https://dx.doi.org/10.1590/1678-460X202238249731

D E L T A

Articles

The learner corpus path: a worthwhile methodological challenge

A trajetória da compilação de um corpus de aprendiz: um desafio metodológico compensador

Deise Prina Dutra[1]
Bárbara Malveira Orfanò[2]
Annallena de Souza Guedes[3]
Jessica Ceritello Alves[4]
João Gabriel Fekete[5]

ABSTRACT

Corpus compilation is a challenging research endeavor that many researchers decide to pursue. Few learner corpora, however, can be easily accessed (e.g.,the International Corpus of Learner English), and none of them carry a variety of text registers written by English learners

1. Universidade Federal de Minas Gerais. Belo Horizonte - Brasil. http://orcid.org/0000-0001-5799-5174. E-mail: deisepdutra@gmail.com.
2. Universidade Federal de Minas Gerais. Belo Horizonte - Brasil. https://orcid.org/0000-0002-5761-6303. E-mail: barbara.orfano@gmail.com.
3. Instituto Federal da Bahia - Campus Ilhéus. Bahia - Brasil. https://orcid.org/0000-0002-5246-4443. E-mail: annallenaguedes@ifba.edu.br.
4. Universidade Federal de Minas Gerais. Belo Horizonte - Brasil. https://orcid.org/0000-0001-5562-7271. E-mail: jessicaceritello@gmail.com.
5. Universidade Federal de Minas Gerais. Belo Horizonte - Brasil. https://orcid.org/0000-0002-2439-1742. E-mail: jfekete@ufmg.br.

at different proficiency levels studying in the Brazilian university context. Therefore, the aim of this paper is to present the compilation of a learner corpus, much needed in our research and teaching context, pointing out the advantages of building this type of corpus for the understanding of learners' needs as well as for pedagogical decision-making based on sound data. Presenting a detailed rationale of the corpus compilation, this article reveals the various decisions made in order to guarantee that fair comparisons can be made. To exemplify the value of building a carefully designed corpus, results of previous studies are compared. Some of the conclusions reached refer to the need for discipline-specific tasks to propel writing proficiency and for authorship skills to be developed in English for Academic Purposes classes to foster academic success.

Keywords: *learner corpus; academic writing; EAP; corpus linguistics.*

RESUMO

A compilação de corpus é uma empreitada de pesquisa desafiadora que muitos pesquisadores decidem realizar. Poucos corpora de aprendizes, entretanto, podem ser facilmente acessados (por exemplo, o International Corpus of Learner English), e nenhum deles carrega uma variedade de registros textuais escritos por aprendizes de inglês de níveis diferentes de proficiência e que estudam no contexto universitário brasileiro. Nesse sentido, o objetivo deste artigo é apresentar a compilação de um corpus de aprendiz, muito necessário em nosso contexto de pesquisa e ensino, evidenciando as vantagens de construir este tipo de corpus para a compreensão das necessidades dos aprendizes, bem como para as tomadas de decisões pedagógicas baseadas em dados sólidos. Apresentando a fundamentação detalhada para a compilação do corpus, este trabalho revela as várias decisões tomadas, a fim de garantir que comparações justas possam ser feitas. Algumas conclusões obtidas referem-se à necessidade de tarefas específicas por área para impulsionar a proficiência na escrita, e para o desenvolvimento das habilidades de autoria nas aulas de Inglês para Fins Acadêmicos para fomentar o sucesso acadêmico.

Palavras-chave: *corpus de aprendiz; escrita acadêmica; IFA; linguística de corpus.*

1. Introduction

When linguists select their research questions and choose how to investigate what they are interested in, so many decisions have to be made. Undoubtedly, the methodology has to be appropriate for the study. Using a corpus linguistics methodological perspective may be the best choice if the questions are related to how people use language in different contexts, as Crawford and Csomay highlight:

> While understanding variation and contextual differences is a goal shared by researchers of other areas of linguistic research, corpus linguistics describes language variation and use by looking at large amounts of texts that have been produced in similar circumstances. (Crawford & Csomay, 2016, p.5)

This empirical approach allows results to be generalized, as a well-designed corpus will adequately represent a register, which can be understood as "a variety associated with a particular situation of use (including particular communicative purposes)" (Biber & Conrad, 2009, p. 6). The researcher, then, should consider the situational characteristics of a register: the participants, the relations among participants, the channel used, and the production circumstances (Biber & Conrad, 2009). Having established clear research questions to answer and the characteristics of the register or registers the linguist is interested in investigating, it is time to choose the corpus to be used. Would a readily available corpus be suitable, or would a corpus compilation be necessary?

In our research area, learner language, there are few corpora that can be accessed, for instance, the *International Corpus of Learner English* (ICLE)[6], the Louvain International Database of Spoken English Interlanguage (LINDSEI)[7], Louvain Corpus of Native English Essays (LOCNESS), the Michigan Corpus of Upper-Level Student Papers

6. ICLE released its third version in 2020 (Granger et al., 2020). It is a corpus of argumentative essays written by learners of English from upper intermediate to advanced levels of English and from 25 different language backgrounds. It has over 5.5 million words and is hosted on a web-based interface. https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html.

7. LINDSEI is a corpus of interviews gathered from learners of English speakers of 11 different native languages. (Gilquin et al.,2010). https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html.

(MICUSP)[8] and the British Academic Written English corpus of English texts (BAWE)[9]. Each of these corpora was designed with a specific purpose. While ICLE and LINDSEI were compiled to allow access to English learners' interlanguage[10], MICUSP and BAWE focused on high grade written papers of different genres and LOCNESS on essays written by American and British university students. However, there are several similarities regarding the situational characteristics involved in the compilation of these corpora. The participants are all students at higher education institutions. They are authors who write or speak, either in a context where what is produced is being assessed or not being assessed, including or excluding time constraints. Besides the fact that the production circumstances may vary, the addressors are students and can be considered novice or apprentice writers.[11] The main difference among these corpora is that participants have different first language backgrounds. After reflecting on these characteristics, a researcher may wonder how suitable such corpora would be for their study. In our case, our research context is a Brazilian university; consequently, some questions would remain unanswered if our studies are limited to these corpora. Despite the two facts that the corpora are all quite large and that ICLEV3 has a subcorpus of essays written by Brazilian students (Br-ICLE[12]), we ultimately found them insufficient for our needs, particularly to deeply investigate linguistic variation across text genres, across academic levels (undergraduate and graduate), across disciplines and across proficiency levels to understand the users' choices with cross-sectional or longitudinal data perspectives. Such aspects cannot be fully covered with Br-ICLE data. Furthermore, making a new corpus available to other researchers has also been one of our goals. A CQPweb

8. MICUSP, a written corpus, has about 2.6 million words. Corpus information is available at http://micusp.elicorpora.info/
9. The BAWE corpus contains 2761 pieces of proficient assessed student writing. It can be accessed through the Oxford Text Archive https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2539.
10. The term *interlanguage* was coined by Selinker (1972, p. 214): "… the existence of a separate linguistic system based on the observable output which results from a learner's attempted production of a TL norm. This linguistic system we will call 'interlanguage' (IL)."
11. Scott and Tribble (2006, p.133) prefer to use the terms 'apprentice' and 'expert' writers rather than 'learner' and 'native speaker'. "Expert texts may most easily be identified on the crude basis of their having been published, or their having been disseminated to specific readeships within bureaucratic, commercial, professional or other organizations".
12. Br-ICLE, coordinated by Tony Berber-Sardinha, has 200.000 words.

framework will soon be available for searchers on our corpus[13] with tools such as keyword search, collocate list, with different association measures, and visualization of occurrence dispersion.

To the best of our knowledge, there was no comprehensive learner corpus of Brazilian university learners' English written texts compiled in the classroom context and available for the studies our research group was aiming at. Therefore, in 2013, as Section 3 lays out comprehensively, a Brazilian learner academic English corpus was designed (CorIFA[14]). Our efforts were motivated, fundamentally, by our desire to improve learners' use of academic English. Corpus analysis allows the studying of a particular group and the corpus compilation seemed to be a great challenge to be pursued as, ultimately, it would be the basis for developing appropriate materials and new courses. Accordingly, the aim of this paper is to present the compilation of a learner corpus, much needed in our research and teaching context, pointing out the advantages of building this type of corpus for the understanding of learners' needs as well as for pedagogical decision-making based on sound data. The following sections will deal, firstly, with the literature, which is the basis of our work, secondly, with the methodological paths taken to compile the academic English learner corpus and, thirdly, with studies based on CorIFA.

2. Theoretical background

Varieties of English as an additional language[15] far outnumber native-speaker varieties,[16] inspiring the study of non-native spoken

13. Besides the university learner corpus described in the article, we will also make available other corpora organized by our CNPq research group, *Grupo de Estudos de Corpora Especializados e de Aprendizes* (GECEA), such as *CALIEMT*:*Corpus de Aprendizes da Língua Inglesa do Ensino Médio* Técnico (Xavier et al., 2019; Oliveira et al., 2017) and CorAChem (Corpus of Articles in Chemistry) and CorAAL(Corpus of Articles in Applied Linguistics) (Dutra et al., 2020).
14. CorIFA stands for *Corpus de Inglês para Fins Acadêmicos.*
15. An additional language is understood as the language someone learns and uses which is not their first language. (Leffa & Irala, 2014)
16. There are 378 million native speakers and 743 million non-native speakers in the world. https://lemongrad.com/ - English Language Statistics – an Exhaustive List. Accessed on June, 22, 2020.

or written English. This is the focus of 'learner corpus research'[17] (LCR), an umbrella term (Granger et al., 2015) to refer to interlanguage investigations that are based on corpus linguistics. A corpus is a "collection of pieces of language text in electronic form" (Sinclair, 2005: 19) compiled according to some criteria, and representing a language or language variety. Corpus compilation requires establishing precise and broadly-inclusive criteria for the consideration of the mode (e.g.,written), the type (e.g.,a research article) and the domain (e.g., academic) of the texts, the language or language varieties (e.g.,learner English), location of texts (e.g.,compiled in Brazil) and text production dates (e.g., from 2015-2020), according to Sinclair (2005). With a focus on description of language use, corpus linguistics, using a range of linguistic software tools, allows both quantitative and qualitative approaches to learner language. Among its advantages are the capacity to deal with a considerable amount of data, and generalizability of results across similar groups. A well-designed corpus, therefore, is representative of a population. As Biber points out:

> Any selection of texts is a sample. Whether or not a sample is 'representative', however, depends first of all on the extent to which it is selected from the range of text types in the target population; an assessment of this representativeness thus depends on a prior full definition of the 'population' that the sample is intended to represent, and the techniques used to select the sample from that population. (Biber, 1993, p. 243)

A careful corpus design enables generalizations of results, as statistical tests are often used to treat data. In this section, we will highlight some learner corpus research, focusing on their design characteristics and how they have coped with representativeness to be able to make comparisons across registers or groups, for instance.

The design of two of the largest learner corpora compiled in the 1990s (Granger, 1998) are worth mentioning: the International Corpus of Learner English (ICLE) and the Longman Learners' Corpus (LLC), especially due to how they have dealt with language, task and learner-related features (Tono, 2003). Language-related features are mode,

17. Other research areas, such as second language acquisition (SLA) (i.e. Mitchell et al., 2013) and psycholinguistics (i.e., Fernández et al., 2017), also investigate English non-native speakers' production (oral/written).

genre, style and topic. These corpora encompass texts in the written mode with slight differences in the other features: mostly argumentative essays on a previously defined list of topics in ICLE and essays and exam scripts on a variety of topics in LLC. Task-related characteristics concern (a) data collection: sectional rather than longitudinal, (b) type of elicitation: spontaneous as contrasted to prepared or edited texts, (c) production time: either fixed or timed or untimed and done as homework, and (d) use of references allowed, for instance, dictionaries, with such information recorded. As for learner-related features, ICLE is a corpus with texts produced by university level students, while LLC allows for participation of different academic level groups. Both corpora have texts written by learners from a variety of first language backgrounds. While ICLE has high-intermediate to advanced material, LCC allows for the submission of texts at all levels. ICLE was "the first large collection of computerized learner data to be made available for research" while LLC "has been commercially available for research" (Tono, 2003, p. 800). Whereas ICLE has recently released a new version presenting over 5.5 million words (Granger et al., 2020); LLC, with 10 million words (Tono, 2003) does not have such updated information on their website.[18]

Several learner corpus studies use Contrastive Interlanguage Analysis (CIA), which can be understood as the analysis that "involves the comparison between learner language and the target language" (Granger, 2015, p. 13). Learner language has been called 'Interlanguage Varieties' (ILV), especially highlighting "the highly variable nature of interlanguage" (Granger, 2015, p. 18). This approach may compare students' oral or written texts with native speakers' texts (ILV vs. NS) (De Cock et al., 1998, on word combinations in a English learner corpus of French as a first language vs. an English native speaker corpus), interlanguage variety with learners' first language and with native speakers' (ILV vs. L1 vs. NS) (Altenberg & Taper, 1998, on adverbial connectors in Sw-ICLE[19], in a Swedish as L1 corpus and in LOCNESS) or two interlanguage varieties or more (ILV vs ILV) (Bohórquez, 2015, on lexical bundles on Ch-ICLE and Dt-ICLE).[20]

18. http://global.longmandictionaries.com/longman/corpus#aa
19. Sw-ICLE stands for the Swedish subcorpus of ICLE.
20. Ch-ICLE is the Chinese subcorpus of ICLE and Dt-ICLE is the Dutch one.

CIA study results can enhance practitioners' understanding of their students' needs; nevertheless, teachers may choose to collect their own class corpus as a Do-It-Yourself Corpus (DIY corpus)[21] or create data-driven learning (DDL) (Johns, 1991) activities based on ready-made corpora (e.g.,COCA, BNC[22]) or on their DIY corpus. It is worth mentioning that DDL (Johns, 1991, 1994) allows students access to corpus data and concordancing softwares as part of their language-learning process. Using the figure of Sherlock Holmes as a metaphor, Johns (1997) explains that learners are seen as detectives as they are encouraged, for example, to search and identify grammatical rules, vocabulary meaning, collocations and lexico-grammatical patterns, to name a few. Following Johns' (1991, p. 4) DDL format "identify-classify-generalize", learners who participated in Lee's research (2011) had the opportunity to learn and practice prepositions through the analysis of concordance lines. They explored a corpus comprising texts from J.K.Rowling's book *Harry Potter and the Philosopher' Stone*, concentrating on verb-preposition collocates. DDL activities raised students' awareness of the use of prepositions while it helped them to figure out the uses and functions of certain phrasal verbs. DDL contributed to students' language acquisition, proving to be a good way to prepare for exams, as it created a learning context with more enthusiasm and student autonomy (Lee, 2011).

Researchers compiling their own corpus need to adopt strict design criteria. According to Gilquin (2015, p. 16), in the case of learner corpus creation, the rules adopted are "even more crucial, given the highly heterogeneous nature of interlanguage" and such design criteria will be fully addressed in the next section. Moreover, few studies investigated students' texts produced in their own class writing contexts (Staples & Reppen, 2016), a gap that *CorIFA* studies serves to fill as the corpus is compiled from class activities in "English for Academic Purposes." With detailed gathering of learner metadata, and careful consideration of task and language variables, as described in the next section, *CorIFA* allows for thorough studies on classroom contextualized learner writing.

21. Check this page for detailed instructions on how to prepare your own corpus https://www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l04_top.htm.
22. There are several online corpora that can be sources of real language use (e.g., https://www.english-corpora.org/coca/; https://www.english-corpora.org/bnc/).

3. CorIFA: a learner corpus

According to Reppen (2010, p. 33), building a corpus requires a significant time investment as it involves a set of highly interconnected procedures. From collecting the texts to saving, storing, marking-up and adding metadata, the researcher is faced with a vast number of decisions that need to be considered when compiling a corpus. CorIFA was originally created in 2013 at a Brazilian public university. An overview of its compilation history, challenges and shifts is presented in this section. Above all, the corpus objective has remained the same: to describe Brazilian university students' written English interlanguage, as produced in a pedagogical context.

Data collection was inspired, at first, by the International Corpus of Learner English (ICLE). CorIFA and ICLE carry similarities regarding task variables, such as, task medium, genre, topic and task setting. CorIFA compilation started with written tasks: argumentative essays, such as those in ICLE. Students were asked to write essays based on previously chosen topics, such as the internet, feminism, science and technology. Teachers asked the students to write their essays in class or at home, submitting them by email. Essay length, another task variable, was different from ICLE, since the latter required texts to be from 500 to 1000 words; whereas CorIFA allowed, at that time, 200 to 300-word essays. Regarding the learner variables, ICLE and CorIFA participants' age range and learning context are similar; the data for both corpora come from university students who have learned English in a non-English speaking country. The main differences between the two corpora, however, entail first language, academic level, discipline and language proficiency. At this point, we will refer only to first language as the other characteristics will be fully discussed in the following section. ICLE encompasses subcorpora with English texts from speakers of a variety of languages (e.g.,Chinese, Turkish, Portuguese, French, etc.) while CorIFA's participants are mainly Portuguese speakers[23]. Finally, the consent forms were in printed format when the corpus started being collected with basic information from the students, for instance, name

23. CorIFA has a subcorpus of a few texts written by speakers of other languages, such as Spanish, French, that were written by foreign students in exchange programs in Brazil. This subcorpus is not included in the issues addressed in this article.

and enrollment number. As consent forms were modified later on by the research group to collect more metadata, the texts collected in 2013 were discarded.

In 2014 another attempt to collect student texts and compile a learner corpus was made. As the primary goal of having students write texts was pedagogical, many decisions were taken by the subject teachers and, thus, most essays were handwritten. The time and effort demanded to transform the written texts into a digital format led the research group not to include 2014 texts into CorIFA. The group considered that, despite transcriber training, there were risks of misspelling or grammar errors being modified by the person digitizing the documents or by computer spell checkers, leading to texts that would not reflect students' real English level. Since the experience of receiving paper-based texts did not facilitate the process of corpus compilation, from 2015 on, the texts have been collected only in digital format, where students fill in an online form.

In 2015, there was a compilation of texts written in controlled and uncontrolled time settings. First, students submitted texts as part of in-class mock tests, to capture students' skills in writing under time constraints and with a proposed topic. The mock tests were taken by B1 and B2 level students. All of them presented the same instructions regarding text production. Students had to write a 300-word essay (minimum) based on a set topic in 30 minutes. Since digital text collection worked well, the compilation process became standardized, and, from 2016 on, students have submitted texts with distinct registers through online forms, according to their proficiency level, as described in Table 1. Systematic corpus compilation has allowed the research group to keep a sound learner corpus, which will be described in the following section. Before sending their texts, learners are asked to fill in a digital form through *Google Forms* with their information and to read a consent form for their participation in the research, with which they may choose to agree or disagree. This form comprises students' information in a way that helps researchers keep better records of participants' social and linguistic backgrounds, and specificities of the task. Such a consent letter is provided in the Appendix.

CorIFA is composed of texts written by undergraduate and graduate students from different courses at the Federal University of Minas Gerais. These students are registered in one of the five English for Academic Purposes (IFA)[24] subjects created in 2012 as part of a set of initiatives to expand and enhance the internationalization process of the university. Students register according to proficiency levels, ranging from intermediate to advanced (B1-C1), following the Common European Framework of Reference[25] for Languages. As part of each subject's assessment, students from each level are required to write to a specific academic register (Table 1). Before being accepted in one of IFA subjects, students' proficiency level must be checked. Students may either submit scores on a standard proficiency test, such as TOEFL or IELTs, or take an internal placement test.

Table 1 – Academic registers written for each IFA subject.

Subject	Register	Proficiency level
IFA I	Statement of Purpose/Summary	B1
IFA II	Abstract	B1+
IFA III	Argumentative Essay	B2
IFA IV	Literature Review	B2+
IFA V	Research Article/Literature Review	C1

The corpus carries an array of academic registers from statement of purpose to research paper. Students write their texts as course requirements, following the teachers' instructions in terms of number of words and topics. The registers have been gradually distributed, starting with, for example, statement of purposes or summaries in IFA I and, ending with a research paper or literature review in IFA V. After students turn in the first draft text, teachers adjust the teaching of that register to their students' needs. Each subject design includes several exercises on each register and the opportunity for text editing. Students

24. IFA (*Inglês para Fins Acadêmicos*) is the Portuguese equivalent of English for Academic Purposes (EAP).

25. Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press, 2001.

then submit a second and/or third draft that is edited and graded. For the corpus compilation, these texts are categorized into unedited and edited versions.

One of the text variables for CorIFA is length in words. IFA teachers may determine text word ranges based on their experience with the students' level and on register characteristics. Average word length is kept as presented in Table 2, which also shows the total number of words per register and in the whole corpus.

Table 2 – Average word length and total words per register

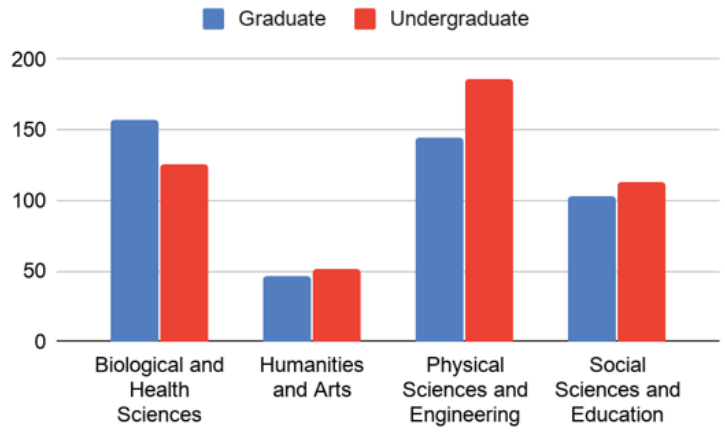
Register	Average number of words	Number of texts	Total number of words
Abstract	208.85	547	114,243
Statement of Purpose	457.26	420	192,051
Summary	225.39	89	21,64
Research Article	1,564.50	20	31,290
Literature Review	552.71	245	135,415
Argumentative Essay	414.94	507	210,375
TOTAL	**553.30**	**1828**	**705,01**

The corpus shows (Table 2) great differences among registers as far as word average length. It consists of six written registers, each ranging in length from 225.39 to 1,564.50 words. The one which surpasses all the others in terms of length in words is the research article. This reflects the reality of the linguistic features among registers, especially related to their nature, which includes physical mode, setting, production circumstances, etc. (Biber & Finegan, 1994), in which some registers do require more words than others. Moreover, since writers must include several sections in a research article, which Biber and Finegan (1994, p. 131) call "standard four-part organization (Introduction, Methods, Results and Discussion - IMRD)" the number of words will certainly surpass other registers in our corpus. The shortest type of text produced by CorIFA participants was the abstract. This register, which could also

be understood as part of a research article, has a clear communicative purpose to sum up the article, presenting its aims, methodology, results, and conclusion. Oftentimes, journals and conferences limit the number of words in an abstract (Swales & Feak, 2009) imposing on the writer the need to be concise.

Another essential aspect carefully planned during corpus compilation was to account for texts per academic level, a learner variable, since participants may either be undergraduate or graduate students (see Figure 1). As the IFA subjects are elective, students at both academic levels can be registered.

Figure 1 – Students per academic level and area

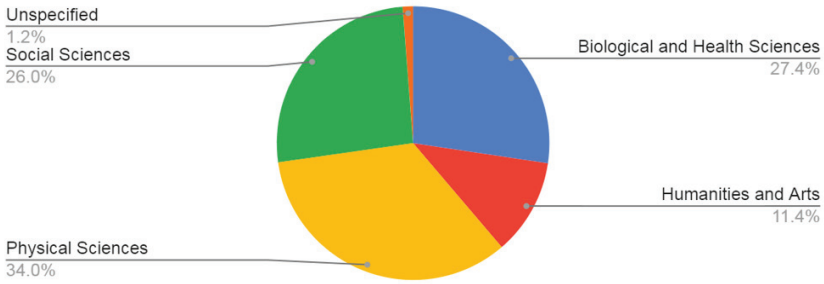


From the first semester of 2016 on, the number of collected samples was quite higher than in the previous year, a situation that remains. The reason that 2015 had the smallest number of samples is due to the compilation process, which were through mock tests, as previously mentioned. Not all students produced the task, many did not give consent to have their texts included in the corpus and some texts did not achieve the minimum number of words. Therefore, fewer texts remained to be part of the corpus.

The greatest number of samples in the corpus comes from students enrolled in courses from the following areas: Physical Sciences and Engineering and Biological and Health Sciences. Humanities and

Arts, on the other hand, constitute the discipline area with the smallest number of samples (see Figure 2). This corpus characteristic is very likely due to the total number of students from Humanities and Arts and Social Sciences and Education enrolled in English for Academic Purposes disciplines being considerably lower than those in the Biological and Hard Sciences fields.

Figure 2 – Texts per students' discipline area



Another characteristic related to the corpus design is its potential for longitudinal studies, as its data can help researchers better understand the relationship between students' writing development and proficiency level. There is a paucity of learner corpus studies from a longitudinal perspective (except for Biber et al., 2020; Goutéraux, 2013; Littré, 2015; Meunier & Littré, 2013). Up to 2019, 217 students submitted texts for more than one semester. Among these, 197 submitted them for a period of one year (two semesters), and 20 for more than one year (3 semesters or more). Interestingly, CorIFA data may come from students who started taking IFA classes as undergraduate students and continued to register after starting a graduate program.

Compiling a learner corpus in a pedagogical context has been challenging, since, for a couple of years, data collection procedures changed, as described. After establishing consistent compilation parameters, the corpus grew steadily. Its task and learner variable complexities allow for a multitude of investigations, some of which will be shown in the next section.

4. Studies based on CorIFA

In this section, we survey the backdrop of studies that employ CorIFA in their research. In the past six years, since the beginning of its compilation, the corpus has been a rich linguistic database for Brazilian researchers. Hitherto, research has mainly centered on learners' language description and on contrasting learners' written interlanguage with data from other corpora using CIA (as pointed out in section 2). Relying on two main academic genres: argumentative essays and abstracts, the studies focus on the understanding of learners' use of English as they are at different proficiency levels and also on detecting their underuse or overuse of specific linguistic features. Most studies use a reference corpus composed of well-evaluated non-native speakers' essays or native-speakers' texts. The topics encompass linking adverbials, collocations, *that*-clauses, conjunctions, noun phrases, and passive constructions. For organizational reasons, first, we bring an overview of one descriptive study and five investigations that fall under the CIA perspective. This section ends with a CorIFA-based study that highlights a pedagogical intervention, leading to reflections on applications of learner corpus research.

Focusing on the interlanguage itself, Queiroz (2019) explores CorIFA deeply to shed light on novice writers' use of noun phrases. These phrases have been regarded as a common linguistic feature in expert academic texts, mainly research articles (Biber et al., 2009; Parkinson & Musgrave, 2014; Gray, 2015; Biber & Gray, 2016). To investigate the grammatical complexity of noun phrases (NPs), the study analyzes general topic essays and specific topic essays[26] from a CorIFA subcorpus and provides a thorough description of pre and post-modification of the types: *adjective + noun* and *noun + prepositional phrase*. Two of the study's findings are particularly noteworthy. First, surprisingly, the subcorpus analysis of upper-intermediate level texts

26. "General topic texts were those in which the EAP instructor presented a topic or question, e.g., Does technology make us more alone? to all students to write an argumentative essay about. Many of these topics were similar to the ones used in English proficiency tests. On the other hand, specific topic texts were those in which students were allowed to choose a topic of their preference to write about. Many wrote essays about their graduate studies, such as one dentistry student who wrote about periodontal disease and premature delivery." Queiroz (2019, p. 57).

revealed a higher use of complex NPs (59.3%) than simple ones (35%). Second, NPs were more frequent in the specific topic essays, which is interpreted as quite positive writing practice as complex NPs are characteristic of academic registers: "it can be assumed that Brazilian learners due to their proficiency level [B2], the academic context of writing, and the probable contact with specialized texts in English from their own disciplines, are capable of using structurally complex and compressed phrasal structures, often characteristic of professional academic writing" (Queiroz, 2019, p. 112). Above all, this last result shows that discipline-specific tasks can propel writing at the university level that is more suitable to the academic context. This research makes evident the potential of descriptive corpus-based research to contribute to applied linguistics, in this case, EAP. The corpus design allowed a task variable (specific topic versus general topic task) to emerge as the one that affected NP use by Brazilian upper-intermediate level university novice writers.

Three of the CIA studies carried out based on CorIFA are on learners' use of conjunctions, which have been considered key features in text coherence (Halliday & Hasan, 1976; Chen, 2006; Liu, 2008; Zihan, 2014). Yet, these linguistic features "are not always needed and (...) they have to be used with discrimination" (Altenberg & Tapper, 1998, p. 80), posing difficulties to learners as they "tend[s] to vary from one language and culture to another" (Altenberg & Tapper, 1998, p. 81). Dutra et al. (2017 and 2019) and Santos (2008) used a CorIFA subcorpus of B1 level argumentative essays, contrasting it to LOCNESS. Despite the fact that these investigations dealt with different conjunctions [Dutra et al. (2017) addition words (*besides* and *also*), Dutra et al. (2019) result markers (*thus*, *so* and *therefore*) and Santos (2018) contrasting connectors (*but* and *however*)], they all detected either substantial quantitative differences (underuse or overuse as compared to LOCNESS or MICUSP), sentence position disparities, as well as discourse function inadequacies on the part of the learners. For instance, the marker *so* is used three times more in CorIFA than in *LOCNESS*, assuming beginning sentence functions of initiating a topic or announcing that an idea is going to be presented again (DUTRA et al., 2019), which have been attested as oral discourse markers (Carter & McCarthy, 2006). An interpretation shared by the three CorIFA investigations based on learners' overuse of conjunctions in sentence

initial position is that register awareness needs to be better addressed in the Brazilian university context.

The use of verbs was the focus of two other CorIFA CIA studies. While Guedes (2017) concentrates on the most frequent academic verbs present in argumentative essays in CorIFA, comparing verb usage to that in British Academic Written English corpus (*BAWE*), Nunes and Orfanò (2020) analyze verbs in research abstracts within the system of transitivity in passive *that-clauses*, contrasting the learner corpus results to a *Lingua Franca Corpus*[27] composed of abstracts from soft and hard sciences. Little action verb variation and low frequency of *adverb + verb* collocations in CorIFA essays, as compared to BAWE, showed Brazilian learners' lack of familiarity with the academic register (Guedes, 2017). As for the verbs used in abstracts, Nunes and Orfanò (2020) also found the learners' choices of verbs inadequate. They use more mental verbs (*conclude, verify* and *assume*) while expert writers show preference for relational verbs (*show, demonstrate* and *reveal*). Another observed result is that learners delete the conscious agent and thus remain distant from the findings they discuss. Researchers, on the other hand, seem to be more actively conscious of the object of study they investigate, showing more intellectual authority. Developing authorship is a skill that needs to be included in EAP classes; corpus-driven linguistic analysis like this one and the study on verb variation and collocations can inform teachers in their practice.

The studies with our learner corpus described up to this point show the extent to which specific corpus analysis can shed light on the understanding of learners' linguistic needs, considering the register they wrote, their proficiency level and type of tasks. The ultimate general goal of developing such studies is catering to learners' exact difficulties because more precisely designed activities can be prepared and the course syllabus redesigned. A good example of a combination of interlanguage analysis and classroom activities was developed by Alves and Pinto (2018). First, the study investigated how results and conclusions were reported in abstracts in two apprentice corpora: CorIFA and MICUSP. CorIFA analysis provided real examples of

27. The reference corpus called *English Lingua Franca* Corpus (Nunes & Orfanò, 2020) consists of abstracts taken from journals belonging to three main areas: Life Sciences, Exact Sciences and Human Sciences.

learner language and the access to MICUSP allowed students to experience a corpus linguistics pedagogical practice: *Do-it-Yourself* (DIY) corpora (McEnery *et al.*, 2006). Students compiled their own study corpus and were able to raise their awareness on how to improve the final two rounds of abstracts. MICUSP was a suitable corpus for DIY since they are formed by well-evaluated university papers and the online framework allows for the user to choose discipline-specific texts.

The agenda for the future is promising and shall contemplate among other issues, advances in research methodologies, granting better access to students' interlanguage, longitudinal studies and description variation across registers and discipline. Furthermore, it is of utmost importance that DDL activities (Johns, 1991) are more present in the language classroom, providing information on their advantages and limitations, and thus transferring research data into direct application, as in Almeida et al. (in press).

5. Conclusion

The overall aim of this paper was to present the rationale behind building an academic learner corpus, making the case that such a process is paramount for revealing traces of learners' written interlanguage. The main principles regarding corpus compilation were presented associated with the design criteria adopted for CorIFA. After that, we outlined the methodological procedures that were followed in the compilation process. The texts included in the corpus and the register associated with the students' proficiency level were also explained indicating a vast range of topics for future studies.

Subsequently, descriptive and CIA research based on data from CorIFA were presented to illustrate the contributions that the corpus has already provided for researchers interested in learner interlanguage. The studies pinpointed in this paper serve to reinforce the claim that compiling and observing a learner corpus can be an invaluable resource for language teachers keen to enhance their understanding of learners' output, enabling them to make more accurate pedagogical decisions for their classes.

Acknowledgements

This article presents research developed by the Grupo de Estudos em Corpora Especializados e de Aprendizes (GECEA)

Conflict of interests

The authors declare they have no conflict of interest.

Credit Author Statement

We, Deise Prina Dutra, Bárbara Malveira Òrfano, Annallena de Souza Guedes, Jessica Ceritello Alves, and João Gabriel Fekete, hereby declare that we do not have any potential conflict of interest in this study. We describe below how each author participated in the preparation of the article: 1) conceptualization of the study (Deise Prina Dutra, Bárbara Malveira Òrfano, Annallena de Souza Guedes); 2) literature review (Deise Prina Dutra; Bárbara Malveira Orfanò, Annallena de Souza Guedes; Jessica Ceritello Alves); 3)formal data analysis (Deise Prina Dutra; João Gabriel Fekete, Jessica Ceritello Alves) and 4) article writing and final editing (Deise Prina Dutra; Bárbara Malveira Orfanò, Annallena de Souza Guedes; João Gabriel Fekete; Jessica Ceritello Alves). All authors approve the final version of the manuscript and are responsible for all aspects, including the guarantee of its veracity and integrity.

References

Almeida, V. C., Orfanò, B. M., & Dutra, D. P. (in press). In V. Viana (Ed.), Is there a better choice? Raising learners' awareness of academic collocations. *Teaching English with corpora: A resource book*. Routledge.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80-93). Pearson Education.

Alves, A. L. L., & Pinto, P. T. (2018). A utilização de that-clauses em abstracts escritos por alunos-pesquisadores brasileiros. *Entrepalavras*, *8*(2), 288-303. http://dx.doi.org/10.22168/2237-6321-21112.

Anthony, L. (2016). AntConc (Version 3.4.3) [Computer Software]. Waseda University.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*(4), 243-257. https://doi.org/10.1093/llc/8.4.243.

Biber, D., Reppen, R., Staples, S., & Egbert, J. (2020). Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. *International Journal of Learner Corpus Research*, *6*(1), 38-71. https://doi.org/10.1075/ijlcr.18007.bib.

Biber, D., & Conrad, S, (2009). *Register, genre and style.* Cambridge University Press.

Biber, D., & Finegan, E. (1994). Intra-textual variation within medical research articles. In N. Oostdijk, & P. de Haan (Eds.), *Corpus-based research into language* (pp. 201-221). Rodopi.

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Biber, D., Grieve, J., & Iberri-Shea, G. (2009). Noun phrase modification. In G. Rohdenburg, & Schlüter, J. (Eds.), *One language, two grammars? Differences between British and American English* (pp. 182-193). Cambridge University Press.

Bohórquez, C. G. (2015). *Eliminação de pacotes lexicais relacionados ao tópico e de pacotes lexicais em contexto de sobreposição: uma proposta metodológica para os estudos da linguística de corpus* [Unpublisehd Master's thesis]. Universidade Federal de Minas Gerais.

Carter, R., & McCarthy, M. (2006). *Cambridge Grammar of English: Spoken and written English grammar and usage*. Cambridge University Press.

Chen, C. W. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics*, *11*(1),113–130. https://doi.org/10.1075/ijcl.11.1.05che.

Common European Reference Framework for languages: learning, teaching, assessment. (2001). Council of Europe. https://rm.coe.int/1680459f97 (accessed December 7, 2021).

Crawford, W. J., & Csomay, E. (2016). *Doing corpus linguistics*. Routledge.

De Cock, S., Granger, S., Leech, G., & McEnery, T (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67-79). Pearson Education.

Dutra, D., Queiroz, J. M. S., & Alves, J. C. (2017). Adding information in argumentative texts: a learner corpus-based study of additive linking adverbials. *Estudos Anglo-Americanos*, *46*(1), 9-32.

Dutra, D. P., Orfanò, B. M., & Almeida, V. C. (2019). Result linking adverbials in learner corpora. *Domínios de Lingu@gem*, *13*(1), 400-431. https://doi.org/10.14393/DL37-v13n1a2019-17.

Dutra, D. P., Queiroz, J. M. S., Macedo, L. D. de, Costa, D. D., & Mattos, E. (2020). Adjectives as nominal premodifiers in Chemistry and Applied Linguistics Corpora. In U. Römer, V. Cortes, & E. Friginal (Eds.), *Advances in Corpus-based Research on Academic Writing Effects of discipline, register, and writer expertise* (pp. 205-226). John Benjamins Publishing Company. https://doi.org/10.1075/scl.95.09dut.

Fernández, E. M., Souza, R., A., & Carando, A. (2017). Bilingual innovations: Experimental evidence offers clues regarding the psycholinguistics of language change. *Bilingualism*: *Language and Cognition*, *20*(2), 251-268. https://doi.org/10.1017/S1366728916000924.

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 9-34). Cambridge University Press.

Gilquin, G., De Cock, S., & Granger, S. (2010). *Louvain international database of spoken English interlanguage*. Presses Universitaires de Louvain.

Goutéraux, P. (2013). Learners of English and Conversational Proficiency. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of Learner Corpus Research: Looking Back, Moving Ahead* (pp.197-210). Presses Universitaires de Louvain.

Granger, S. (1998). The computerized learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer*. Longman.

Granger, S. (2015a). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 486-510). Cambridge University Press.

Granger, S. (2015b). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7-24. https://doi.org/10.1075/ijlcr.1.1.01gra.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *International corpus of learners English: Version 3*. Presses Universitaires de Louvain.

Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only art of the story*. John Benjamins Publishing Company.

Guedes, A. de S. (2017). *Verbos do inglês acadêmico escrito e suas colocações: um estudo baseado em um corpus de aprendizes brasileiros de inglês* [Unpublished PhD Dissertation]. Universidade Federal de Minas Gerais.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Johns, T. (1991). Should You Be Persuaded: Two Examples of Data-driven Learning. In T. Johns, & P. King (Eds.), *Classroom Concordancing*. *ELR Journal*, *4*, 1-16.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 293-313). Cambridge University Press.

Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100-115). Routledge.

Lee, H. (2011). In defense of concordancing: An application of data-driven learning in Taiwan. *Procedia-Social and Behavioral Sciences*, *12*, 399-408. https://doi.org/10.1016/j.sbspro.2011.02.049.

Leffa, V. J., & Irala, V. B. (2014). O ensino de outra(s) língua(s) na contemporaneidade. In V. Leffa, & V. B. Irala (Eds.), *Uma espiadinha na sala de aula: ensinando língua adicionais no Brasil* (pp. 21-48). Educat.

Littré, D. (2015). Combining Experimental Data and Corpus Data: Intermediate French-speaking Learners and the English Present. *Corpus Linguistics and Linguistic Theory*, *11*(1), 89-126.

Liu, D. (2008). Linking adverbials: An across-register corpus study and its implications. *International Journal of Corpus Linguistics*, *13*(4), 491-518. https://doi.org/10.1075/ijcl.13.4.05liu.

McEnery, T.; Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge.

Meunier, F., & Littré, D. (2013). Tracking Learners' Progress. Adopting a Dual 'Corpus Cum Experimental Data' Approach. *The Modern Language Journal*, *97*(1), 61-76.

Mitchell, R., Myles, F., & Marsden, E. (2013). *Second language learning theories*. Routledge.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. *How to use corpora in language teaching*, *12*, 125-156. https://doi.org/10.1075/scl.12.11nes.

Nunes, L. P., & Orfanò, B. (2020). Investigating the system of transitivity in passive that-clauses of research abstracts. In N. Kenny, & L. Escobar (Eds.), *The changing face of ESP in today's classroom and workplace* (pp. 63-177). Vernon Press.

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Oliveira, S. B., Fonseca, M. M. S., & Marques, D. S. (2017). Collaborative writing: the English language learners gaze on Art. *Fórum Linguístico*, *14*, 21-52. https://periodicos.ufsc.br/index.php/forum/issue/view/2513/showToc (accessed June 23, 2020).

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, *14*, 48-59. https://doi.org/10.1016/j.jeap.2013.12.001.

Queiroz, J. M. S. (2019). *The Grammatical Complexity of English Noun Phrases in Brazilian Learner's Academic Writing: A Corpus-based Study*. [Unpublished master thesis]. Universidade Federal de Minas Gerais. http://www.poslin.letras.ufmg.br/defesas/1980M.pdf (accessed 15 December, 2021).

Reppen, R. (2010). Building a corpus. *The Routledge Handbook of Corpus Linguistics.* Routledge.

Santos, M. A. (2018). *Descrição do uso das conjunções but e however em redações acadêmicas em língua inglesa de nível B1 com base em corpus*. [Unpublished master thesis]. Universidade Estadual Paulista Júlio de Mesquita Filho. https://repositorio.unesp.br/handle/11449/154076 (accessed 15 December, 2021).

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins Publishing.

Selinker, L. Interlanguage. (1972). *International Review of Applied Linguistics*, *10*(3), 209-231. http://dx.doi.org/10.1515/iral.1972.10.1-4.209.

Sinclair, J. (Ed.). (1993). *Collins COBUILD English Grammar.* Collins.

Sinclair, J. (2005). Corpus and text-basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Guide to Good Practice (pp. 1-16). Oxbow Books.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of second language writing, 32*, 17–35. https://doi.org/10.1016/j.jslw.2016.02.002.

Swales, J., & Feak, C. (2009). *Abstracts and the writing of abstracts*. Michigan University Press.

Tono, Y. (2003). Learner corpora: design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 800-809).

Xavier, A. D., Oliveira, S. B., & Souza, E. L. M. (2019). A construção de memes como ferramenta de ensino da língua inglesa*. Periferia, 11*, 140-161. https://doi.org/10.12957/periferia.2019.36440.

Zihan, Y. (2014). *Linking adverbials in English*. [Unpublished PhD Dissertation]. Victoria University of Wellington.

Recebido em: 16/07/2020
Aprovado em: 15/12/2021

Appendix

CARTA DE CONSENTIMENTO LIVRE E ESCLARECIDO:
Para os participantes

Caro(a) Senhor(a)

A coordenação das disciplinas de "Inglês para Fins Acadêmicos" da UFMG conduz pesquisas que tem por objetivo estudar o desenvolvimento das habilidades de leitura, de escrita, de audição e de fala de aprendizes de inglês para fins acadêmicos. Cada projeto de pesquisa está devidamente autorizado pela Câmara de Pesquisa da Faculdade de Letras da UFMG.

A fim de que os projetos possam ser desenvolvidos, é necessária a sua autorização, vez que as pesquisas constarão da coleta das suas redações produzidas enquanto aluno do curso. A sua participação nesta pesquisa é voluntária e não determinará qualquer risco nem tratá desconfortos. Além disso, sua participação é importante para o aumento do conhecimento a respeito dos processos de aquisição e desenvolvimento das quatro habilidades supracitadas por alunos universitários brasileiros, podendo beneficiar outros alunos futuramente na melhoria do ensino de língua inglesa no nível superior.

Informamos que o/a Sr(a). tem a garantia de acesso, em qualquer etapa dos estudos, sobre qualquer esclarecimento de eventuais dúvidas. Se tiver alguma consideração ou dúvida sobre a ética da pesquisa, entre em contato com a coordenação do programa (3409-3839) ou com o Comitê de Ética em Pesquisa (CoEP) da Universidade Federal de Minas Gerais, situado na Av. Antônio Carlos, 6627. Unidade Administrativa II - 2º andar - Campus Pampulha, telefone 3409-4592 / 3409-4027.

Também é garantida a liberdade da retirada de consentimento a qualquer momento e deixar de participar do estudo.

Fica também garantido que as informações obtidas serão analisadas em conjunto com as de outras pessoas, não sendo divulgada a identificação de nenhum dos participantes.

O/A Sr(a). tem o direito de ser mantido atualizado sobre os resultados parciais das pesquisas e caso seja solicitado, todas as informações que solicitar lhe serão fornecidas.

Não existirão despesas ou compensações pessoais para o participante em qualquer fase dos estudos. Também não há compensação financeira relacionada à sua participação.

Os participantes das pesquisas comprometem-se a utilizar os dados coletados somente para pesquisa, e os resultados serão veiculados através de artigos científicos, em revistas especializadas e/ou em encontros científicos e congressos, sem nunca tornar possível a sua identificação.

Abaixo se encontra o Termo de Consentimento Livre e Esclarecido, para concordância caso não tenha ficado qualquer dúvida.

Deise Prina Dutra - Coordenadora Geral do IFA/UFMG

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO *

Acredito ter sido suficientemente informado a respeito dos estudos conduzidos pela coordenação do Programa Inglês para Fins Acadêmicos/UFMG. Ficaram claros para mim quais são os propósitos dos estudos, os procedimentos a serem realizados, as garantias de confidencialidade e de esclarecimentos permanentes. Ficou claro, também, que a minha participação é isenta de despesas e que tenho garantia do acesso aos resultados e de esclarecer minhas dúvidas a qualquer tempo. Concordo voluntariamente em participar e estou ciente de que poderei retirar o meu consentimento a qualquer momento sem penalidade ou prejuízo ou perda de qualquer benefício que eu possa ter adquirido.

- ◯ Concordo
- ◯ Discordo