KEY-FEATURE ANALYSIS FOR DUMMIES: A COMPREHENSIVE STEP-BY-STEP GUIDE FOR NOVICE RESEARCHERS

Análise de Atributos-Chave para Iniciantes: Um Guia Passo a Passo para Pesquisadores em Formação

DOI: 10.14393/LL63-v41S-2026-12

Carolina Bohórquez Grondona*

Luciana Aguiar de Oliveira**

ABSTRACT: Corpus linguistics enriches academic writing instruction by comparing different text registers. This study presents a detailed guide to Key-Feature Analysis (KFA), a powerful methodology designed to identify both lexico-grammatical differences and similarities across texts. Recognizing that many linguistics students are deterred by complex statistics, this paper aims to make KFA more accessible. It provides a comprehensible explanation of the methodology, its calculations, and relevant software. To demonstrate its application, the study uses KFA to compare the introductions of academic articles and dissertations, written in English by academics and Brazilian graduate students, respectively. Key findings reveal that articles feature significantly more split auxiliaries and infinitives, which enables movements of persuasion and emphasis, while dissertations predominantly use demonstrative pronouns, which reveal a necessity of reference connections. Ultimately, this work seeks to empower students to confidently apply KFA in their own research.

KEYWORDS: Register Analysis. Academic Writing. Key-feature Analysis. Statistics for novice researchers. Academic Writing Introductions.

RESUMO: A linguística de corpus enriquece o ensino da escrita acadêmica ao possibilitar comparações entre diferentes registros textuais. Este estudo apresenta um guia detalhado da Análise de *Key-Features (KFA*), metodologia eficaz para identificar semelhanças e diferenças léxico-gramaticais entre textos. Considerando que muitos estudantes de Linguística se sentem intimidados por estatísticas complexas, este trabalho busca tornar a KFA mais acessível. São oferecidas explicações sobre a metodologia, seus cálculos e softwares utilizados. Para exemplificar sua aplicação, a KFA é empregada na comparação entre introduções de artigos acadêmicos e teses, escritos em inglês, por acadêmicos e por doutorandos brasileiros, respectivamente. Resultados mostram que introduções dos artigos apresentam uso predominante de *Split auxiliaries and infinitives* (viabilizando movimentos de persuasão e ênfase), enquanto as introduções das teses apresentam uso predominante de *Demonstrative Pronouns* (revelando necessidade de referenciação). O principal objetivo deste trabalho é incentivar estudantes a aplicarem a metodologia com confiança em suas próprias pesquisas.

PALAVRAS-CHAVE: Análise de registro. Escrita acadêmica. Key-feature Analysis. Estatística para iniciantes. Introduções acadêmicas.

* Mestre em Estudos Linguísticos. Universidade Federal de Minas Gerais. ORCID: 0009-0007-7483-2933. E-mail para contato: carolinaboho1983(AT)gmail.com.

** Mestre em Estudos Linguísticos. Universidade Federal de Minas Gerais e Universidade do Estado de Minas Gerais. ORCID: 0009-0007-3818-2611. E-mail para contato: luciana.oliveira(AT)uemg.br.

1 Introduction

Research in the fields of corpus linguistics (LC) and Academic Writing has proven to be effective in generating results that significantly benefit the teaching of writing to university students (Biber; Conrad; Cortes, 2004; Flowerdew, 2015; O'keeee; Mccarthy; Carter, 2007; Swales; Freak, 2012; Tribble; Wingate, 2013). These advancements are primarily achieved through studies that compare linguistic features of academic discourse with those of other registers. Such comparisons enable the identification and description of both differences and similarities between registers (Biber; Conrad, 2019), providing a solid foundation for informing academic writing pedagogy.

Within this objective of assisting academic writing teaching through register analysis, there are three main methodological routes to be followed, which take the text as their starting point: Multidimensional analysis (MDA) (Biber, 1988), Key-word Analysis (Scott, 1997) and Key-feature Analysis (KFA) (Egbert; Biber, 2023). Key-feature analysis, which is the one explored in this study, is the most recent and arguably more straight-forward when compared to MDA. The latter was created to detect register variation while KFA can work well to detect differences between closely related registers or texts within the same register. The KFA methodology was designed to identify functionally relevant linguistic items through the application of statistical procedures and it is classified as an exploratory analytical technique (Egbert; Biber, 2023).

Although the statistics used in KFA may be considered less complex than those used in MDA, for instance, the detailing of the former can set the basic ground for understanding what is behind the numbers. This might be encouraging, particularly for novice researchers and students in the humanities who frequently experience insecurity and anxiety when encountering complex mathematical calculations and statistical analyses in scholarly literature (Ballo; Pauli; Worrell, 2016; Chew; Dillon, 2014; Murtonen; Lehtinen, 2003). Such challenges can act as barriers, hindering their ability to fully understand, replicate, or apply these methodologies in their own research.

A study by Silva et al. (2002) underscores this issue, revealing that humanities students often exhibit negative attitudes toward mathematics and statistics. The study emphasizes the importance of creating supportive academic environments to facilitate students' comprehension of these disciplines. Furthermore, it highlights a positive correlation between

students' understanding of foundational concepts in mathematics and statistics and their willingness to engage with these subjects. In other words, as students gain proficiency in basic principles, their confidence and interest in these exact sciences increase, fostering a more positive academic experience.

Given that corpus linguistics is inherently tied to empirical and numerical data, statistics plays a crucial role in enabling efficient quantitative analysis. Such analysis can reveal patterns and variations across different registers, offering valuable insights into linguistic phenomena (Brezina, 2018). The present study seeks, therefore, to contribute to corpus linguistics, mainly register studies, by providing a step-by-step guide designed to help linguistics students engage more confidently with statistical methods. More specifically, the guide focuses on the application of KFA aiming at enhancing their research capabilities.

This article, therefore, offers a detailed exploration of the methodology outlined in Egbert and Biber's (2023) study. It provides clear and accessible, step-by-step explanations of the calculations involved, addressing fundamental questions such as their purpose, significance, and the interpretation of their individual components. The mathematical procedures are intentionally presented within the body of the text as they constitute an integral part of the methodological exposition and are essential for readers to fully grasp the analytical process. In doing so, the study functions as both an explanatory guide and a practical manual for researchers wishing to apply KFA independently. Additionally, it presents preliminary findings from a KFA analysis of introductions in academic articles and PhD dissertations and offers practical insights for future research.

2 A step-by-step guide for KFA

The initial step for KFA involves the selection of corpora for analysis. This paper examines the introduction sections of PhD dissertations (Dutra et al., 2024, in preparation;

Dutra; Berber Sardinha, 2024)[1] and scientific articles (Ramos et al., 2024; Braga et al., 2024)[2] within the field of Agrarian Sciences. Two distinct corpora were compiled for this purpose and each of them have 21 texts: Corpus 1 (PHDI), comprising 18,795 words from PhD dissertation introductions, and Corpus 2 (SAI), consisting of 19,318 words from scientific article introductions. PHDI texts were written, in English, by Brazilian university students and the SAI texts were published in high impact journals. As can be seen in Table 1 below, the corpora present the same number of texts, approximately the same number of words/tokens, and very similar average text-length. The source corpora (see footnotes 3 and 4) were previously preprocessed, cleaned, and formatted as .txt files. To ensure data comparability between the PHDI and SAI corpora, only files containing very similar word counts were selected.

Table 1: Corpora metadata

	Corpus 1: PHDI	Corpus 2: SAI
Text/file Count	21	21
Token Count	18795	19318
Average Text Length	895	919.9

Source: The authors.

After the corpora were prepared, a situational analysis was conducted. For this study, the framework developed by Biber and Conrad (2019) was employed. This approach is particularly valuable as it allows for a functional interpretation of the items/features selected for the key-feature analysis, a step that will be detailed later in this paper. The framework provides an analytical and descriptive structure that potentially aids in the feature's meaningful interpretation within the context of the analysis. As a sample, the situational classification of Corpus 1: PHDI is shown in Chart 1 below:

[1] The whole corpus of master's thesis and PhD dissertation is called CROPS (Corpus of Agrarian Sciences Postgraduate Studies- Costa et al., in preparation). For this study, we used a subcorpus of CROPS consisting of the introduction sections. Because we selected a subsection of CROPS, we refer to this subcorpus of CROPS as a corpus of PhD dissertation Introduction sections: PHDI.

[2] The whole corpus of Agrarian Sciences research articles is called CorAgrarian (Corpus of Agrarian Sciences Research Articles). For this study, we used a subcorpus of CorAgrarian consisting of the introduction sections. Therefore, we refer to this subcorpus of CorAgrarian as Scientific Article Introduction corpus: SAI.

Chart 1: PhD dissertations' situational analysis

I. Participants
   A. Addressor(s): Author.
      1. Single.
      2. Social characteristics: A graduate student completing a PhD in agrarian sciences.
   B. Addressees
      1. Unenumerated.
      2. The research community, especially within agrarian sciences, and the PhD evaluation panel.

II. Relations among participants
   A. Interactiveness: Low.
   B. Social roles: Unequal. The author is a junior researcher or expert-in-training, while the addressees (evaluation panel, established researchers) are senior experts with greater institutional power.
   C. C. Personal relationship: Distant and professional (e.g., strangers or professional colleagues).
   D. Shared knowledge: High degree of shared specialist knowledge is assumed. Little to no shared personal knowledge is expected.

III. Channel
   A. Mode: Writing.
   B. Specific medium:
      Permanent: Digital document (e.g., PDF) hosted in an online thesis repository.

IV. Production circumstances: Revised and edited.

V. Setting
   A. Are the time and place of communication shared by participants? No.
   B. Place of communication
      1. Public.
      2. University or institutional website / online repository.
   C. Time: Contemporary.

VI. Communicative purposes
   A. General purposes: Contextualize, justify, outline.
   B. Specific purposes: Present PhD research.
   C. Factuality: Factual.
   D. Expression of stance: Epistemic stance.

VII. Topic
   A. General topical domain: Science/ academic.
   B. Specific topic: A specific research area within agrarian sciences.

Source: The authors.

The next step in the analysis is the annotation of the texts using a language-specific tool to automatically categorize every word in the corpus and assign it a specific lexico-grammatical tag. The Multi-Feature Tagger of English (MFTE) (Le Foll; Shakir, 2024) was employed for this purpose. The MFTE is derived from Andrea Nini's Multidimensional Analysis Tagger (MAT) (Nini, 2019), which is an open-source implementation of the Biber Tagger (1988). As its final output, the MFTE[3] calculates the frequency of each feature per text, generating a comprehensive data frame.

Once the texts were processed and the data frame was generated by the tagger, the next step involved selecting the lexico-grammatical features to be analyzed in the study. Egbert and Biber (2023) emphasize that such features should be chosen functionally, taking into account the specific registers under investigation and grounding the selection in prior literature. The lexico-grammatical features selected were those identified in the literature review conducted by Kitjaroenpaiboon et al. (2023), which identifies 24 categories relevant to academic writing and research. We chose to work with this literature review since the categories in it were derived from and supported by studies within the CL field, ensuring their validity and applicability to the analysis of academic discourse. For illustrative purposes, the following are some of the categories and studies considered in the review. Kitjaroenpaiboon et al. (2023, p.100 apud Halliday, 2013) cite the following about verb tenses and their aspect and one subcategory (present simple):

> Tenses and aspects are the most discussed features, expressing time at, during, or over which a state or action denoted by a verb occurs. The change of tense choices can indicate a change in meaning. Tense use is not only about transforming one verb form to another but it is also a temporal implicature (Kitjaroenpaiboon et al., 2023, p. 100 apud Halliday, 2013).
>
> Present simple provides two main communicative functions in the research article. One is to situate a particular event and another is to mark a particular proposition as a generalization (Swales, 2004). In the latter case, the use of 'present simple' indicates that the propositional information is valid regardless of time. (Kitjaroenpaiboon et al., 2023, p. 100)

[3] According to the authors, MTFE presents an overall accuracy rate of 97.83% for texts within the academic writing register (Le Foll; Shakir, 2024).

The selected features,[4] accompanied with the corresponding tags in parenthesis, for the present study are: Words; Average Word-Length (AWL); Type-Token Ratio (TTR); Lexical Density (LDE); Be as Main Verbs (BEMA); Coordinating Conjunctions (CC); Demonstrative Pronouns and Articles (DEMO); Downtoners (DWT); Emphatics (EMPH); Adverbs of Frequency (FREQ); Hedges (HDG); Attributive Adjectives (JJAT); Predicative Adjectives (JJPR); Necessity Modals (MDNE); Nouns (NN); Present Perfect (PEAS); Pronoun 'It' (PIT); Adverbs of Place (PLACE); Other Adverbs (RB); Split Auxiliaries and Infinitives (SPLIT); 'That' Deletion (THATD); 'That' Complement Clauses (THSC); Adverbs of Time (TIME); Past Simple (VBD); Present Simple (VPRT); Wh-Clauses (WHSC); Negation (XX0); Possibility Modals (MDPOSSCall); Prediction Modals (MDPREDall); Passive Voice (PASSall); First Person Pronoun (PP1all); Third Person Pronoun (PP3all); Adverbs having no additional tag (RBother).

It is important for researchers working with automated tagging tools to consider that, most likely, no tagger will perfectly align with the specific tagging needs of a given study. It falls upon the researcher to critically evaluate the tagging output and devise appropriate solutions. These solutions may involve manual adjustments or automated processing using programming languages. However, it is not the focus of this study to provide a thorough description of how decisions were made regarding tagging.

Table 2 below shows part of the spreadsheet[5] with the data provided by the tagger. As shown, column A lists the filenames of the texts included in the corpus, while each subsequent column represents a specific linguistic feature, identified by its corresponding tag. These columns display the frequency of each feature across individual texts. For example, Column B shows that the text *CorAgrarian.AgE.agrmachine.06_introduction_mcl.txt* contains 892 words (or tokens), whereas *CorAgrarian.AnS.annutrition.01_introduction_mcl.txt* contains 941.

[4] The features were selected based on an effort to combine the lexical grammatical-features listed in Kitjaroenpaiboon et al. (2023) to the tags we had available from MFTE.

[5] MTFE will produce dataframes that can be opened as most types of spreadsheet software. Users have the freedom to choose their preferred one.

Table 2: Features raw frequency in Corpus 2: SAI

	A	B	C	D	E	F	G
1	Filename	Words	BEMA	CC	DEMO	DWNT	EMPH
2	CorAgrarian.AgE.agrmachine.06_introduction_mcl.txt	892	18	54	5	2	9
3	CorAgrarian.AgE.largecrop.03_introduction_mcl.txt	848	9	60	7	1	1
4	CorAgrarian.AgE.largecrop.04_introduction_mcl.txt	903	10	59	10	0	2
5	CorAgrarian.AgE.largecrop.07_introduction_mcl.txt	859	3	46	8	1	0
6	CorAgrarian.AgE.largecrop.09_introduction_mcl.txt	986	11	49	6	2	2
7	CorAgrarian.Agr.ecology.26_introduction_mcl.txt	912	11	48	2	0	1
8	CorAgrarian.Agr.soil.11_introduction_mcl.txt	937	8	52	2	3	1
9	CorAgrarian.AnS.anbreeding.07_introduction_mcl.txt	858	20	26	3	2	1
10	CorAgrarian.AnS.annutrition.01_introduction_mcl.txt	941	13	46	1	0	1

Source: dataframe generated by MFTE tagger.

Most taggers will provide the frequency of each tag both raw and normalized. Normalization is a crucial step in corpus-based studies, as it ensures that frequency counts are comparable across texts of varying lengths. As Biber, Conrad, and Reppen (1998) explain, normalization adjusts raw frequencies by dividing them by the total number of words in a text and multiplying by a common base—typically 1,000 or 1,000,000. The choice of base depends on the size of the corpus and the analytical scope of the study. A base of 1,000 words is commonly used in smaller corpora or when comparing shorter texts, as it provides interpretable values without overinflating results. This approach is particularly suitable when dealing with texts that range from a few hundred to a few thousand words, as in the present study. In contrast, normalization per 1,000,000 words is standard in large-scale corpora, such as the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA), where macro-level patterns are the focus. In Table 3 below, raw frequencies were replaced by normalized frequencies in the spreadsheet file:

Table 3: Features normalized frequency in Corpus 2: SAI

Filename	Words	BEMA	CC	DEMO	DWNT	EMPH
CorAgrarian.AgE.agrmachine.06_introduction_mcl.txt	892	20.179	60.538	5.605	2.242	10.09
CorAgrarian.AgE.largecrop.03_introduction_mcl.txt	848	10.613	70.755	8.255	1.179	1.179
CorAgrarian.AgE.largecrop.04_introduction_mcl.txt	903	11.074	65.338	11.074	0	2.215
CorAgrarian.AgE.largecrop.07_introduction_mcl.txt	859	3.492	53.551	9.313	1.164	0
CorAgrarian.AgE.largecrop.09_introduction_mcl.txt	986	11.156	49.696	6.085	2.028	2.028
CorAgrarian.Agr.ecology.26_introduction_mcl.txt	912	12.061	52.632	2.193	0	1.096
CorAgrarian.Agr.soil.11_introduction_mcl.txt	937	8.538	55.496	2.134	3.202	1.067
CorAgrarian.AnS.anbreeding.07_introduction_mcl.txt	858	23.31	30.303	3.497	2.331	1.166
CorAgrarian.AnS.annutrition.01_introduction_mcl.txt	941	13.815	48.884	1.063	0	1.063

Source: dataframe generated by MFTE tagger.

Once the frequency numbers are normalized, it is possible to start working with Cohen's d formula, presented below (Figure 1). The formula will be broken down into its core components to provide a clearer understanding of the underlying mathematical processes involved. The explanation draws on a range of complementary sources that discuss these statistical concepts from different perspectives. For further detail, see Brezina (2018); Cohen (2013); Gries (2016); Larson-Hall (2009); Plonsky and Oswald (2014).

Figure 1: Cohen's d formula

$$d = \frac{M_1 - M_2}{SD_{pooled}}$$

Source: The authors.

Cohen's d is a statistical measure used to quantify the standardized difference between two groups—in this case, two corpora. Rather than simply indicating whether a difference exists, Cohen's d tells us how substantial that difference is. When comparing Corpus 1 and Corpus 2 in terms of the frequency of a particular lexico-grammatical feature (such as passive voice, modal verbs, or noun density), the key questions are: Is this feature more frequent in one corpus than the other? If so, is the difference considered small, moderate, or large according to standard interpretive guidelines? More on that is explained later in the article.

In the formula, the means (M1: Corpus 1 mean and M2: Corpus 2 mean) reflect the average frequency of the key linguistic feature under analysis within each corpus. As measures of central tendency, the means indicate where most data points lie for each group. Cohen's d quantifies the difference between these two central values. This component captures the direction and magnitude of the average difference between the corpora. Without the means, a basis for assessing the extent of difference in feature usage between the two corpora would be lacking.

Table 4 below shows the mean additions. The final row of the second column, for example, presents the mean number of words in the entire corpus. This value was calculated by summing the word counts of all individual texts and dividing that total by the number of

texts analyzed. In the spreadsheet, this was done using the built-in =mean function provided by the spreadsheet software.

Table 4: Added mean values in Corpus 1: PHDI

Filename	Words	BEMA	CC	DEMO	DWNT	EMPH
CorAgrSc.Ans.APhD.Int.2017.035.txt	912	10.965	70.175	10.965	0	3.289
CorAgrSc.Ans.APhD.Int.2021.005.txt	869	3.452	48.331	4.603	0	0
CorAgrSc.FEg.APhD.Int.2019.032.txt	848	8.255	38.915	18.868	1.179	4.717
CorAgrSc.FEg.APhd.Int.2019.019.txt	979	21.45	37.794	11.236	2.043	1.021
CorAgrSc.FEg.APhd.Int.2020.018.txt	890	21.348	39.326	10.112	2.247	0
CorAgrSc.FEg.APhd.Int.2022.012.txt	898	20.045	40.089	13.363	0	4.454
CorAgrSc.For.APhD.Int.2020.014.txt	681	17.621	51.395	8.811	0	2.937
CorAgrSc.For.APhD.Int.2021.015.txt	864	12.731	49.769	12.731	0	0
Mean	895	13.875	49.43061905	10.02571429	0.8851904762	2.067142857

Source: The authors.

In the formula, M1 - M2 is divided by SD (Standard Deviation). SD is a statistical measure that indicates how much the values in a dataset tend to vary or "spread out" from the average. In simple terms, it tells how consistent or variable the data are. If the values in a dataset are closely clustered around the average, the standard deviation will be small. Conversely, if the values are widely dispersed, the standard deviation will be larger. For example, imagine two classes of students taking the same test. In both classes, the average score is 75 out of 100. However, in Class A, most students scored between 70 and 80, while in Class B, scores ranged from 50 to 100. Although the averages are identical, the standard deviation for Class B would be much larger, reflecting greater variability in student performance. In the context of this study, standard deviation helps in understanding how consistently a feature (e.g., modal verb usage) appears across texts within a corpus. A small standard deviation indicates that most texts use the feature at similar rates, while a large standard deviation suggests wide differences in usage from one text to another. Cohen's d is used to quantify the size of the difference between the texts in the two corpora, in a standardized way. It is not just based on raw differences, but it also takes into account the natural variability in the data. This is where standard deviation plays a crucial role. While the means of two groups tell us the average value of a particular feature in each group (e.g., the average frequency of modal verbs in two corpora), they do not tell us how spread out or consistent the data are within each group. That is precisely what the standard deviation reveals.

The standard deviation type used in Cohen's d formula is a pooled one. To calculate the pooled standard deviation, several steps are involved (see Figure 2). As mentioned earlier, the standard deviation represents the overall variability from both corpora combined. However, it is not obtained by simply averaging the two standard deviations. Instead, each standard deviation is first squared—a step that removes negative signs and emphasizes larger differences, ensuring that all variations contribute positively. These squared values are then added together and divided by 2 to calculate the average squared deviation. Finally, the square root of this result is taken. This last step is important because it converts the squared value back into the original unit of measurement, allowing the pooled standard deviation to be expressed in the same scale as the means. This makes it meaningful and directly usable in calculating Cohen's d, where standardizing the difference between the two means depends on having a comparable unit of variation.

Figure 2: Pooled Standard Deviation formula

$$SD_{\text{pooled}} = \sqrt{\frac{{SD_1}^2 + {SD_2}^2}{2}}$$

Source: The authors.

To enhance further clarity, Figure 3 displays both formulas, illustrating how the pooled standard deviation is derived and integrated into the overall calculation of Cohen's d.

Figure 3: Breaking down Cohens'd formula.

$$SD_{\text{pooled}} = \sqrt{\frac{{SD_1}^2 + {SD_2}^2}{2}} \qquad d = \frac{M_1 - M_2}{SD_{pooled}}$$

Source: The authors.

In the spreadsheet software, the standard deviation (not the pooled standard deviation) will be used. The reason for that is explained later in this session. The standard deviation can be calculated automatically using the spreadsheet software, through the =STDEV built-in function. This is done by selecting all relevant text cells corresponding to each key feature and

entering the formula in the appropriate output cell. Table 5 shows the added final row.

Table 5: Added Standard Deviation values in Corpus 1: PHDI

Filename	Words	BEMA	CC	DEMO	DWNT	EMPH
CorAgrSc.Ans.APhD.Int.2017.035.txt	912	10.965	70.175	10.965	0	3.289
CorAgrSc.Ans.APhD.Int.2021.005.txt	869	3.452	48.331	4.603	0	0
CorAgrSc.FEg.APhD.Int.2019.032.txt	848	8.255	38.915	18.868	1.179	4.717
CorAgrSc.FEg.APhd.Int.2019.019.txt	979	21.45	37.794	11.236	2.043	1.021
CorAgrSc.FEg.APhd.Int.2020.018.txt	890	21.348	39.326	10.112	2.247	0
CorAgrSc.FEg.APhd.Int.2022.012.txt	898	20.045	40.089	13.363	0	4.454
CorAgrSc.For.APhD.Int.2020.014.txt	681	17.621	51.395	8.811	0	2.937
CorAgrSc.For.APhD.Int.2021.015.txt	864	12.731	49.769	12.731	0	0
Mean	895	13.875	49.43061905	10.02571429	0.8851904762	2.067142857
Standard Deviation	69.06011874	5.111383335	8.744446961	3.862277322	0.952720873	1.895432702

Source: The authors.

As an example, the standard deviation for *'Be' as Main Verb* (BEMA) feature in the PHDI corpus is 5.111383335 (5.11), while the standard deviation for the *Coordinating Conjunctions* (CC) feature in the PHDI corpus is 8.744446961 (8.74). These values represent the extent of variation within each corpus for the respective features.

The comparison between these two SD values reveals that the use of *Coordinating Conjunctions* (CC) varies more widely across texts in the PHDI corpus than *'Be' as Main Verb* (BEMA). In other words, some texts in the corpus contain a much higher or lower number of *Coordinating Conjunctions* compared to others, indicating greater dispersion. In contrast, the *'Be' as Main Verb* usage is more consistent across texts, with less fluctuation in frequency.

To summarize, in the formula for Cohen's d, the difference between the means of the two corpora (M1 - M2) is divided by the pooled standard deviation, which is an average of the standard deviations of both corpora. By dividing the mean difference by this pooled value, Cohen's d tells us how large the difference is compared to the typical amount of variation in the data. By expressing the result in standard deviation units, Cohen's d helps us understand whether the difference between the groups is small, moderate, or large. At this point, all the numerical data required for the calculations related to Corpus PHDI have been completed. The next step is to record, in the same spreadsheet, the data and corresponding calculations for Corpus SAI.

With all the necessary data available, the next step is to calculate the d value. This can be done efficiently, as the formula can be entered directly into the desired spreadsheet cell and applied to each item in the study. The formula = (MEAN1 - MEAN2) / SQRT((POWER(SD1, 2) + POWER(SD2, 2)) / 2) is applied as follows: MEAN1 represents the average frequency of the item in Corpus 1 (PHDI), and MEAN2 the average in Corpus 2 (SAI). Similarly, SD1 and SD2 correspond to the standard deviations of the item in Corpus 1 and Corpus 2, respectively.

The standard deviation, not the pooled standard deviation, was used in the spreadsheet since the built-in formula incorporates the SDpooled internally—using squaring (POWER) and the square root (SQRT). Therefore, there is no need to compute the pooled standard deviation as a separate step. As a result of this calculation, each lexico-grammatical item will be assigned a positive or negative d value. Positive values indicate that the item is more representative and significantly used in Corpus 1, while negative values indicate higher representation in Corpus 2. Table 6 below shows the result within the last rows in the spreadsheet:

Table 6: Cohen's d values for Corpus PHDI and Corpus SAI

	Words	AWL	TTR	LDE	BEMA	CC
Cohen's d	0.4635306051	0.039982117	0.416680446	-0.4802286676	0.2070329055	-0.5113448947

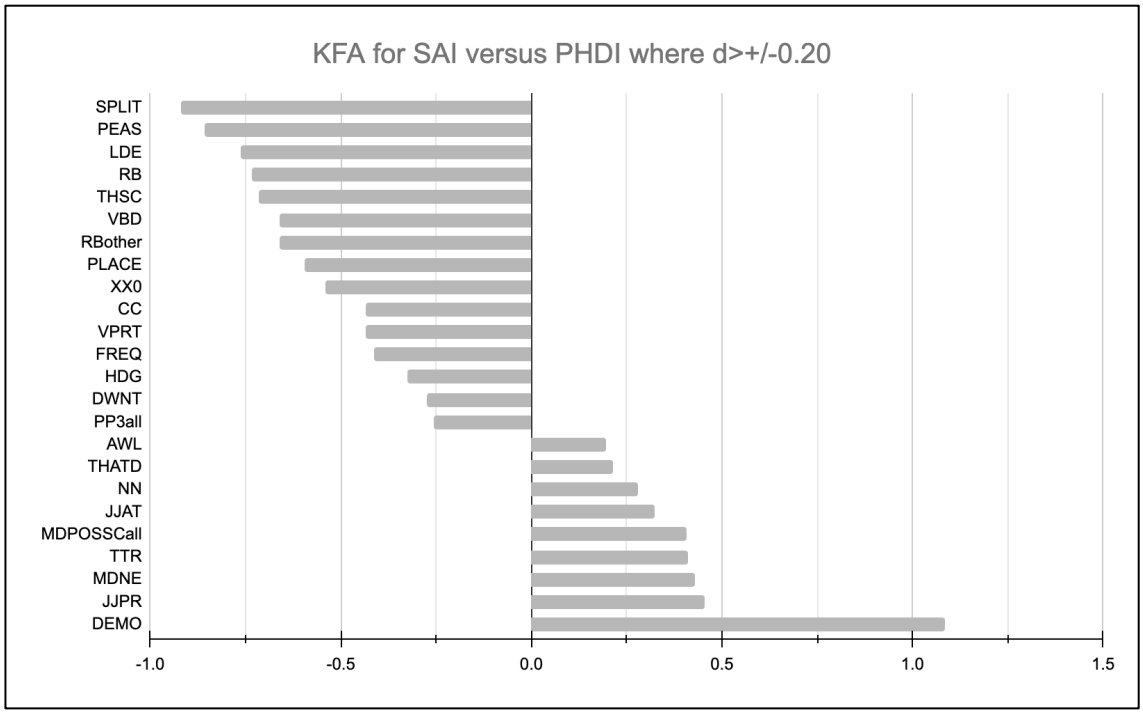
Source: The authors.

A graph can then be generated to visually represent the d-values associated with each selected feature. Figure 4 below presents the results of the Key-feature analysis for the PHDI and SAI corpora. On the right-hand side of the graph, positive d-values correspond to Corpus 1 (PHDI), while negative values on the left indicate features more prominent in Corpus 2 (SAI).

Although Egbert and Biber (2023) note that fixed benchmarks for interpreting the magnitude of Cohen's d are not strictly necessary, as the values can be interpreted along a continuous scale, this study adopts general thresholds for effect size: a d-value greater than 0.20 (whether positive or negative) is considered small; a d-value above 0.50 is regarded as medium; and a d-value exceeding 0.80 represents a large effect size. Based on these criteria, for the purposes of the sample analysis presented in this paper, all d-values below ±0.20 were excluded from the graph. The interpretation of results follows the logic of standard deviation

units. For instance, a feature with d = -1.0 indicates that its mean frequency (across texts) is one standard deviation higher in Corpus 2 than in Corpus 1 (Egbert; Biber, 2023).

Figure 4: KFA for SAI versus PHDI



Source: The authors.

The final step of the KFA methodology, as stated by Egbert and Biber goes as follows:

> Once lexico-grammatical features have been ranked according to their Cohen's d values, the quantitative portion of the key feature methods are finished. It is then up to the researcher to provide linguistic interpretations for the features that are deemed to be strongly associated with the two corpora. These linguistic interpretations can be based on lexico-grammatical functions established in previous literature or reference grammars. Importantly, all interpretation should be supported by textual patterns and examples from the two corpora. (Egbert; Biber, 2023, p. 124)

Owing to space constraints and, more crucially, to the specific objective of this study—namely, a detailed examination of the KFA methodology—the following section provides an illustrative analysis. Accordingly, only two linguistic features were selected for demonstration. The first, *Split auxiliaries and infinitives (SPLIT)*, constitutes the most salient feature in Corpus SAI, exhibiting a large effect size (d = -1.00). The second, *"Demonstrative pronouns" (DEMO)*, likewise displays a large effect size (d = +1.02) and represents the most prominent feature in

Corpus PHDI. A comprehensive analysis encompassing the full set of linguistic features is currently underway and will be reported in a separate publication.

4 Sample linguistic interpretation

The first feature analyzed is SPLIT (Split auxiliaries and infinitives), which produced a large-effect d-value of -1.00, and indicates their significantly higher frequency in Corpus SAI (Scientific Article Introductions). As Biber (1988) and Quirk (1985) point out, split auxiliaries are used to make explicit marking of the writers' own persuasion or argumentative discourse designed to persuade the readers as well as create emphasis. This linguistic feature is realized when an adverb is placed between an auxiliary and a following verb, e.g., *would actually increase*. In Corpus SAI, examples of this use are:

1. *(…) discrete rainfall events can also trigger a rapid pulse of soil respiration;*
2. *(…) leading to plant mortality and competitive release of less palatable species that may consequently encroach or become invasive;*
3. *Marked changes in any of these biotic and abiotic factors could profoundly alter the global C cycle and feedbacks to climate change;*
4. *It is paramount to consider the effects of scale in river management as evidence collected at the micro scale does not always extrapolate to the larger scale and vice versa.*

This pattern may support the hypothesis that article introductions, being characterized by a specialized tone rather than an apprentice tone, require a more persuasive style in order to effectively cause the reader to engage in the writer's rationale through reasoning or argumentation. The use of split auxiliaries might also create emphasizing effects.

The second feature analyzed in this sample is DEMO (Demonstrative Pronouns). It is the most productive feature in corpus PHDI (PhD dissertation introductions), with a large effect size (d = +1.02). One of the main uses of demonstrative pronouns (this, that, these and those) is to aid in the establishment of shared knowledge between readers and the authors (Kanoksilapatham, 2003 Apud Kitjaroenpaiboon et al., 2023). Other authors highlight that important communicative functions of this feature are to signal a high focus on the referent to which the writer wants to draw the reader's attention; to signal a focus and topicality in texts;

to reduce ambiguities that often result from the use of pronominal 'this' and also to give the text a more professional style; to be used as pronouns as well as determiners; and to refer to a complex predication (Rodman, 1994; Swales, Feak, 2012; Biber; Gray, 2010). In Corpus PHDI, examples of these uses are:

1. *This thesis is a first effort to discover theoretical references in tambaqui breeding (…);*
2. *However, lignocellulosic biomass needs a step that helps to reduce the material's recalcitrance.;*
3. *But when adding those components in a system's perspective, (…);*
4. *To explore these four dimensions, different sensory methods were used.*

A hypothesis for the prominent use of 'DEMO' in this corpus might be connected to the fact that PhD candidates need to engage in full literature connections to establish a meaningful and appropriate territory for their own research. Article writers, on the other hand, need to rely much less on other writer's concepts and might reference other's work less, since the focus is their own new investigation.

Regarding the Situational Analysis presented in the methodological session, it is possible to state that linguistic choices are shaped by the situational context of communication, specifically factors such as purpose, audience, medium, and level of expertise. The contrast observed between the two corpora aligns well with these principles. Together, these results reinforce the value of the Situational Analysis framework in explaining register variation in academic discourse. They show that even what may seem relatively subtle linguistic differences (like demonstrative pronouns) can be meaningfully interpreted through the lens of contextual variables that shape genre-specific communication strategies.

5 Final considerations

This study suggests that detailing the methodological procedures of the Key-Features Analysis (Egbert; Biber, 2023), as applied here, provides a valuable resource for students and researchers seeking to implement this framework in their own work. While the analysis was carried out using only spreadsheet tools, the steps involved are fully automatable through

scripting, offering possibilities for customization and broader applicability, which is an objective aligned with the future goals of the authors of this text.[6]

Importantly, this work also responds to a gap often found in academic publications for novice corpus linguistics researchers: the absence of detailed descriptions of statistical procedures used in the methodology of the study. This issue is particularly relevant in the humanities, where students often report feelings of anxiety or insecurity when confronted with mathematical or statistical content (Ballo; Pauli; Worrell, 2016; Chew; Dillon, 2014; Murtonen; Lehtinen, 2003).

Finally, the sample results presented here have the potential to inform further investigations which may contribute to the development of pedagogical materials designed to support academic writing instruction. Also, it would be important to replicate this investigation using larger corpora so that the size variable could be evaluated.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

BALLO, K.; PAULI, R.; WORRELL, M. Undergraduate students' experiences of anxiety in statistics courses. **Psychology Learning & Teaching**, *[S. l.]*, v. 9, n. 3, p. 274–285, 2016.

BIBER, D.; GRAY, B. Challenging stereotypes about academic writing: complexity, elaboration, and explicitness. **Journal of English for Academic Purposes**, *[S. l.]*, v. 9, n. 1, p. 2–20, 2010.

BIBER, D. **Variation across Speech and Writing**. 1. ed., [s.l.] : Cambridge University Press, 1988. DOI: 10.1017/CBO9780511621024. Disponível em: https://www.cambridge.org/core/product/identifier/9780511621024/type/book. Acesso em: 8 jul. 2025.

BIBER, D.; CONRAD, S. **Register, Genre, and Style**. [s.l.] : Cambridge University Press, 2019. 423 p. Google-Books-ID: x7OQDwAAQBAJ.

[6] Few studies have used KFA and compared several corpora, combining this analysis with cluster analysis (e.g. Berber Sardinha et al. 2024).

BIBER, D.; CONRAD, S.; CORTES, V. If you look at …: Lexical Bundles in University Teaching and Textbooks. **Applied Linguistics**, *[S. l.]*, v. 25, n. 3, p. 371–405, 2004. DOI: 10.1093/applin/25.3.371.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus Linguistics: Investigating Language Structure and Use**. [s.l.] : Cambridge University Press, 1998. 324 p. Google-Books-ID: 2h5F7TXa6psC.

BRAGA, J.; MARQUES, C.; DUTRA, D.; TEIXEIRA, G.; OLIVEIRA, S. O Uso de Present Simple, Present Perfect, Past Simple e Past Perfect nas Introduções de Artigos Científicos, Teses e Dissertações Escritos em Inglês na Área de Ciências Agrárias. *In*: XVI Encontro de Linguística de Corpus e XIII Escola Brasileira de Linguística Computacional, 2024, Brasília. **Anais eletrônicos** [...] Associação Brasileira de Linguística de Corpus, 2024. p. 59 – 60. Disponível em: https://www.elc-ebralc.net.br/anais. Acesso em: 9 dez. 2025.

BREZINA, V. **Statistics in Corpus Linguistics: A Practical Guide**. Cambridge: Cambridge University Press, 2018. DOI: 10.1017/9781316410899. Disponível em: https://www.cambridge.org/core/books/statistics-in-corpus-linguistics/4E530F86B328B2287681AD240796D2CF. Acesso em: 8 jul. 2025.

CHEW, P.; DILLON, D. Statistics Anxiety Update: Refining the Construct and Recommendations for a New Research Agenda. **Perspectives on Psychological Science**, *[S. l.]*, v. 9, n. 2, p. 196–208, 2014. DOI: 10.1177/1745691613518077.

COHEN, J. **Statistical Power Analysis for the Behavioral Sciences**. 2. ed., New York: Routledge, 2013. 567 p. DOI: 10.4324/9780203771587.

DUTRA, D.; BERBER SARDINHA, T. The role of Corpus Linguistics in EAP. *Em*: **English For Academic Purposes: Reflections, Description & Pedagogy**. 1. ed., Porto Alegre: Zouk, 2024.

DUTRA, D.; BOCORNY, A.; COSTA, D.; TEIXEIRA, G.; MARQUES, C. Desafios Metodológicos na Compilação de Textos Acadêmicos das Ciências Agrárias. *In*: XVI Encontro de Linguística de Corpus e XIII Escola Brasileira de Linguística Computacional, 2024, Brasília. **Anais eletrônicos** [...] Associação Brasileira de Linguística de Corpus, 2024. p. 39 – 40. Disponível em: https://www.elc-ebralc.net.br/anais. Acesso em: 9 dez. 2025.

DUTRA, D.; BOCORNY, A.; COSTA, D.; TEIXEIRA, G.; MARQUES, C. Introducing CROPS: A specialized corpus of Agrarian Sciences graduate theses from Brazilian universities. *[S. l.]*, in preparation.

EGBERT, J.; BIBER, D. Key feature analysis: a simple, yet powerful method for comparing text varieties. **Corpora**, *[S. l.]*, v. 18, n. 1, p. 121–133, 2023. DOI: 10.3366/cor.2023.0275.

FLOWERDEW, L. Corpus-based research and pedagogy in EAP: From lexis to genre. **Language Teaching**, *[S. l.]*, v. 48, n. 1, p. 99–116, 2015. DOI: 10.1017/S0261444813000037.

GRIES, S. **Quantitative Corpus Linguistics with R: A Practical Introduction**. 2. ed., New York: Routledge, 2016. 286 p. DOI: 10.4324/9781315746210.

KITJAROENPAIBOON, W.; FAHKRAJANG, S.; FONGSARUN, P.; PLOYLERMSAENG, W. A Review of Lexico-grammatical Features and their Functions in an Academic Discourse. **Journal of Multidisciplinary in Social Sciences**, *[S. l.]*, v. 19, n. 2, p. 99–112, 2023.

LARSON-HALL, J.; LARSON-HALL, J. **A Guide to Doing Statistics in Second Language Research Using SPSS**. New York: Routledge, 2009. 440 p. DOI: 10.4324/9780203875964.

LE FOLL, E.; SHAKIR, M. The Multi-Feature Tagger of English (MFTE): Rationale, description and evaluation. **Research in Corpus Linguistics**, *[S. l.]*, v. 13, n. 2, p. 63–93, 2024. DOI: 10.32714/ricl.13.02.03.

MURTONEN, M.; LEHTINEN, E. Difficulties Experienced by Education and Sociology Students in Quantitative Methods Courses. **Studies in Higher Education**, *[S. l.]*, v. 28, n. 2, p. 171–185, 2003. DOI: 10.1080/0307507032000058064.

NINI, A. **MAT The Multidimensional-Analysis-Tagger**. 2019. Disponível em: https://www.scribd.com/document/670853671/MAT-The-multidimensional-analysis-tagger. Acesso em: 30 jun. 2024.

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. **From Corpus to Classroom: Language Use and Language Teaching**. Cambridge: Cambridge University Press, 2007. (Cambridge Professional Learning). DOI: 10.1017/CBO9780511497650.

PLONSKY, L.; OSWALD, F. How Big Is "Big"? Interpreting Effect Sizes in L2 Research. **Language Learning**, *[S. l.]*, v. 64, n. 4, p. 878–912, 2014. DOI: 10.1111/lang.12079.

QUIRK, R.; GREENBAUM, S.; LEECH, G.; SVARTIVIK, J. **A grammar of contemporary English language**. Harlow: Longman, 1985.

RAMOS, C.; OLIVEIRA, S.; DUTRA, D.; BRAGA, J.; VICTOR, A. Nominalization: Corpus-Based Study of Discussion Sections in Forestry Research Articles. *In*: XVI Encontro de Linguística de Corpus e XIII Escola Brasileira de Linguística Computacional, 2024, Brasília. **Anais eletrônicos** [...] Associação Brasileira de Linguística de Corpus, 2024. p. 53 – 54. Disponível em: https://www.elc-ebralc.net.br/anais. Acesso em: 9 dez. 2025.

RODMAN, L. The active voice in scientific articles: frequency and discourse functions. **Journal of Technical Writing and Communication**, *[S. l.]*, v. 24, n. 3, p. 309–331, 1994.

SCOTT, M. PC analysis of key words — And key key words. **System**, *[S. l.]*, v. 25, n. 2, p. 233–245, 1997. DOI: 10.1016/S0346-251X(97)00011-0.

SILVA, C.; BRITO, M.; CAZORLA, I.; VENDRAMINI, C. Atitudes em relação à estatística e à matemática. **Psico-USF**, *[S. l.]*, v. 7, p. 219–228, 2002. DOI: 10.1590/S1413-82712002000200011.

SWALES, J.; FREAK, C. **Academic Writing for Graduate Students: Essential Tasks and Skills**. Ann Arbor: The University of Michigan Press, 2012.

TRIBBLE, C.; WINGATE, U. From text to corpus – A genre-based approach to academic literacy instruction. **System**, *[S. l.]*, v. 41/1, n. 2, p. 307–321, 2013.

Recebido em: 11.07.2025 Aprovado em: 26.01.2026