPES. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. **Plataforma Sucupira: Dados Cadastrais do Programa**. 2025b. Available at: <https://sucupira->

legado.capes.gov.br/sucupira/public/consultas/coleta/programa/listaPrograma.jsf. Retrieved 24 Sep., 2025.

COLLINS, Luke. **Corpus linguistics for online communication: A guide for research**. Chapter 2: Designing and building corpora. New York: Routledge, 2019. p. 29-50.

DUTRA, Deise Prina; ORFANÒ, Bárbara Malveira; GUEDES, Annallena de Souza; ALVES, Jessica Ceritello; FEKETE, João Gabriel. The learner corpus path: a worthwhile methodological challenge. **D.E.L.T.A.**, v. 38, n. 2, p. 1-24, 2022. <https://dx.doi.org/10.1590/1678-460X202238249731>.

FLOWERDEW, John; HABIBIE, Pejman. **Introducing English for research publication purposes**. New York: Routledge, 2022.

GRAY, Bethany. On the complexity of academic writing: Disciplinary variation and structural complexity. In: CORTES, Viviana.; CSOMAY, Eniko. (ed.). **Corpus-based research in applied linguistics: Studies in honor of Doug Biber**. John Benjamins, 2015. p. 49-78.

HYLAND, Ken. **Disciplinary discourses: Social interactions**. The University of Michigan Press, 2004.

HYLAND, Ken. Corpora and academic discourse. In: HYLAND, Ken; HANDFORD, Michael; CHAU, Meng Huat. **Corpus applications in applied linguistics**. 2012, p. 30-46, 2012.

HYLAND, Ken. **Academic publishing: Issues and challenges in the construction of knowledge**. Oxford: Oxford University Press, 2016.

HYLAND, Ken; BONDI, Marina. Chapter 1: Introduction. In: HYLAND, Ken; BONDI, Marina (ed.). **Academic discourse across disciplines**. S.l.: Peter Lang, 2006.

LILLIS, T.; CURRY, M. J. **Academic Writing in a Global Context: The Politics and Practices of Publishing in English**. London: Routledge, 2010.

MCENERY, Tony; BROOKES, Gavin. Building a written corpus: what are the basics? In: O'KEEFFE, Anne; MCCARTHY, Michael J. (ed.). **The Routledge Handbook of Corpus Linguistics**. 2nd Edition. Routledge, 2022. p. 35-47.

NASSI-CALÒ, Lilian. Estudo aponta que artigos publicados em inglês atraem mais citações [online]. **SciELO em Perspectiva**, 2016.

PARKINSON, Jean. Representing own and other voices in social science research articles. **International Journal of Corpus Linguistics**, v. 18, n. 2, p. 199-228, 2013.

REPPEN, Randi. Building a corpus: what are key considerations? In: O'KEEFFE, Anne; MCCARTHY, Michael J. (ed.). **The Routledge Handbook of Corpus Linguistics**. 2nd Edition. Routledge, 2022. p. 13-20.

SHAMSI, Ahmad Fawzi; OSAM, Ulker Vanci. Challenges and support in article publication: perspectives of non-native English speaking doctoral students in a “publish or no degree” context. **Sage Open**, v. 12, n. 2, 2022.

SHI, Huimin; WANNARUK, Anchalee. Rhetorical Structure of Research Articles in Agricultural Science. **English Language Teaching**, v. 7, n. 8, p. 1-13, 2014.

SINCLAIR, John. Corpus and Text. *In*: WYNNE, M (ed.). **Developing Linguistic Corpora: A Guide to Good Practice**. Oxford: Oxbow Books, 2005.

SWALES, John, Malcom; FEAK, Christine, Bell. **Academic Writing for Graduate Students: Essential Tasks and Skills**. 3rd ed., Michigan ELT, 2012.

WEISSER, Martin. **Practical Corpus Linguistics: An Introduction to Corpus-based Language Analysis**. Malden, MA: Wiley-Blackwell, 2016.

APPENDIX 1 – List of institutions in Brazil (CAPES, 2025b) offering Graduate Programs (Master's/PhD) in Agrarian Sciences within the five subject areas comprised in CROPS (Animal Science, Agricultural Engineering, Agronomy, Food Engineering, and Forestry), divided according to Brazil's geographical regions. Institutions represented in CROPS are marked with an asterisk (*).

Central-West

- INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO
- UNIVERSIDADE ANHANGUERA
- UNIVERSIDADE DE BRASÍLIA*
- UNIVERSIDADE DE RIO VERDE
- UNIVERSIDADE ESTADUAL DE GOIÁS
- UNIVERSIDADE ESTADUAL DE MATO GROSSO DO SUL
- UNIVERSIDADE FEDERAL DA GRANDE DOURADOS
- UNIVERSIDADE FEDERAL DE GOIÁS
- UNIVERSIDADE FEDERAL DE JATAÍ
- UNIVERSIDADE FEDERAL DE MATO GROSSO
- UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
- UNIVERSIDADE FEDERAL DE RONDONÓPOLIS

Northeast

- EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA ALGODÃO
- EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA MANDIOCA E FRUTICULTURA
- UNIVERSIDADE DO ESTADO DA BAHIA
- UNIVERSIDADE ESTADUAL DA PARAÍBA
- UNIVERSIDADE ESTADUAL DA REGIÃO TOCANTINA DO MARANHÃO
- UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
- UNIVERSIDADE ESTADUAL DE SANTA CRUZ
- UNIVERSIDADE ESTADUAL DO MARANHÃO
- UNIVERSIDADE ESTADUAL DO SUDOESTE DA BAHIA
- UNIVERSIDADE ESTADUAL VALE DO ACARAÚ
- UNIVERSIDADE FEDERAL DA BAHIA
- UNIVERSIDADE FEDERAL DA PARAÍBA (including CAMPUS AREIA, CAMPUS BANANEIRAS)
- UNIVERSIDADE FEDERAL DE ALAGOAS*
- UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
- UNIVERSIDADE FEDERAL DE SERGIPE
- UNIVERSIDADE FEDERAL DO AGRESTE DE PERNAMBUCO
- UNIVERSIDADE FEDERAL DO CEARÁ*
- UNIVERSIDADE FEDERAL DO MARANHÃO
- UNIVERSIDADE FEDERAL DO PIAUÍ
- UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
- UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
- UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO
- UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO*
- UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO*

North

- INSTITUTO NACIONAL DE PESQUISAS DA AMAZÔNIA
- UNIVERSIDADE FEDERAL DE RORAIMA
- UNIVERSIDADE FEDERAL DO ACRE
- UNIVERSIDADE FEDERAL DO AMAPÁ
- UNIVERSIDADE FEDERAL DO NORTE DO TOCANTINS
- UNIVERSIDADE FEDERAL DO OESTE DO PARÁ*
- UNIVERSIDADE FEDERAL DO PARÁ
- UNIVERSIDADE FEDERAL DO TOCANTINS
- UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA*
- UNIVERSIDADE NILTON LINS

Southeast

- FUNDO DE DEFESA DA CITRICULTURA
- INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO ESPÍRITO SANTO
- INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO SUDESTE DE MINAS GERAIS
- INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO TRIÂNGULO MINEIRO
- SECRETARIA DE AGRICULTURA E ABASTECIMENTO DO ESTADO DE SÃO PAULO - INSTITUTO AGRONÔMICO DE CAMPINAS
- SECRETARIA DE AGRICULTURA E ABASTECIMENTO DO ESTADO DE SÃO PAULO - INSTITUTO DE ZOOTECNIA
- UNIVERSIDADE BRASIL
- UNIVERSIDADE DE SÃO PAULO/CENTRO DE ENERGIA NUCLEAR NA AGRICULTURA/ESCOLA SUPERIOR DE AGRICULTURA LUIZ DE QUEIROZ*
- UNIVERSIDADE DO OESTE PAULISTA
- UNIVERSIDADE ESTADUAL DE CAMPINAS
- UNIVERSIDADE ESTADUAL DE MONTES CLAROS
- UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO*
- UNIVERSIDADE ESTADUAL PAULISTA JÚLIO DE MESQUITA FILHO (CAMPI ARARAQUARA/BOTUCATU/ILHA SOLTEIRA/JABOTICABAL/SÃO JOSÉ DO RIO PRETO)
- UNIVERSIDADE FEDERAL DE ALFENAS
- UNIVERSIDADE FEDERAL DE LAVRAS*
- UNIVERSIDADE FEDERAL DE MINAS GERAIS*
- UNIVERSIDADE FEDERAL DE SÃO CARLOS
- UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI
- UNIVERSIDADE FEDERAL DE UBERLÂNDIA*
- UNIVERSIDADE FEDERAL DE VIÇOSA*
- UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO*
- UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
- UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
- UNIVERSIDADE PROFESSOR EDSON ANTÔNIO VELANO

South

- FUNDAÇÃO UNIVERSIDADE DO ESTADO DE SANTA CATARINA
- UNIVERSIDADE DE PASSO FUNDO
- UNIVERSIDADE ESTADUAL DE MARINGÁ*
- UNIVERSIDADE ESTADUAL DE PONTA GROSSA
- UNIVERSIDADE ESTADUAL DO CENTRO-OESTE
- UNIVERSIDADE ESTADUAL DO NORTE DO PARANÁ
- UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ
- UNIVERSIDADE FEDERAL DE PELOTAS
- UNIVERSIDADE FEDERAL DE SANTA CATARINA (including CAMPUS CURITIBANOS*)
- UNIVERSIDADE FEDERAL DE SANTA MARIA*
- UNIVERSIDADE FEDERAL DO PARANÁ*
- UNIVERSIDADE FEDERAL DO RIO GRANDE
- UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL*
- UNIVERSIDADE PARANAENSE
- UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ*