THE RELEVANCE OF AI-POWERED TOOLS IN THE ENGLISH ACADEMIC WRITING OF BRAZILIAN SCHOLARS IN APPLIED LINGUISTICS AND IN THE VISUALISATION OF RESEARCH DATA

**Rogério Yamada**

**ABSTRACT**

The recent advent of new-generation Artificial Intelligence (AI) powered tools, such as ChatGPT and Google Bard, is bound to enable a series of possibilities of content creation, including Academic Writing. In spite of its current limitations, massive investment in the technology from leading AI enterprises and the world’s ever growing demand for productivity are expected to increase the adoption of the technology to a point where its use is not going to be an option. This study aims to developing a better understanding about the influence AI-powered tools might have on texts composed by Brazilian Applied Linguistics English for Research Publication Purposes (ERPP) authors of a Higher Education institution's graduate programme via a *corpus*-based approach known as Multi-Dimensional Analysis (MDA) (BERBER SARDINHA, 2000, p. 100) over a period of three terms (from the second term of 2023 to the second term of 2024). Its insights are considered to be critical for designing strategies for better preparing Brazilian ERPP researchers and students for the highly demanding requirements of mainstream English-only journals and fostering their participation in the global research community. Exploring data visualisation technologies, the research project is going to create computer-generated visual representations of its outcomes in order to enable the assessment of complex relationships between variables and also to streamline the interchange of research data. A preliminary plan aimed to complete the research project before the end of the first term of 2025 is proposed.

**Keywords**: Artificial Intelligence; *Corpus* Linguistics; Data Visualisation; English for Research Publication Purposes; Multi-Dimensional Analysis

# 

# INTRODUCTION

The adoption of Artificial Intelligence (AI) technologies has recently reached a new level with the release of large language models capable of assisting humans with content creation.

Considering the research question ‘What is the relevance of the use of AI-powered tools for Brazilian Applied Linguistics English for Research Publication Purposes (ERPP) authors?’, the present study aims to employ a *Corpus* Linguistics approach known as Multi-Dimensional Analysis (MDA) (BERBER SARDINHA, 2000, p. 100) in the context of a group of Brazilian Applied Linguistics ERPP writers in a Higher Education institution's graduate programme.

The insights produced by this proposed research project are considered to be essential for enhancing the participation of Brazilian scholars in the global research community especially in Humanities. Furthermore, they could also be a key asset to inform new pedagogical approaches for better preparing students of ERPP for the highly demanding requirements of the mainstream English-only journals.

The project is going to contribute to the development of innovative initiatives for AI-driven data visualisation within the scope of the CNPq[[1]](#footnote-0)’s institutional project ‘*Inteligência artificial na pesquisa em linguagem e discurso: Galerias multimodais e multidimensionais de visualização*’[[2]](#footnote-1) seeking to explore computer-generated visual representations of data. It will also add to the creation of data visualisation galleries as specified in the aforementioned institutional project.

In the next section, the theoretical basis is going to be outlined.

# LITERATURE REVIEW

Despite the continuing advance of the English Language (EL) as *lingua franca* in a multitude of contexts of the human activity, off-network English as an International Language (EIL) scholars continue to be strikingly disadvantaged in the competition for publication in mainstream English-only journals (BELCHER, 2007), which stifles their participation in the global research community.

Belcher (2007) questions to which degree the rules of the Anglosaxon publication system are responsible for this monocultural trend. In her study, Belcher (2007) argues that EIL academics who eventually have their papers published have to undergo a cumbersome process of revision and resubmission in a "story of reviewer patience and author persistence".

In a piece of research dedicated to study how multilingual Brazilian writers reacted to the demands to comply with the mainstream linguistic and cultural conventions, Monteiro and Hirano (2020) reported that "scholars from the [H]umanities and [S]ocial [S]ciences publish substantially less in international journals, reporting challenges particularly with language".

In that respect, the recent emergence of AI-powered large language models, such as ChatGPT and Google Bard, represents an opportunity to users of English for Research Publication Purposes (ERPP) for their academic production. Despite its currently known limitations, “[g]enerative AI tools can produce a wide variety of credible writing in seconds, then respond to criticism to make the writing more fit for purpose.” (WHAT IS CHATGPT, DALL-E, AND GENERATIVE AI? | MCKINSEY, 2023).

Judging by the massive investment the development of the technology has been receiving from leading AI enterprises and as its use becomes widespread, it can be assumed that those AI-based tools will soon become reliable enough to influence a large volume and variety of text genres produced in human activities. The world’s ever growing demand for productivity is unlikely to pose the use of this technology as an option.

According to an article by McKinsey (WHAT IS CHATGPT, DALL-E, AND GENERATIVE AI? | MCKINSEY, 2023), generative AI algorithms are trained on massive amounts of data, which makes their responses appear highly creative and even more lifelike with the inclusion of random elements.

Considering the specific case of ERPP academics, their writing would probably become a combination of their own creation and excerpts composed with an AI-powered tool (please refer to [APPENDIX A - Writing review via ChatGPT sample](#_xru4lg3snxk5)). Therefore, from a linguistic point of view, the use of AI-powered tools may introduce distinctive language features that could lead to a register variation in their pieces of text.

A register variation study via MDA is expected to produce insights on the nature of variation introduced in the writers’ Academic Writing. According to Berber Sardinha (2000, p. 100), the approach known as MDA, originally proposed by Biber in 1988, allows for the analysis of a large volume of texts of a certain register, identifying patterns of co-occurring grammatical features (also known as functional dimensions) that characterise the variation of that register.

Biber (2009, p. 823) argues that registers can be defined by their situational (e.g. informality, orality, etc.) and linguistic characteristics:

Although registers are defined in situational terms, they can also be compared with respect to their linguistic characteristics: the study of register variation. Register variation is inherent in human language: a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar reflecting a range of situational factors. (BIBER, 2009, p. 823)

A complete MDA follows eight methodological steps (BIBER, 2009, p. 825-826):

1. An appropriate *corpus* is designed based on previous research and analysis. Texts are collected, transcribed (in the case of spoken texts), and input into the computer;
2. Research is conducted to identify the linguistic features to be included in the analysis, together with functional associations of the linguistic features;
3. Computer programs are developed for automated grammatical analysis, to identify or ‘tag’ all relevant linguistic features in texts;
4. The entire *corpus* of texts is tagged automatically by computer, and all texts are edited interactively to ensure that the linguistic features are accurately identified;
5. Additional computer programs are developed and run to compute normed frequency counts of each linguistic feature in each text of the *corpus*;
6. The co-occurrence patterns among linguistic features are identified through a factor analysis of the frequency counts;
7. The ‘factors’ from the factor analysis are interpreted functionally as underlying dimensions of variation;
8. Dimension scores for each text are computed; the mean dimension scores for each register are then compared to analyse the salient linguistic similarities and differences among registers.

In a study about discourse of academia from a multidimensional perspective, Berber Sardinha (2020, p. 302-303) proposed the use of a derivation of MDA known as lexical MDA which

Unlike a “traditional” MD Analysis, which uses mostly grammatical features and whose general goal is to identify functional dimensions of register variation, a lexical MD Analysis uses lexical units (words, n-grams, collocations) to identify the dimensions. (BERBER SARDINHA, 2020, p. 302-303)

According to Berber Sardinha (2000, p. 103-104), MDA, due to its cumulative nature, is an appropriate approach for analyses of *corpora* that grow incrementally.

Research in *Corpus* Linguistics requires judicious design and preparation of *corpora*. Sanchez (1995, p. 8-9 *apud* BERBER SARDINHA, 2018, p. 338) defines a *corpus* as

A set of linguistic data (pertaining to the oral or written use of the language, or both), systematised according to certain criteria, sufficiently extensive in breadth and depth, so that they are representative of the totality of the linguistic use or of any of its scopes, arranged in such a way that they can be processed by computer, in order to provide various and useful results for the description and analysis. (SANCHEZ, 1995, p. 8-9 *apud* BERBER SARDINHA, 2018, p. 338, author’s translation)[[3]](#footnote-2)

Berber Sardinha (2018, p. 338-339) enumerates four prerequisites for preparing a *corpus*:

1. The *corpus* should consist of authentic texts in natural language, not specifically produced for linguistic research nor artificially produced by computer programmes;
2. If the texts were produced by non-native speakers, the *corpus* should be qualified as a learner *corpus*;
3. The *corpus* should be fit for purpose in the sense that it should be designed according to the requirements of its corresponding research;
4. The *corpus* should be representative. Although this is a requisite that cannot be objectively determined, a larger *corpus* is generally more representative than a smaller one because it may contain more instances of low frequency linguistic features (BERBER SARDINHA, 2018, p. 343).

According to Berber Sardinha (2018, p. 345), it should be the researcher’s responsibility to demonstrate how representative and adequate a *corpus* is for their study.

The size of a *corpus* is defined by three dimensions (BERBER SARDINHA, 2018, p. 344):

1. Number of words;
2. Number of texts;
3. Number of registers.

Having defined the theoretical basis of this study, the methodological procedures are going to be detailed in the next section.

# METHODS

The study is going to be designed to test the following hypothesis: The revision via an AI-powered tool (ChatGPT or Google Bard) of papers authored by Brazilian ERPP scholars in the area of Applied Linguistics should present a magnitude of register variation that should be larger than the magnitude obtained from applying the same AI-powered tool on papers of the same area that were recognised by mainstream journal reviewers as ready for publication.

Therefore, the relevance of the AI-powered ​​tool will be as great as the register variation obtained by its use when compared with the register variation obtained in a paper that was accepted by one of the mainstream journals.

To that end, the research project is going to consider the following procedures:

1. A *corpus* of Applied Linguistics Published Papers (ALPP) is going to be designed and have its dimensions analysed via MDA;
2. The texts of the ALPP *corpus* are going to be reviewed via an AI-powered tool and used to design a *corpus* of AI-revised Applied Linguistics Published Papers (AI-ALPP). Its dimensions are going to be analysed via MDA;
3. The magnitude of register variation verified in steps 1 and 2 should be named ‘AI-assisted Variation - Published Reference’ (AI-VPR);
4. A *corpus* of Applied Linguistics Brazilian-authored Papers (ALBP) is going to be designed and have its dimensions analysed via MDA;
5. The texts of the ALBP *corpus* are going to be reviewed via an AI-powered tool and used to design a *corpus* of AI-revised Applied Linguistics Brazilian-authored Papers (AI-ALBP). Its dimensions are going to be analysed via MDA;
6. The magnitude of register variation verified in steps 4 and 5 should be named ‘AI-assisted Variation - Brazilian-authored Reference’ (AI-VBR);
7. AI-VPR and AI-VBR are going to be analysed for interpretation of results.

The four *corpora* should be designed according to the high-level requirements depicted in Table 1.

Table 1 - *Corpora* design

|  | **Requirement** | **Description** |
| --- | --- | --- |
| 1 | Source | ALPP: Applied Linguistics published papers  AI-ALPP: AI-revised Applied Linguistics published papers  ALBP: Applied Linguistics Brazilian-authored papers  AI-ALBP: AI-revised Applied Linguistics Brazilian-authored papers |
| 2 | Purpose | Study of the register variation |
| 3 | Mode | Written |
| 4 | Time | Contemporary |
| 5 | Selection | Sample *corpus* |
| 6 | Content | Specialised: English for Research Publication Purposes |
| 7 | Authorship | ALPP: ERPP writers  ALBP: Brazilian Applied Linguistics ERPP writers in a Higher Education institution's graduate programme |

Source: Author’s source based on Berber Sardinha (2018, p. 338-342)

Even though a *corpus* should consist of authentic texts in natural language, not produced by computer programmes, texts produced by ERPP writers assisted by AI-powered tools are going to be considered authentic for the purpose of this study.

The ALPP and ALBP *corpora* are going to be sourced and designed to meet the required level of representativeness and adequacy according to the criteria to be determined and validated with the supervisor.

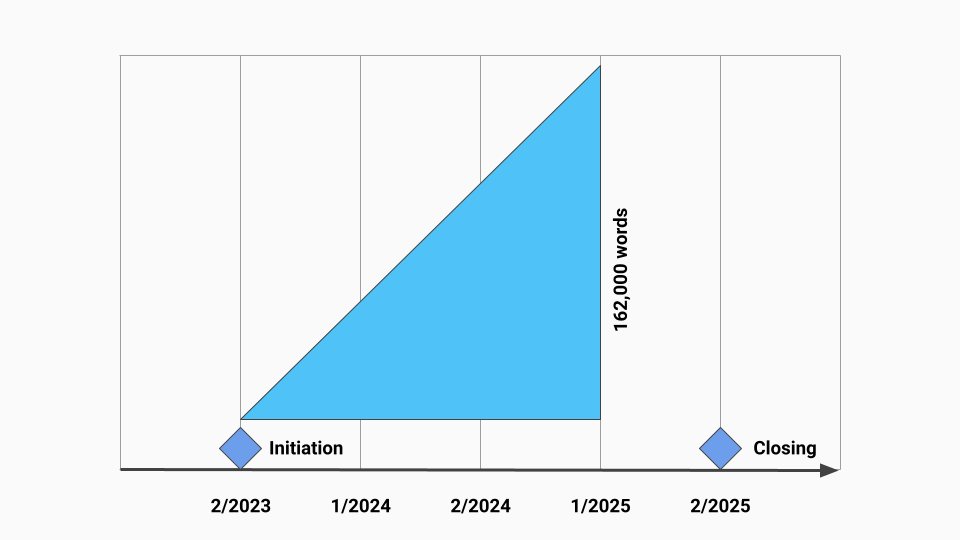
A possible strategy for the ALBP *corpus* would be to incrementally build it from the second term of 2023 to the second term of 2024. It is expected that there is going to be one group of 20 ERPP writers per term that will compose papers with 1,350 words on average, resulting in a final word count estimated at 162,000 words as shown in Table 2. This strategy might be reconsidered according to the mentioned criteria of *corpora* representativeness and adequacy.

Table 2 - Estimated ALBP *corpus* size

| **Source** | **Interval** | **Texts** | **Word count** |
| --- | --- | --- | --- |
| ERPP writers 1 | 2/2023 | 40 | 54,000 |
| ERPP writers 2 | 1/2024 | 40 | 54,000 |
| ERPP writers 3 | 2/2024 | 40 | 54,000 |
|  |  | 120 | 162,000 |

Source: Author’s source

Figure 1 - ALBP *corpus* incremental growth



Source: Author’s source

In principle, the study is going to employ the traditional (functional) MDA. This assumption should be validated at the establishment of the research project.

## DATA VISUALISATION

The research project is going to explore data visualisation technologies to create computer-generated visual representations of its outcomes in order to enable the assessment of complex relationships between variables and also to streamline the interchange of research data.

A few readily available open source libraries such as D3.js, ECharts, Vega and deck.gl may be employed for this purpose (WHAT IS DATA VISUALIZATION? | IBM, 2023).

Details about how the research project is expected to be executed are the subject of the next section.

# PROJECT SCHEDULE

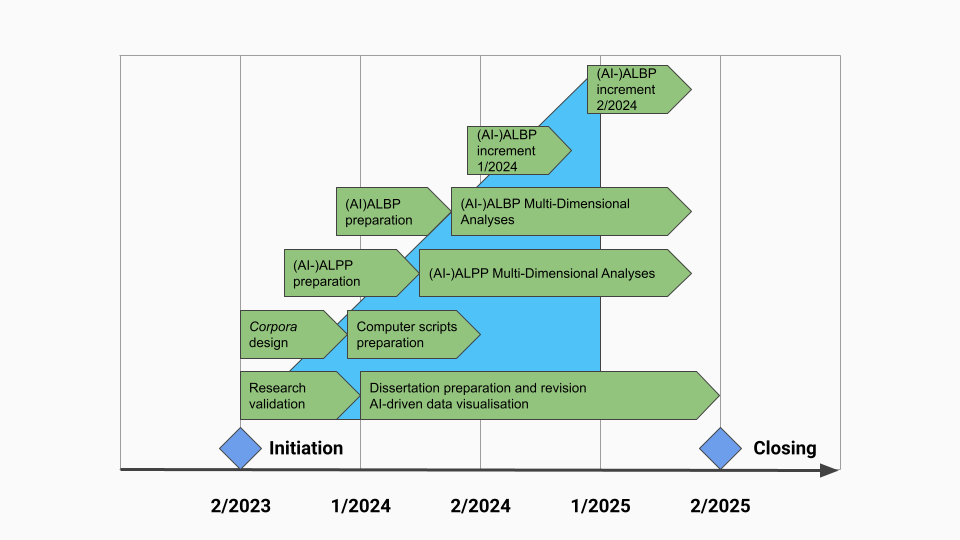
The project is expected to start in the second term of 2023 with an initial validation with the supervisor. The first task should be the *corpora* design.

The ALBP *corpus* is going to be designed and populated initially with the texts produced in the second term of 2023. As the project progresses it is expected to incrementally grow.

The required computer scripts are going to be prepared in parallel. The Multi-Dimensional Analyses are going to be performed, refined and expanded as the project progresses according to Figure 2.

The project is expected to be concluded by the end of the first term of 2025 with the presentation of the final revision of the dissertation.

Figure 2 - Project schedule with the ALBP *corpus* growth in the background



Source: Author’s source

**REFERENCES**

BELCHER, Diane D. Seeking acceptance in an English-only research world. **Journal of Second Language Writing**, [*s. l.*], vol. 16, no. 1, p. 1–22, 2007.

BERBER SARDINHA, Tony. Análise multidimensional. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, [*s. l.*], vol. 16, no. 1, p. 99–127, 2000.

BERBER SARDINHA, Tony. Discourse of academia from a multidimensional perspective. *In*: FRIGINAL, Eric; HARDY, Jack A. (eds.). **The Routledge handbook of corpus approaches to discourse analysis**. London; New York: Routledge, 2020. (Routledge handbooks in applied linguistics).

BERBER SARDINHA, Tony. Linguística de Corpus: histórico e problemática. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, [*s. l.*], vol. 16, no. 2, p. 323–367, 2018.

BIBER, Douglas. Multi-dimensional approaches. *In*: LÜDELING, Anke; KYTÖ, Merja (eds.). **Corpus linguistics: an international handbook**. Berlin; New York: Walter de Gruyter, 2009. (Handbooks of Linguistics and Communication Science). vol. 2, p. 822–855.

MONTEIRO, Kátia; HIRANO, Eliana. A periphery inside a semi-periphery: The uneven participation of Brazilian scholars in the international community. **English for Specific Purposes**, [*s. l.*], vol. 58, p. 15–29, 2020.

WHAT IS CHATGPT, DALL-E, AND GENERATIVE AI? | MCKINSEY. [*S. l.*], 2023. Available at:<https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai>. Accessed: 17 Apr. 2023.

WHAT IS DATA VISUALIZATION? | IBM. [*S. l.*], [*s. d.*]. Available at:<https://www.ibm.com/topics/data-visualization>. Accessed: 7 Jun. 2023.

# 

# APPENDIX A - Writing review via ChatGPT sample

Date of enquiry: 8th May, 2023

Author: Dear ChatGPT, I am reviewing a research article. Would it be possible for you to improve the writing of certain passages considering the generally accepted standards of English for Academic Purposes?

ChatGPT: Of course! I would be happy to assist you with improving the writing of certain passages in the research article. Please provide the specific passages you would like me to work on and any specific requirements or guidelines you would like me to follow.

Author: Although registers are defined in situational terms, they can also be compared with respect to their linguistic characteristics: the study of register variation. Register variation is inherent in human language: a single speaker will make systematic choices in pronunciation, morphology, word choice, and grammar reflecting a range of situational factors. (BIBER, 2009, p. 823)

ChatGPT: Here is a possible revision of the passage that adheres to academic writing standards:

Register variation is a fundamental aspect of human language and can be analyzed by comparing linguistic characteristics across different registers. Registers, defined in terms of situational factors, exhibit systematic choices made by speakers in pronunciation, morphology, word choice, and grammar. These choices reflect the specific situational factors influencing language use. By studying register variation, researchers gain insights into the ways in which language adapts and conforms to different communicative contexts (BIBER, 2009, p. 823).

1. Conselho Nacional de Desenvolvimento Científico e Tecnológico [↑](#footnote-ref-0)
2. ‘Artificial intelligence in language and discourse research: Multimodal and multidimensional visualisation galleries’ [↑](#footnote-ref-1)
3. Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. [↑](#footnote-ref-2)