E-JournALL, EuroAmerican Journal of Applied Linguistics and Languages
Volume 10, Issue 2, December 2023, 9-31
ISSN 2376-905X
http://doi.org/10.21283/2376905X.1.10.2.2748
www.e-journall.org

From learner corpus to data-driven learning (DDL): Improving lexical usage in academic writing

SHARON HARTLE
Università degli Studi di Verona

Received 19 May 2023; accepted after revisions 14 October 2023

ABSTRACT

EN Despite considerable discussion in the literature (Flowerdew & Peacock, 2001; Hyland, 1998; Tang, 2012) competent English academic writing is still a problem which needs to be solved. English for Academic Purposes (EAP) teaching often focuses on specialized lexis, which may, however, be the area where academic writers need least help. The study of a small corpus of C2 level academic writing which consisted of the sub-genres of summary and discussion writing revealed that one key area which is problematic is collocation. This paper presents the results of this small corpus investigation into learner language and how it informed the classroom implementation of data-driven learning (DDL) to increase learner awareness of and ability to use collocations effectively in written academic English. The article briefly describes the corpus and the resulting teaching procedure adopted. The first step of this procedure is familiarization followed by experimentation using Sketch Engine (SkeLL).

Key words: LEARNER CORPORA, DATA-DRIVEN LEARNING, ENGLISH FOR SPECIFIC ACADEMIC PURPOSES (ESAP), ACADEMIC LEXIS ANALYSIS, COLLOCATION

ES A pesar de una considerable discusión en la literatura (Flowerdew & Peacock, 2001; Hyland, 1998; Tang, 2012), la escritura académica competente en inglés sigue siendo un problema que necesita ser resuelto. La enseñanza de Inglés con Fines Académicos (EAP) a menudo se centra en léxico especializado, que, sin embargo, podría ser el área donde los escritores académicos necesitan menos ayuda. El estudio de un pequeño corpus de escritura académica de nivel C2, que constaba de los subgéneros de resumen y de discusión, reveló que una de las áreas problemáticas clave es la colocación. Este artículo presenta los resultados de esta pequeña investigación de corpus sobre la lengua de los y las aprendices, y explica cómo informó la implementación en el aula del aprendizaje basado en datos (Data driven learning-DDL) para aumentar la conciencia del aprendiz y su capacidad para usar colocaciones de manera efectiva en el inglés académico escrito. El artículo describe brevemente el corpus y el procedimiento de enseñanza resultante adoptado. El primer paso de este procedimiento es la familiarización, seguida de la experimentación con el uso de Sketch Engine (SkeLL).

Palabras claves: CORPUS DE APRENDICES, APRENDIZAJE BASADO EN DATOS, INGLÉS CON FINES ACADÉMICOS ESPECÍFICOS (ESAP), ANÁLISIS DEL LÉXICO ACADÉMICO, COLOCACIÓN

IT Nonostante l'esistenza di un'ampia discussione in letteratura (Flowerdew & Peacock, 2001; Hyland, 1998; Tang, 2012), la competenza nella scrittura accademica in inglese rimane ancora un problema da risolvere. L'insegnamento dell'Inglese per Scopi Accademici (EAP) spesso si concentra su lessico specializzato, che potrebbe, tuttavia, essere l'area in cui gli accademici necessitano di meno aiuto. Lo studio di un piccolo corpus di scrittura accademica di livello C2, composto dai sottogeneri di riassunto e discussione, ha rivelato che una delle aree problematiche è la collocazione. Questo articolo presenta i risultati di questa indagine su un piccolo corpus di lingua appresa e come essa ha orientato l'attuazione in classe dell'apprendimento guidato dai dati (Data driven learning - DDL) per aumentare la consapevolezza degli apprendenti e la capacità di utilizzare le collocazioni in modo efficace nell'inglese accademico scritto. L'articolo descrive brevemente il corpus e la procedura di insegnamento risultante adottata. Il primo passo di questa procedura è la familiarizzazione, seguita da sperimentazione con l'uso di Sketch Engine (SkeLL).

Parole chiave: CORPUS DI APPRENDENTI, APPRENDIMENTO GUIDATO DAI DATI, INGLESE PER SCOPI ACCADEMICI SPECIFICI (ESAP), ANALISI DEL LESSICO ACCADEMICO, COLLOCAZIONE

✉ **Sharon Hartle,** University of Verona
sharon.hartle@univr.it

1. Introduction

Academic work submitted to peer review often meets with requests for non-native speaker authors to seek native speaker revision of the English. The aim of this article is not to debate the whys and wherefores of peer review but, nevertheless, at least one interesting conclusion may be drawn from this common request. This is the fact that the need for both university students and academics to improve their level of written English is still very much an issue. This has been discussed widely in the broader ESP literature (Flowerdew & Peacock, 2001; Hyland, 1998; Tang, 2012) but is still a problem, which is far from being solved (Ennis & Mikel Petrie, 2020; Hartle et al., 2022; Littlewood, 2014). The use of appropriate lexis is of particular importance when establishing an academic register, and awareness of academic lexis is key when developing academic skills such as reading (Hyland & Tse, 2009). This is certainly the case when writing, as academic writing is a register that needs to be learned since there are no "native users of academic language" (Lew et al., 2018). English for Academic Purposes (EAP) teaching often focuses on the development of lexis but this tends to be the specialized lexis required for specific disciplines, which may be the area where academic writers need least help. What may cause considerable difficulty, on the other hand, is lexical usage (Hyland 2006; Flowerdew, 2015). For this reason, the Learner Corpus 22 (LC22), which this paper draws on, was developed at the University of Verona, Department of Foreign Languages and Literatures, as the central component of a learner language production study, undertaken to inform the development of a learning design model for academic English writing courses coordinated by our department. It is a small corpus of C2 level academic writing, produced by post-graduate students, which consists of the sub-genres of summary and discussion writing. Although the project is still in its infancy and the corpus is to be extended to provide a diachronic view of learner production, it provided us with a cross-sectional snapshot of some of the strengths and weaknesses of learner writing. Initial findings revealed the problematic usage of lexis, and in particular effective word choice and collocation. This, in turn, led us to choose a data-driven learning (DDL) approach, which we hoped would also develop learner independence, for our courses.

This paper focuses on the key, initial results of this small corpus investigation into learner-generated language and the way in which the findings, subsequently, informed the classroom implementation of data-driven learning. The aim was to help students of EAP improve their awareness of and ability to use collocations effectively in written academic English. The article briefly describes the research design of the academic writing study itself, together with the corpus, its compilation, and its analysis. The second part of the article describes the resulting two-step learning design, which was implemented in English for Specific Academic Purposes (ESAP) courses. In this approach, learners are firstly familiarized with common collocation errors and secondly, they are introduced to corpus interfaces designed for language learning. A guided discovery approach (Bruner, 1961) is used to help them to experiment with DDL to improve their own use of collocations when writing. The main tool used for this is the Sketch Engine for language learning (SkeLL)[1], a freely accessible interface, which enables tailored web searches for a range of collocations and synonyms.

2. Collocation in language learning

Linguists traditionally define collocation as co-occurrence over a range of a few words to either side of a specific item (Halliday, 1994; Sinclair, 1991), without necessarily considering their semantic properties (Macis & Schmitt, 2017). From a pedagogical viewpoint, however, the focus is rather on collocation as phraseological, lexical combinations which co-occur, but which also have a reciprocal relationship of varying degrees. These are largely determined by convention, such as 'miss' as a verb which collocates with the noun 'train', rather than 'lose' which Italian L1 speakers may choose, transferring 'perdere' into the English 'lose' without considering that the meaning will vary in the collocation. Mutual reciprocity, first suggested by Firth, with his widely cited phrase "You shall know a word by the company it keeps" (Firth, 1957) may refer to different distributions, but the most useful notion for learners is that the meanings are created by the reciprocity of the collocates. This reciprocity may be considered to fall on a continuum from weak to strong (Conzett, 2001), or free to restricted (Nesselhauf, 2003), where, for instance, 'a friendly dog' has weak reciprocity. In an expression such as to 'throw in the towel', on the other hand, the reciprocity is strong, depending on what Conzett refers to as the expectation created by one element in the collocation that the other will co-occur.

[1] Sketch Engine for Language Learning (SkeLL) https://www.sketchengine.eu/skell/skell-web-interface-for-english-language-learning/ (last accessed April 8, 2023).

2.1. Advanced L2 learner Usage Problems

L2 learner difficulty with collocation has often been seen to be problematic in language production (Bahns & Eldaw, 1993; Durrant & Schmitt, 2009; Granger & Bestgen, 2014) particularly at intermediate to advanced levels. Two aspects which emerge from the literature, are, firstly, the question of frequency, in that low frequency, or rare collocations, tend to be under-used (Granger & Bestgen, 2014). Secondly, restriction, as mentioned above, is also an issue (Durrant & Schmitt, 2009). What is noticeable from the results of such research, is, also, that learners tend to mismatch those collocations that are in the middle of the free-restricted range. This refers to the choice of constituent parts that can combine with other elements as well, such as 'missing' and 'trains', where the collocation cannot be considered fixed. 'Miss', indeed, may collocate with a range of items, and has different meanings resulting from the reciprocity relationships: 'miss (feel nostalgic) my friends', 'miss (not do something in time) the deadline', 'miss (not reach) a target'. Conzett (op. cit., p. 70) advises focusing pedagogically on medium strength items, which, when combined, are possibly more useful for learners than extremely rare or fixed items, questioning a tendency to focus on teaching low frequency items at advanced levels. The question of which items to focus on is also problematic (Timmis, 2008); therefore, it may be more appropriate to increase learner independence, so that they can exercise their own agency in choosing the collocations that they themselves need.

2.2. EAP in a digital post COVID-19 age

Discussion of the effectiveness of technology and digital tools in language teaching, over the years, has often referred to the twin aspects of learner autonomy and agency – which have emerged also from Emergency Remote Teaching (ERT) studies (Green et al., 2020; Whittle et al., 2020). Definitions of autonomy range from Holec's notion of complete responsibility for learning being in the learner's own hands (Holec, 1981) to more nuanced interpretations (Benson, 2007; Little, 1991). Agency is linked to autonomy, and implies, to some extent, the learner's investment in their own learning (Bourdieu & (Translated by) Nice, 1984; Norton, 2013). In the case of our learners the investment is particularly strong as their future careers may depend on their ability to publish in academic English. Ahearn views agency as "the socioculturally mediated capacity to act" (Ahearn, 2001 p. 112) and in a pedagogical context this means both providing opportunities for learner involvement in the process, and developing mutual mediation between teachers and learners in that process (Hartle, 2020; Larsen-Freeman, 2019; Van Lier, 2008). It has long been recognized that an effective lexical repertoire means more than simply having knowledge of word meanings in particular contexts, but of developing an awareness of lexical complexity (Lewis, 1993; I. S. P. Nation, 2001; Shin & Nation, 2008; Timmis, 2008). Familiarizing learners, therefore, with digital tools that may increase independence in the development of their own lexical repertoires, may aid competence as well as increasing both agency and investment and, ultimately, lead to improved performance on academic writing tasks. Our courses aim to foster a DDL approach to the study of lexis for precisely this reason.

2.3. Data-driven Learning

The term DDL, coined by Tim Johns at the end of the twentieth century (Johns, 1986, 1991), advocates an enquiry-based approach involving learner corpora searches designed to answer questions related to language usage. This involves the use of computational tools to analyze corpora data, allowing learners to identify and study frequent word combinations and patterns of language use (Chen & Baker, 2010). DDL has proved to be particularly effective for learning academic language, as it enables identification and use of key academic lexis and collocation patterns (Biber et al., 2004; Hyland, 2008). Although the approach was initially met with enthusiasm (Boulton, 2017; Sinclair, 2004), this waned at the turn of the century due partly to restricted access to many corpora, and to the investment into developing the skills required for teachers and learners to use the corpus analysis tools available at that time. Considerable effort was needed to choose appropriate, naturally occurring language for pedagogical use, particularly in international contexts (Prodromou, 1996; Widdowson, 1991), which also detracted from its popularity. The advent of digital interfaces, however, that enable ease of access to corpora, may make DDL once more attractive nowadays both to teachers and learners (Boulton, 2017). User-friendly interfaces, which are widely available, may be extremely valuable for learners seeking to develop their awareness and use of lexical patterning, particularly where collocation is concerned.

3. Background to the study

As mentioned in the Introduction, our original hypothesis regarding the problematic use of academic English by academics at all levels, was that the issues were probably related not only to grammar but also to lexis. One measure that may alleviate this is intervention at an early stage, helping young academics to develop effective strategies that support their writing (Barnau & Ferková, 2022; Basturkmen & Wette, 2016). Academic English courses provided by our department tend to be limited, 40-hour intensive courses, which are offered annually to postgraduate students, although general English courses are also available to them in the university language centre. As a result of this limited duration and from the reflections of past course participants, (Hartle & Cavalieri, forthcoming) a key aspect of such short courses, is one of increasing learner autonomy, an essential consideration for our learning design. Before we could develop the teaching model itself, however, the target content also needed to be identified clearly. This was a case of identifying both strengths that could be encouraged in learner writing and weaknesses that needed to be addressed. Consequently, it was decided to conduct a pilot study based on a local corpus, which aimed to answer two main research questions:

1) What are the principal grammatical and lexical strengths in advanced, learner academic writing?
2) What are the principal grammatical and lexical weaknesses in advanced, learner academic writing?

At a later stage another question was added to these to aid the development of our teaching model:

3) What resources may be developed to aid lexical acquisition and competence?

Drawing on both Corpus Linguistics and thematic analysis (Clarke & Braun, 2014), the research methodology adopted was a mixed methods approach, chosen because, as stated by Dörnyei (2007) "a mixed methods inquiry offers a potentially more comprehensive means of legitimizing findings than do either QUAL or QUAN methods alone." (p. 62). The quantitative analysis, in our case, was the study of the corpus data. The theoretical framework adopted to conduct the study was Computer aided error analysis (CEA) (Dagneaux et al., 1998), which was adapted to meet local needs. Our methodology was an adaptation of traditional CEA, focused on identifying errors, which we altered because we aimed to determine effective, as well as problematic, language choices. In our annotation of the corpus, therefore, we focused not only on errors but also on effective language choices that were higher than the production expected at a B1 level. The annotation process will be described in greater detail in Section 4.2.

3.1. The corpus LC22 and the reference corpora

The LC22 corpus itself is small, which is characteristic of learner corpora for various reasons including the labour intensive nature of CEA annotation, the fact that it was part of a local, pilot study but also because specialized, smaller corpora may reveal "context-specific aspects of discourse, which are not always evident in larger ones" (O'Keefe et al., 2007, p. 182). Small corpora are, indeed, widely used to investigate specific discourse patterns, including patterns to inform learning design (Bondi, 2001). In our case, the corpus data required interpretation, if the results were to reflect the reality of the learners involved in producing the texts. Corpora, such as the BNC or COCA, for instance, tend to classify Latin-derived lexis as having less frequent occurrences, but in our context, where the main L1 being used is Italian, Latin-derived lexis is common, even though its choice may not always reflect appropriate usage. Frequency has generally been considered an effective indicator of difficulty for the acquisition of lexis for some time now, as it is thought to be "a rational basis for making sure that learners get the best return for their vocabulary learning effort" (Nation & Waring, 1997, p. 15). This, however, would mean correlating less frequent occurrences with increased difficulty, which, with Latin-derived items, such as "incapable", "integration" or "radical"[2] to list just a few, may not be the case in our context.

The next question was which reference corpus to choose. McEnery and his colleagues (2019) speak of a disconnect between Second Language Acquisition (SLA) research and Learner Corpus Research (LCR). They refer to a 2019 special edition of *The Modern Language Journal*[3] entitled *SLA Across Disciplinary Borders*. This issue considers SLA across different disciplines including corpus linguistics (Duff & Byrnes, 2019). The volume

[2] Items extracted from the LC22 corpus.
[3] The Special Issue is available in open access at this link: (last accessed July 14, 2023).

consisted of 15 contributions in all, from which three papers refer to corpus linguistics. Despite the widely held belief that learner corpora should inform language learning and teaching, not one of these articles referred to learner corpora. This suggests that the reference corpora that educators and course developers commonly refer to are still native speaker (NS) ones. Identifying effective language production may, indeed, not be a question of complying with often unrealistic NS norms, but of being able to communicate effectively (Graddol, 2006; Kirkpatrick, 2007). Identifying effective language use, however, is complex as a boundary has to be established. Effective communication for those with a lower level of linguistic competence, such as A2 for instance, is not the same as it is for those with a higher level of competence. Most of the language produced by our participants was actually classified at a B1 level and, for this reason, the boundary for the level of language production considered to be particularly effective was set at B1+ in our study. Our benchmark for effective lexical production was both NS and non-native speaker (NNS) production, and three sources were used as reference corpora to establish this: the British National Corpus (BNC)(BNC Consortium, 2007), the Corpus of Contemporary American English (COCA)(Davies, 2020), and the English Vocabulary Profile (EVP)[4]. The latter is an interface which draws on a range of NNS texts, both written and spoken, and was analyzed by means of search options provided by Text Inspector[5]. This is an interface which enables searches of both NS and NNS data, which we used to determine the CEFR productive levels of participants, compared against frequencies of NS usage. The choice to have three reference corpora was determined by the aim of referencing our local data not only with NS but also with NNS norms. The specific methodology applied to the analysis of corpus data will be described in Section 4.

The findings from the corpus analysis were then supplemented by means of the thematic, qualitative analysis of interviews, conducted with the participants for the purposes of triangulation[6], in a constructivist approach where the researcher drew on what Timmis (2008) refers to as "professionally informed intuition" (p. 7) to interpret the results. The study, therefore, involved three main stages: firstly the testing of the participants and the text production stage, followed by the compilation and analysis of the corpus itself, and finally the development of the course learning design based on the results from the corpus analysis together with the findings from the interviews.

3.2. Participants and texts

The participant sample, which, being a pilot study, was very small, consisted of 21 learners, (16 females, 5 males) whose level of English was C2 according to the Common European Framework Guidelines (Council of Europe, 2001). They were recruited by means of convenience sampling as they were all post graduate students attending English language MA courses at the University of Verona, and their level was tested at the beginning of the study. A univariate analysis of test scores was carried out before potential participants were admitted to the study. The participants took the university language centre, English C2 level test, which focuses on productive, academic writing and presentation skills at this level[7]. This testing was conducted to ensure consistency in the levels of the participants. All 21 participants scored overall marks, which fell between 60 and 95 percent, and were therefore within an acceptable C2 range. The mean score was 78.713 and there was a standard deviation of 11.36, meaning that within the range there was some variance. This, however, was to be expected and was useful for the study as it facilitated comparison between higher- and lower-level performances.

[4] This publication has made use of the English Vocabulary Profile. This resource is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile program, which aims to provide evidence about language use that helps to produce better language teaching materials. See http://www.englishprofile.org for more information. The English Vocabulary Profile has been compiled from both the Cambridge Learner Corpus (CLC) (Lexical Computing Limited on behalf of Cambridge University Press and Cambridge English Language Assessment, 2017) and the Cambridge English Corpus (see https://www.cambridge.es/en/about-us/cambridge-english-corpus). Additional sources for the C levels research have included reference lists relevant to academic English and frequency data on idioms. The EVP is a work in progress whose aim is descriptive rather than prescriptive.

[5] Text Inspector is available at: https://textinspector.com/ (last accessed April 9, 2023).

[6] The analysis of the interviews is beyond the scope of this paper.

[7] In this language test, although the focus is on the productive skills of writing and speaking, test-takers are provided with input which may be audio, audiovisual or text based, so it involves integrated skills.

In the preparatory stage of the study, the participants were shown a short video documentary[8] and asked to write a short 200-word summary. They were then shown the documentary again and asked to write a longer 800-word discursive discussion of one main idea that they found interesting, and they were asked to adopt a neutral, academic style. These texts formed the content of the LC22 corpus and were grouped into two sub-corpora: summaries and discussions. Written consent was provided by all the participants, enabling their texts to be used anonymously both in the study and the resulting dissemination of the research findings.

4. Methodology: compiling and annotating the corpus

4.1. The LC22 corpus: three stages

At the end of the preparatory stage described in Section 3.2, 42 texts in total had been generated. The LC22 corpus was then compiled by subdividing the learner texts into the two main sub-corpora. As this was a pilot study, developed mainly to inform local learning design, the resulting corpus was small, as discussed in Section 3.1., counting 19,193 tokens (17,121 words). Only 2,219 of the words, however, were discrete items due to repeated instances of many of the words, since all the participants discussed the same subject matter from the documentary.

Compilation of the corpus involved three main stages, and actually resulted in three different corpora. The first corpus was compiled in plain text and analyzed with the aid of Text Inspector to ascertain the levels of lexical sophistication (Jimènenz Catalán & Fernandez Fontecha, 2019; Kyle & Crossley, 2015)[9], using the EVP, BNC and COCA as cross-referencing benchmarks. The texts were analyzed for a range of features such as readability, lexical diversity or metadiscourse markers. Our primary interest was in the lexical diversity and sophistication features, which enabled us to determine the overall scores and corresponding CEFR levels of the lexis in the texts. This, in turn, enabled us to ensure that the levels of the texts reflected the participants' levels before undertaking the study itself.

The second stage was to create a 'raw' corpus in the Sketch Engine (Kilgarriff et al., 2014): the text entered was plain text and this was then annotated with automatic part of speech (POS) tagging. This was to enable basic searches such as corpus size, accurate frequency of occurrences, wordlists and to conduct searches for keyness. This made a descriptive analysis of discrete tokens, such as specific concordances, possible but did not allow for structural exploration, which was of primary interest for our study of learner language production, levels and specific strengths or weaknesses. Consequently, a third corpus was then compiled, which was annotated manually in XML and uploaded to the Sketch Engine, where it was automatically tagged for POS. This XML annotated corpus provided us with the opportunity to conduct searches for specific patterns using structural corpus query language (CQL) exploration, so that searches could be made for specific patterns such as effective verb/noun collocation or problematic word choice. All of these had been identified by qualitative analysis during the annotation phase, which is worth describing in more detail.

4.2. XML annotation

Coding the data in XML and then uploading the corpus into the Sketch Engine enabled us to provide overarching metadata such as the sub-corpus, the participants' L1, the identity code number of the texts, which were also related to the participants so that T8D1, for instance, was the discussion text provided by participant number eight. Metadata regarding the participant's L1, gender, mark on the initial pre-test and the year was also included, since this corpus will be expanded to cover other years as well as 2022. This can be seen in Figure 1. The different metadata enable a range of different analysis options such as searching, for instance, for specific participants, gender or level ranges.

[8] The Documentary was the BBC2 broadcast "History of Now: the Story of the Noughties" https://www.bbc.co.uk/programmes/b00pyn3l (last accessed April 10, 2023), and provided on DVD for Unit 6 of *Speakout Advanced* (Clare & Wilson, 2012).

[9] Lexical sophistication often refers to the frequency of unusual words in a text.

```
<?xml version="1.0" encoding="UTF-8"?>
<doc title="T8DI" subclass="Discussion" lang1="Italian" gender="M" mark="95" year="2022">
<text>
    <body>
        There were two themes
        <cat1 macro="LexiS" tag="WC1" Lev1="WPrep1"> about </cat1> this talk,
        <cat1 macro="Misc" tag="PuncT" Lev1="PuWR"> : </cat1>
        <cat1 macro="LexiS" tag="Coll2" Lev1="LP2"> which have given me particular food for thought </cat1>. The first one is the
        <cat1 macro="LexiS" tag="WC2" Lev1="WN2"> pursuit </cat1>
        <cat1 macro="LexiS" tag="Coll2" Lev1="Nprep2"> pursuit of </cat1>
        <cat1 macro="LexiS" tag="Coll2" Lev1="AdjN2">eternal youth </cat1>, that characterizes our society nowadays.
        <cat1 macro="LexiS" tag="Coll2" Lev1="LP2"> What is evident, I think is the fact that </cat1>  youth and the dream of it
        <cat1 macro="LexiS" tag="Coll2" Lev1="VN2"> drive our daily life  </cat1>
        <cat1 macro="Gram" tag="Noun" Lev1="PluOm"> life </cat1>.
```

Figure 1. An extract from the XML annotated corpus

The data was coded into two macro language-production categories: general language production and infelicities. The examples in Figure 1 all belong to the general language production category. These examples were then tagged for problematic or effective language choice. Choices that were tagged as problematic were assigned the number "1" as part of the tag, such as "'tag=WC'", where the 1 means there was an error. In Figure 1, for instance, 'tag="WC1" in the first row, is an example of word choice: in this case the choice made was "about", and it is problematic from a semantic viewpoint, in the context:

"There were two themes **about*** this talk,"
which would be better expressed as
"There were two themes **in** this talk,"

When the tag includes the number 2, such as 'tag="WC2"' in the fourth row, it is an instance of effective language production. In this case, the choice of the noun "pursuit", which is a non-frequent lexical item in both the BNC and COCA and was classified as C2 level in EVP, is used appropriately in the context of the phrase "the pursuit of eternal youth". The third tag, which is 'Lev1', refers to the specific language item being analyzed, so that 'Lev1="AdjN2" in row six refers to a choice of the effective adjective/noun collocation "eternal youth".
The annotation was informed by the system methodology outlined in the Louvain Error Tagging Manual, Version 2.0 (Granger et al., 2022), although, as previously mentioned, this was adapted for our local needs, as the Louvain methodology does not deal with effective language production. One issue that was encountered was the need for the multiple tagging of certain items, because of their language functions in the discourse. "Pursuit", for instance, was tagged as an effective word choice of the noun itself (row four) but then was also tagged as an effective noun/preposition collocation (row five). This enabled us to make precise structural exploration searches, but it also determined the choice of keeping the second corpus, which was POS tagged but not annotated manually, to provide an accurate overview of corpus size and frequency of specific occurrences.

5. Overall findings and discussion

The initial, overall findings confirmed our original hypothesis that problematic areas to focus on in academic English writing courses are related not only to grammar but primarily to lexis, with the caveat that this refers to advanced levels, like the level of the participants in this study. Figure 2 shows that 24% of the effective language production was related overall to effective collocations of various different types. 10% referred to effective word choice. The most problematic area was word choice, with 18% of problematic language production in this category, which refers to the choice of the wrong word because of its meaning. The fifth largest category, which was also problematic, was collocation. What was interesting to see here, was, that although we had expected to see problematic uses with collocation, there was actually considerable evidence of effective usage. There was also both problematic and effective word choice, which tends to suggest that a learning design should highlight the effective usage and also provide learners with the tools to enhance this. A major category of problematic usage was verbs, which is generally related to problems either of morphological form or of inappropriate tense choices.

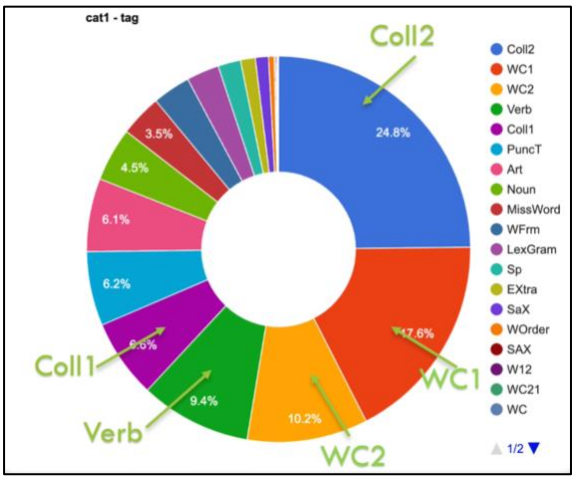


Figure 2. An overview of the general findings

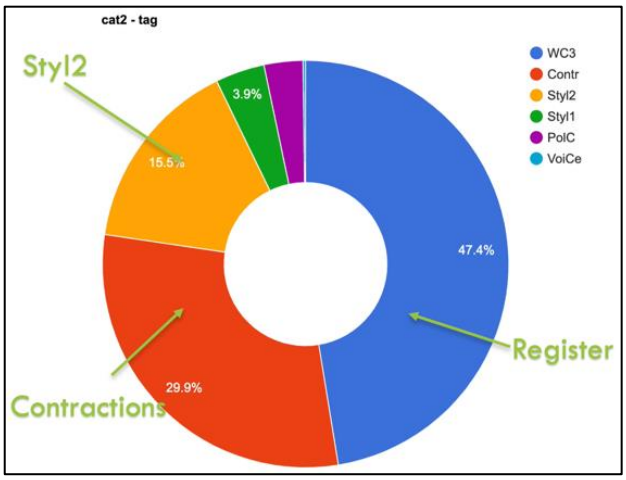


Figure 3. An overview of the infelicities

The second macro category was related to infelicities, and Figure 3 shows an overview of this area. The participants were asked to write in a neutral, academic style, but it was clear from the analysis of their production that they were not able to do this. This was suggested by instances of formal language, often related to conjunctions that were present in texts, where informal choices such as contractions or informal lexical choices were also common. The informal elements were mainly related to lexical choice, such as "little", which, when referring to size, tends to be informal, or discourse features such as the choice of "really" as a pre-modifier. Examples of these choices can be seen in Table 1.

Table 1
Register choices

Over Informal	Over formal
I'm not speaking about…lots of…incredible…really…little…nice	Furthermore…contrary to…thus…in the following… consequently

When applying the lens of level to the exploration, what is revealed is that the main difference between lower and higher-level production is related to the use of collocations, as can be seen in Figures 4 and 5. 59% of effective language production was related to collocations in higher-level production compared with 40% in the lower-level range. Effective and less effective word choice was balanced at higher levels, whereas at the lower ones 12% of the production was problematic from the viewpoint of word choice, and only 4% was classified as being effective (higher than a B1 level). Ineffective collocation accounted for only 5% at higher levels, whereas this was slightly higher, at 8%, in lower-level production. The discussion, in this paper, therefore, will mainly focus on the results of the collocation analysis, even though the areas of lexical choice from the semantic viewpoint would also be worthy of investigation.

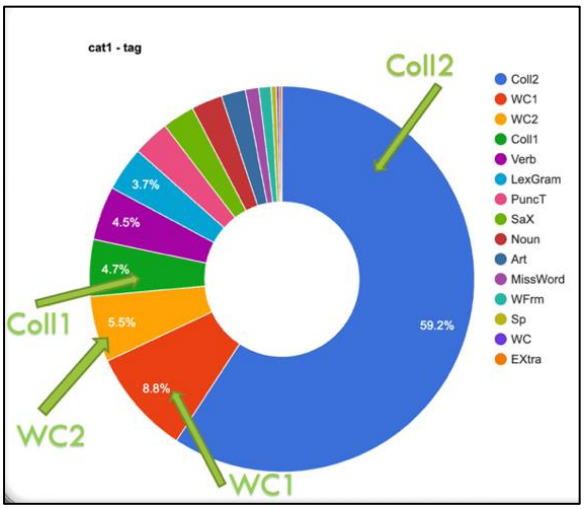


Figure 4. An overview of higher level (92-95% marks)

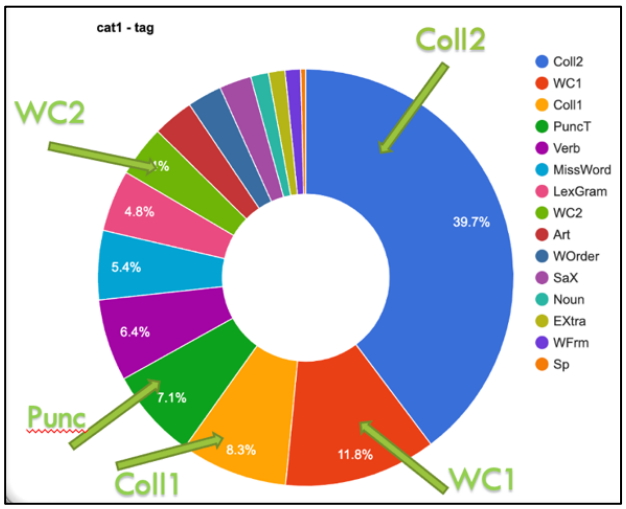


Figure 5. An overview of lower level (60-63% marks)

5.1. Collocation and lexical phrases

Effective collocation, in our corpus, included a range of specific collocation types but the most common patterns tended to be verb/noun and adjective/noun collocations. This category also included lexical phrases or frames (Benson, 1994; Koprowski, 2005; Lewis, 1993; Nattinger & DeCarrico, 1992), that is items with a fixed lexical element that frames an idea. In the first example in Table 2, for instance, the frame is "No matter how good or bad… was". The lexical phrase here: "No matter how good or bad" may remain fixed and then used to frame various ideas. In this case it was "their youth", but it could be a range of items such as "their education, their relationships, their salary", to name just a few. It may be useful for learners to focus on fixed items, such as these lexical frames, to develop the quality of their written production. A more detailed overview of the data retrieved from the two levels of analysis introduced in the previous section, can be seen in Appendix but illustrative instances will be discussed here. Examples are taken from the higher and lower C2 levels in order to show differences which may appear within the range of a single level. The examples in Table 2 show that there seems to be more experimentation with language that is part of the personal repertoire of the learner, such as the use of the frame "No matter how good or bad… was" or the verb/noun collocation "massive change", at higher levels. In contrast, at lower levels, the effective frames tended to be items that may have been presented explicitly in class, such as "What interested me… was", or had been taken directly from the video documentary that the participants had viewed in the preparatory stage. One example of this is "Now the different generations are less aware of each other." The choice of less frequent lexis, possibly indicating a wider lexical repertoire, is also evident at higher levels in the choice of items that are collocated, such as "huge impact", "fixed rendezvous" or "radical changes", in the adjective/noun collocations, whereas lower-level production often relies on the choice of higher frequency items such as "everyday lives", "cultural phenomenon", "parallel world". As was noted earlier, although items such as both "cultural" and "phenomenon" may be classified as low frequency in NS corpora, in our context, where the majority of the participants were Italian L1 speakers, these are Latin-derived items that are easily accessible to them. "Phenomenon" was classified as C1 in the EVP, but is easily comprehensible for our participants.

Table 2
Examples of common effective (tagged as 2) and problematic (tagged as 1) lexical phrase or collocation choices grouped according to level

Language Choice	Higher Level	Lower Level
Lexical Phrase (LP2)	No matter how good or bad their youth was... They tend to live the lifestyle of a teenager	What interested me in particular about this video was... Now the different generations are less aware of each other.
Verb Noun (VN2)	(This generation has) seen a massive change represent a full commitment emulate youngsters	...disorientating our society deal with dangers attach too much importance
Adjective Noun (AdjN2)	huge impact fixed rendezvous radical changes	everyday lives cultural phenomenon parallel world
Verb Noun (VN1)	gains a lot of profit have possibilities	live the present make some experiences
Noun Preposition (NPrep1)	desire of break the rules This tendency to a form of 'perpetual childhood'	The reason of many social changes The same possibilities of young people
Adjective Noun (AdjN1)	a stylish, young garment A consistent amount (money)	deep changes ambiental problems

5.2. Problematic language choices

Greater experimentation is evident at higher levels, although it may lead to problematic choices such as "a stylish young garment", where the presence of "stylish, young" tends to point to a collocation with a person or animate entity, and not "garment". This experimentation may be interpreted as a positive aspect of the learning process, but providing tools for learners to check the appropriacy of their choices may lead to more effective communication, which is required when writing academically at higher levels.

A similar problem can be seen in the choice of a "consistent amount of money", where "consistent" does not usually collocate with "amount" and perhaps "considerable" would have been a better choice. At the lower levels, collocation choices, such as "deep changes" or "ambiental problems", point rather to a more basic transfer from the L1 to the L2, and the morphologically creative adaptation of L1 terms such as "ambientale" to create something the writer considers to sound English (ambiental*), when the learner lexicon is not developed enough for them to be able to express themselves effectively. These initial findings tend to underline the fact that in order to foster effective academic writing habits our learners need to be made aware of their strengths and weaknesses and then to be provided with the tools that will enable them to develop their own lexical repertoires.

6. Focus on the learning design

As a result of these initial findings, in our academic writing courses it was decided to focus particularly on the development of lexis and mainly collocation. A second objective was to develop lifelong learning skills and explore the use of available resources as an aid to lexical choice in writing. This, it was hoped, would foster greater independence and agency (Ahearn, 2001) in our learners. The learning design involves a two-step approach. The first stage is a focus on familiarization and reflection, which aims to introduce learners to the concept of collocation and to reflect on both effective and less effective usage. This is important because a lack of knowledge about collocation itself and its importance in moulding effective communication may hamper learners at the outset. This stage involves guided-discovery (Bruner, 1961) strategies, where learners analyze

and reflect on peer writing, by exploring an annotated, learner-friendly corpus, created in Markin[10]. They are guided inductively, by means of the pedagogically informed (Timmis, 2008) annotation, provided by the teacher but generated from learner production, to notice the strengths and the weaknesses in the discourse produced. The second step is one of experimentation and is where DDL is applied in a more personalized way. Learners are familiarized with a range of freemium or free corpus interfaces, such as SkeLL, which will be considered in more depth in this paper, although they were also introduced to other interfaces such as the English Corpora[11] or Just the Word[12]. This step involves analysis and reflection as well but also experimentation as learners personalize the language they wish to explore: a key factor if investment in the language being learned (Norton, 2009) is to be encouraged.

6.1. The first step: teacher compiled corpora on Markin

This step involves the compilation of another corpus, which is pedagogical in nature. It is compiled by teachers from learner-generated writing. The software used is Martin Holmes' Markin (Holmes, n.d.), which is a desktop programme, that is not new, but is easy to use for both teachers and learners, and has proved popular over the years in our context. Figure 6 shows the interface with default coding options on the left (red for problematic, and green for effective language choice). These default options are fairly generic, however, and a button that just says 'good', for instance, may not be helpful enough for learners to see why a particular choice is effective. The buttons, in fact, can be personalized by teachers according to need and new ones can be added and saved for future use, and comments can also be added for greater granularity. The corpus data itself can simply be entered, by typing or pasting, into the white central space and the programme will automatically convert the final document into html, so no coding is required.

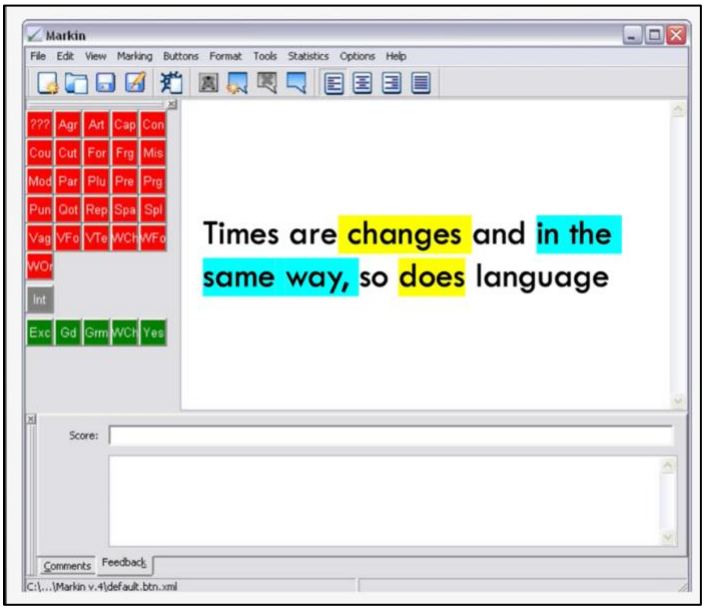


Figure 6. The Markin interface with default coding options

Figure 7 shows an example of a text, which was not part of the LC22 corpus, but produced in class[13]. As can be seen, the texts are anonymized and only certain features of the text have been coded, in the interests

[10] This software was developed by Martin Holmes and was used as a desktop programme for the development of pedagogically friendly corpora. http://www.cict.co.uk/markin/index.php (last accessed April 10, 2023).
[11] This interface enables access to a range of corpora such as British, American and Internet English, as well as the Global Web-based English corpus, which provides data from 20 different countries. The interface is available at https://www.english-corpora.org/coca/ (last accessed April 12, 2023).
[12] Just the Word draws on the BNC but provides clearly organized collocational output that is learner friendly. It is available at http://www.just-the-word.com/ (last accessed April 12, 2023).
[13] Consent was given for publication.

of guiding learner analysis. The red labels indicate problematic areas whereas the green ones are effective features. The numbers refer to notes and, as can be seen in the example, this enables teachers to provide much more precise discussions as to why a choice may not be appropriate. In this case, the choice of "aspects" would collocate better with the idea of language change, and the teacher has explained the semantic prosody of intentionality, more commonly found with "modification".

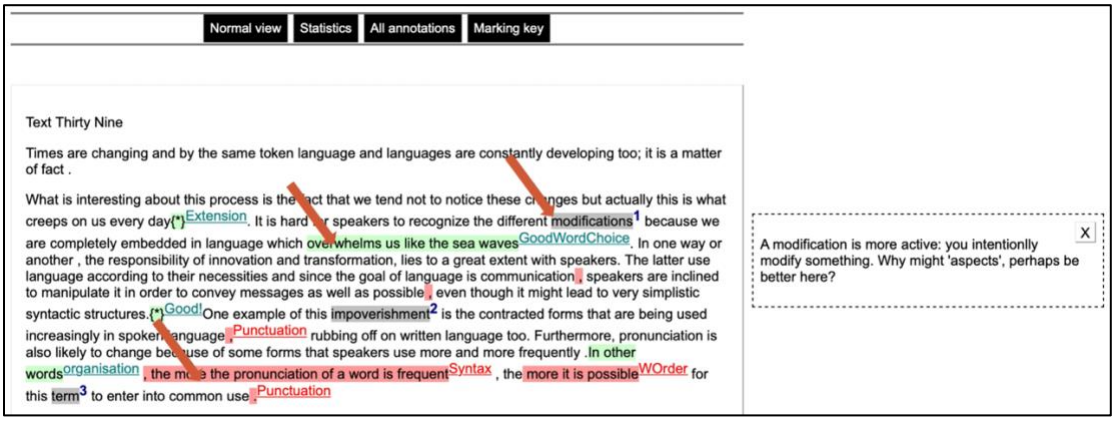


Figure 7. A text in Markin after it has been coded for use in class

6.1.1. Classroom Procedure

The procedure used in the first step involved five stages:

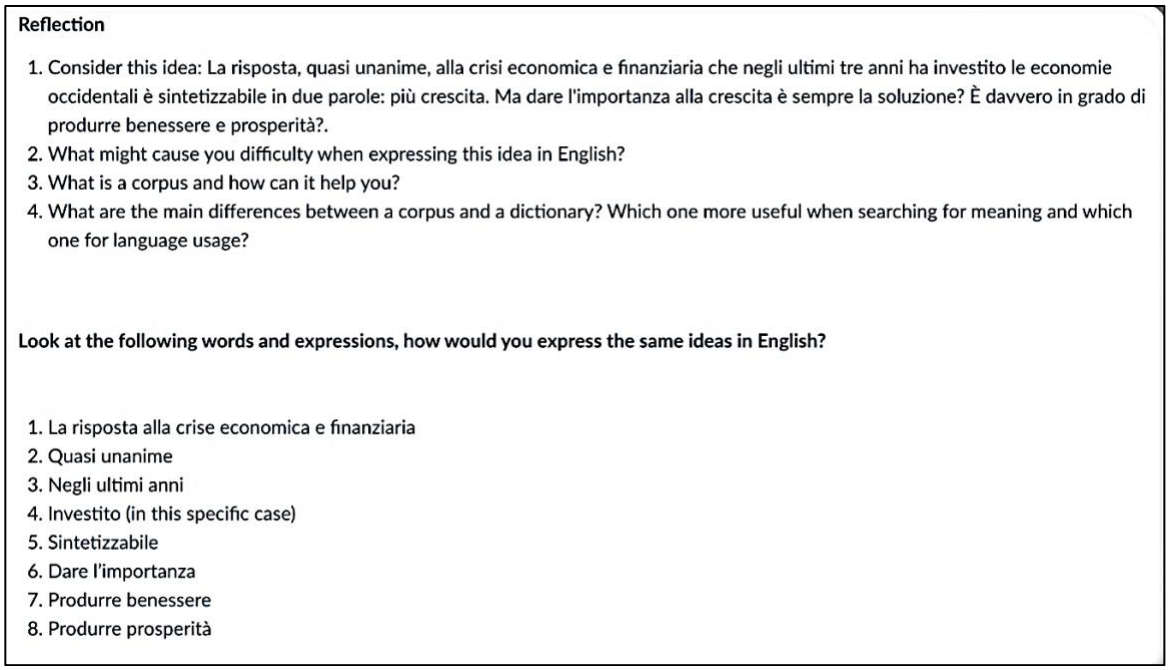
1) Teacher prepares learner texts and determines aspects to highlight for guided discovery;
2) Teachers develop the corpus, which is coded and annotated;
3) Learners analyze the texts either in or outside class. This stage may involve further structuring from the teacher, such as asking the learners to focus on three aspects that they find interesting or useful. These may then be further discussed with peers and the whole class;
4) Learners analyze their own texts and correct or edit them. This may be done individually or together with their peers in pair or group work;
5) Class discussion is held on the most effective or problematic aspects of their writing;
6) This work is extended by setting further writing tasks or discussions such as explorations of emergent problematic language choices or reflection on concepts such as collocation itself, which may require clarification.

This procedure enables learners to invest, in Norton's terms (2010, 2013), in their own agency and their personal learning process by working on their own or their peers' writing. They also exercise their agency by then deciding how they would like to edit their personal work as a result of their analysis, or which language areas are key for them to focus on. The issue of developing learners' lexical repertoires can then be addressed by introducing them to easily accessible online corpus interfaces, such as SkeLL. In order for learners, however, to be able to develop their own repertoires, they must firstly be familiarized with the use of corpus interfaces.

6.2. The second step: deconstruction, reflection and the exploration of online corpus interfaces and developing personalized lexical repertoires

The initial procedure adopted in the second step is also one of reflection and guided discovery: learners are asked to look at a very short text in their L1 and to identify possibly problematic features. As is shown in Figure 8, the text is firstly presented and learners are encouraged to reflect and discuss potential problems in general terms. The first two questions focus specifically on language awareness items but the third and fourth ask them to consider the nature of corpora and how they differ from dictionaries, as this, in the past, has proved to be a valuable reflection. For many learners, indeed, this may be their first encounter with corpora, and it is important that they realise that corpora interfaces are particularly useful when searching for issues of language usage. In the second activity, the text, which has previously been deconstructed into language chunks, can be

analyzed. This approach aims to sensitize learners firstly to the collocation issues that may emerge. "Dare l'importanza", for instance, cannot simply be transferred into "give importance" in English, and knowledge of collocation may help learners to realize this. Learners may not know the answers to these questions, but the aim is to arouse curiosity and whet learner appetites.



Reflection

1. Consider this idea: La risposta, quasi unanime, alla crisi economica e finanziaria che negli ultimi tre anni ha investito le economie occidentali è sintetizzabile in due parole: più crescita. Ma dare l'importanza alla crescita è sempre la soluzione? È davvero in grado di produrre benessere e prosperità?.
2. What might cause you difficulty when expressing this idea in English?
3. What is a corpus and how can it help you?
4. What are the main differences between a corpus and a dictionary? Which one more useful when searching for meaning and which one for language usage?

Look at the following words and expressions, how would you express the same ideas in English?

1. La risposta alla crise economica e finanziaria
2. Quasi unanime
3. Negli ultimi anni
4. Investito (in this specific case)
5. Sintetizzabile
6. Dare l'importanza
7. Produrre benessere
8. Produrre prosperità

Figure 8. Initial guided reflection in step two

Following these reflections, the next stage is to provide learners with the reference tools, that will help them to answer questions about language usage. Figure 9 shows the next stage on this journey, which is one of applying corpus search skills to answer the questions arising from the analysis of the language chunks (see Figure 8). As can be seen, the students are also led to reflect on the type of questions that can be answered by using dictionaries, which often involves searches for meanings rather than answers to the lexico-grammatical problems that can better be answered by the corpus interface.[14] The questions provide a guided procedure which helps learners to familiarize themselves with the range of searches that can be conducted using different resources.

[14] Dictionaries, which include traditional ones such as the Longman dictionary online https://www.ldoceonline.com/ but also more contemporary interfaces such as Reverso Context https://context.reverso.net/translation/ , are also explored as learners use these sites and can be helped to develop a constructive but critical approach to such use. Both resources were last accessed on April 12, 2023.

Dictionaries are very useful for the meaning of words and phrases and for examples but when it comes to usage and collocations there is a better alternative: corpora.

Welcome to SkeLL

Go to Intro to SkeLL

1. Look up '**crisis**' in the **word sketch** to check how this is usually collocated?
2. Which verbs can you use when 'crisis' is a direct object?
3. Look up "years" (**word sketch**) to find out which adjectives can be used here?
4. Which expression is probably the most appropriate for the text above?
5. Look up 'ha investito le economie occidentali' in Reverso Context Be careful because you need to find the meaning that you are interested in.
6. Look up 'plunge' in SkeLL (**word sketch**) to find out how to use it? Try looking at the examples with 'Europe' as a direct object.
7. Look up 'sintetizzabile in due parole' in Reverso Context but look to see which expressions are most common.
8. Look up 'importanza' in Skell (word sketch) Which verbs commonly collocate with importance as a direct object?
9. Look up 'benessere' in Reverso Context then check the meanings of "well-being" and 'welfare' in the Longman Dictionary Online . Which one would be better in the above text?
10. Now look your choice up in SkeLL (**word sketch**) to see which verbs commonly collocate with it?
11. Repeat numbers 9 and 10 for 'prosperita'.

Figure 9. Guided introduction to searching SkeLL

Learners are shown the SkeLL interface and how to access the word sketch feature, initially using the word "crisis", for instance. The resulting output, which is shown in Figure 10, is very clearly organized according to specific POS collocations for exploration, which is appropriate for use by learners.

crisis noun ✓ Show context

verbs with crisis as subject		verbs with crisis as object		adjectives with crisis		modifiers of crisis		nouns modified by crisis	
deepen	crisis deepened	resolve	resolve the crisis	acute	crisis is acute	financial	the financial crisis	hotline	crisis hotline
erupt	crisis erupted	precipitate	precipitated a crisis	unscathed	crisis unscathed .	economic	the economic crisis	management	crisis management .
affect	affected by the crisis	solve	solve the crisis	avoidable		debt	debt crisis	situation	a crisis situation
unfold	crisis unfolded	avert	crisis was averted	imminent	crisis is imminent .	global	the global financial crisis	intervention	crisis intervention ,
hit	crisis hit	deepen	deepening crisis	severe	crisis was severe	humanitarian	humanitarian crisis	prevention	crisis prevention and
loom	crisis loomed	loom	looming crisis	impending		hostage	hostage crisis	pregnancy	crisis pregnancies
escalate	crisis escalated	face	facing a crisis	profound		banking	the banking crisis	counseling	crisis counseling ,
arise	crisis arose	defuse	to defuse the crisis	momentous		identity	an identity crisis	counselor	crisis counselor
worsen	crisis worsened	trigger	crisis triggered	grave		oil	the 1973 oil crisis	center	crisis center
grip	crisis gripping the	escalate	escalating crisis	cyclical		Ukraine	the Ukraine crisis	Core	Crisis Core : Final Fantasy
impact	crisis has impacted	worsen	worsening crisis	unprecedented		mortgage	the subprime mortgage crisis	communication	crisis communications
abate	the crisis abated	provoke	crisis provoked	foreseeable		subprime	the subprime mortgage crisis	hit	crisis hit ,
rock	the crisis rocking the	overcome	overcome the crisis	past	the crisis was past	housing	the housing crisis	aftermath	Crisis Aftermath : The Battle
threaten	crisis threatened	address	address the crisis	unresolved		Asian	the Asian financial crisis	proportion	crisis proportions .
precipitate	crisis precipitated by	tackle	tackle the crisis	hypocritical		energy	the energy crisis	mode	in crisis mode

words and		or crisis	
recession	crisis and recession	recession	recession and financial crisis
disaster	crisis or disaster	saving	the savings and loan crisis
conflict	crises and conflicts	disaster	disasters and crises
emergency	crisis or emergency	conflict	conflict and crisis
war	crisis and war	Savings	the Savings and Loan crisis ,
collapse	crisis , the collapse	emergency	emergency or crisis
downturn	crisis and economic downturn	bubble	bubble and financial crisis
bank	crisis , banks	scandal	scandal and crisis

Figure 10. SkeLL output for the 'crisis' word search

Learners then consider the different verb/noun collocations and they then choose one option to explore further. In one group, on our course, for instance, ”precipitate a crisis” was chosen and learners then analyzed the output in context for that choice (Figure 11).

Figure 11. SkeLL output for ‘precipitate a crisis’ in context

The SkeLL algorithm searches several corpora, uploaded to the Sketch Engine, for up to forty different occurrences of this item, which are all different and learners can then study this output for register, collocation but also for lexical phrases or frames. Since the findings from our study have shown lexical frames to be a key element in what was perceived to be effective language the next step was introduced specifically to foster learner awareness of and competence in the use of frames. This is an experimental phase, where learners are asked to experiment with their own examples.

6.2.1. Experimentation Procedure: working with lexical frames

The procedure used was the following:

1) Learners choose an example from the SkeLL output in context (See Figure 11);
2) The example is analyzed. If the choice is “This event precipitates a family crisis”, for instance, the resulting frame may look like this:
   “This _____ precipitated a _____ crisis”;
3) Learners then generate their own personalized examples such as “The speech precipitated a political crisis”;
4) Examples are discussed in class;
5) Learners then explore other collocations that they may require in their own writing and generate language frames such as “This election precipitated a national crisis”;
6) This work may be extended by learners sharing their findings or testing each other.

7. Conclusion

In conclusion, the learning design of our ESAP course was developed as a direct result of the findings from the pilot study. The study, although limited to our local context, provided us with valuable insights into learner needs and confirmed our initial hypothesis as to the problematic issue of collocation. What also emerged, however, was learners' effective use of collocation, which included lexical frames. Our research informed the DDL pedagogical framework which we developed for our academic English writing courses. The learning design combined elements of awareness raising, reflection and analysis. Our course work did not focus only on the remedial area of error correction, although this was a part of it, but also on the recognition of effective language choices. Our design integrates the skills of analysis, fostered by means of guided discovery, and reflection to enhance greater understanding of the contributing factors when writing academically, and experimentation with the language learners themselves choosing which collocations they need to focus on in their own writing. This enables learners to invest in their own learning process, and such investment may lead to increased lexico-grammatical competence in academic writing skills. When introduced at an early stage, learners such as postgraduate students, at the beginning of their academic writing careers, have time to build their lexical repertoires before possibly becoming professional writers in later life. Interfaces such as SkeLL are both accessible and user-friendly for learners. They may be considered to be valuable tools, which provide language models, in the output generated, that can be built on by learners for their personalized production and are veritable lifelong learning resources.

References

Ahearn, Laura M. (2001). Language and agency. *Annual Review of Anthropology*, *30*, 109–137. https://doi.org/10.1146/annurev.anthro.30.1.109

Bahns, Jens, & Eldaw, Moira (1993). Should we teach EFL students collocation? *System, 21*(1), 101–114. https://doi.org/10.1016/0346-251x(93)90010-e

Barnau, Anna, & Ferková, Nina (2022). Medical English for academic purposes: the impact of a new guide for postgraduates on developing specific communication skills. *ICERI2022 Proceedings, 1,* 1262–1267. https://doi.org/10.21125/ICERI.2022.0329

Basturkmen, Helen, & Wette, Rosemary (2016). English for academic purposes, In Graham Hall (Ed.), *The Routledge Handbook of English Language Teaching* (pp. 164–176). Routledge. https://doi.org/10.4324/9781315676203-15

Benson, Morton (1994). Lexical phrases and language teaching. In *System 2*, 406–408. https://doi.org/10.1016/0346-251X(94)90027-2

Benson, Phil (2007). State of the art article: Autonomy in language teaching and learning. *Language Teaching, 40*(1), 21–40. https://doi.org/10.1017/s0261444806003958

Biber, Douglas, Conrad, Susan, & Cortes, Viviana (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

BNC Consortium (2007). The British National Corpus, XML Edition. Retrieved from http://hdl.handle.net/20.500.12024/2554

Bondi, Marina (2001). Small corpora and language variation Reflexivity across genres. In Mohsen Ghadessy, Alex Henry, & Robert L. Roseberry (Eds.), *Small Corpus Studies and ELT. Theory and Practice* (pp. 135–174). John Benjamins Publishing Company. https://doi.org/10.1075/scl.5.11bon

Boulton, Alex (2017). Data-driven learning and language pedagogy. In Steven Thorne & Stephen May (Eds.), *Language, Education and Technology: Encyclopedia of Language and Education* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-319-02237-6_15

Bourdieu, Pierre (1984). *Distinction: A Social Critique of the Judgment of Taste*. (Richard Nice, Trans.). Routledge & Kegan Paul. (Original work published 1979). https://doi.org/10.1086/446595

Bruner, Jerome Seymour (1961). The act of discovery. *Harvard Educational Review, 31*, 21–32. https://doi.org/10.4324/9780203088609-13

Chen, Yu-Hua & Baker, Paul (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14(*2), 30–49. https://doi.org/ 10125/44213

Clarke, Victoria & Braun, Virginia (2014). Thematic analysis. In Michalos, Alex C., *Encyclopedia of quality of life and well-being research* (p. 283). Springer. https://doi.org/10.1007/978-94-007-0753-5_3470

Conzett, Jane. (2001). Integrating collocation into a reading and writing course. In Michael Lewis (Ed.), *Teaching Collocation* (2nd ed., pp. 70–86). LTP.

Council of Europe (2001*). Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge University Press. http://www.coe.int/lang-CEFR

Dagneaux, Estelle Denness, Sharon & Granger, Sylviane (1998). Computer-aided error analysis. *System, 26*(2), 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

Davies, Mark (2020). The Corpus of Contemporary American English (COCA). Retrieved from https://www.english-corpora.org/coca/.

Dörnyei, Zoltán (2007). *Research methods in applied linguistics.* Oxford University Press. https://doi.org/10.1017/s0272263110000094

Duff, Patricia & Byrnes, Heidi (2019). SLA Across Disciplinary Borders: Introduction to the Special Issue. *Modern Language Journal, 103*, 3–5. https://doi.org/10.1111/modl.12537

Durrant, Philip & Schmitt, Norbert (2009). To what extent do native and non-native writers make use of collocations?. *IRAL – International Review of Applied Linguistics in Language Teaching, 47*(2), 157–177. https://doi.org/10.1515/iral.2009.007

Ennis, Michael, & Mikel Petrie, Gina (2020). A response to disparate/desperate circumstances. In Michael Ennis & Gina Mikel Petrie (Eds.), *Teaching English for tourism: bridging research and praxis* (pp. 1–6). Routledge. https://doi.org/10.4324/9780429032141-101

Firth, John R. (1957). A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (Special Volume of the Philological Society)*, 1952–59, 1–32. https://doi.org/10.1093/ref:odnb/33138

Flowerdew, John, & Peacock, Matthew (2001). Issues in EAP: A preliminary perspective. In John Flowerdew & Matthew Peacock (Eds.), *Flowerdew and Peacock Research Perspectives on English for Academic Purposes* (pp. 8–24). Cambridge University Press. https://doi.org/10.1017/cbo9781139524766.004

Flowerdew, Lynne (2015). Corpus-based research and pedagogy in EAP: From lexis to genre. *Language Teaching, 48*(1). https://doi.org/10.1017/S0261444813000037

Graddol, David (2006). *English Next.* British Council Research. British Council. https://www.teachingenglish.org.uk/sites/teacheng/files/pub_english_next.pdf

Granger, Sylviane, & Bestgen, Yves (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching, 52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Granger, Sylviane, Swallow, Helen, & Thiewissen, Jennifer (2022). *The Louvain Error Tagging Manual Version 2.0*. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain. https://www.researchgate.net/publication/360806408_The_Louvain_Error_Tagging_Manual_Version_20

Green, Jenny, Burrow, Maria, & Carvalho, Lucila (2020). Designing for transition: supporting teachers and students cope with emergency remote education. *Postdigital Science and Education, 2*(3), 906–922. https://doi.org/10.1007/s42438-020-00185-6

Halliday, Michael A. K. (1994). *An introduction to functional grammar (2nd ed.)*. Arnold, Hodder Headline.

Hartle, Sharon (2020). A dove in flight: Agency in 21st century language learning. *Humanising Language Teaching*, 22(3). Retrieved from https://www.hltmag.co.uk/june2020/a-dove-in-flight

Hartle, Sharon & Cavalieri, Silvia (forthcoming). Challenges and opportunities in EAP: Teaching PhD presentation skills. In Brian R. Morrison, Carole MacDiarmid, Ide Haghi, & Anneli Williams (Eds.), *Proceedings of the 2021 BALEAP conference: exploring pedagogical approaches in EAP teaching*. Garnet Publishing.

Hartle, Sharon, Facchinetti, Roberta, & Franceschi, Valeria (2022). Teaching communication strategies for the workplace: A multimodal framework. *Multimodal Communication, 11(*1), 5–15. https://doi.org/10.1515/mc-2021-0005

Holec, Henri (1981). *Autonomy in foreign language learning*. Pergamon.

Holmes, Martin (n.d.). *About Markin.* Retrieved from http://www.cict.co.uk/markin/index.php

Hyland, Ken (2006). *English for academic purposes. An advanced resource book.* Routledge. https://doi.org/10.4324/9780203006603

Hyland, Ken (1998). *Hedging in scientific research articles*. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.54

Hyland, Ken (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, Ken & Tse, Polly (2009). Academic Lexis and Disciplinary Practice: Corpus Evidence for Specificity. *IJES* 9 (2). Retrieved from https://revistas.um.es/ijes/article/view/90781

Jimènenz Catalán, Rosa María, & Fernandez Fontecha, Almudena (2019). Lexical Availability Output in L2 and L3 EFL Learners: Is There a Difference? *English Language Teaching, 12*(2), 77–87. https://doi.org/10.5539/elt.v12n2p77

Johns, Tim (1986). Micro-Concord: A Language Learner's Research Tool. *System, 14*(2), 151–162. https://doi.org/10.1016/0346-251x(86)90004-7

Johns, Tim (1991). Should you be persuaded - two samples of data-driven learning materials. *English Language Research Journal, 4,* 1–16. http://www.lexically.net/wordsmith/corpus_linguistics_links/Tim_Johns_and_DDL.pdf

Kilgarriff, Adam, Baisa, Vit, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychly, Pavel, & Suchomel, Vit (2014). The Sketch Engine. Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kirkpatrick, Andy (2007). *World Englishes*. Cambridge University Press.

Koprowski, Mark (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal, 59*(4), 322–332. https://doi.org/10.1093/elt/cci061

Kyle, Kristopher, & Crossley, Scott (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Larsen–Freeman, Diane (2019). On language learner agency: a complex dynamic systems theory perspective. *Modern Language Journal, 103*, 61–79. https://doi.org/10.1111/modl.12536

Lew, Robert, Frankenberg-Garcia, Ana, Rees, Geraint, Roberts, Jonathan C., Sharma, Nirwan, & Butcher Peter (2018). ColloCaid: A tool to help academic writers find the words they need. In Fanny Meunier, Julie Van de Vyver, Linda Bradley & Sylvie Thouësny (Eds.), *CALL and complexity – short papers from EUROCALL 2019* (pp. 144-150). https://doi.org/10.14705/rpnet.2019.38.1000

Lewis, Michael (1993). The Lexical Approach: The State of ELT and a Way Forward. In *Studies in Second Language Acquisition (Vol. 7)*. Language Teaching Publications.

Little, David (1991). Learner Autonomy 1 : Definitions, Issues and Problems Learner autonomy. In *Research Gate. Authentik*. Retrieved from https://www.researchgate.net/publication/259874253_Learner_Autonomy_1_Definitions_Issues_and_Problems

Littlewood, William (2014). Methodology for teaching ESP. In Vijay Bhatia & Stephen Bremner (Eds.), *The Routledge Handbook of Language and Professional Communication* (pp. 287–303). Routledge. https://doi.org/10.4324/9781315851686.ch19

Macis, Marijana, & Schmitt, Norbert (2017). The figurative and polysemous nature of collocations and their place in ELT. *ELT Journal*, *71*(1), 50–59. https://doi.org/10.1093/elt/ccw044

McEnery, Tony, Brezina, Vaclav, Gablasova, Dana, & Banerjee, Jayanti (2019). Corpus linguistics, learner corpora and SLA. *Annual Review of Applied Linguistics, 39*, 74–92. https://doi.org/10.1017/s0267190519000096

Nation, Paul (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/s0008413100018260

Nation, I.S. Paul, & Waring, Rob (1997). Vocabulary size, text coverage and word lists. In Norbert Schmitt & Michael McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6–19). Cambridge University Press. Retrieved from https://www.lextutor.ca/research/nation_waring_97.html

Nattinger, James R., & DeCarrico, Jeanette S. (1992). Lexical phrases and language teaching. Oxford University Press.

Nesselhauf, Nadja (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223–242. https://doi.org/10.1093/applin/24.2.223

Norton, Bonny (2010). Identity, literacy & English language teaching. *TESL Canada Journal* 28 (1), 1-13. https://doi.org/10.18806/tesl.v28i1.1057

Norton, Bonny (2013). *Identity and language learning: extending the conversation* (2nd Edition). Multilingual Matters. https://doi.org/10.21832/9781783090563

O'Keefe, Anne, McCarthy, Michael, & Carter, Ronald (2007). *From corpus to classroom*. Cambridge University Press. https://doi.org/10.1017/cbo9780511497650

Prodromou, Luke (1996). Correspondence from Luke Prodromou. *ELT Journal, 50*(1), 88–89. https://doi.org/10.1093/elt/50.1.88

Shin, Dongkwang, & Nation, Paul (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339–348. https://doi.org/10.1093/elt/ccm091

Sinclair, John McH. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair John McH. (2004). Introduction In John McH. Sinclair, (Ed.), *How to Use Corpora in Language Teaching* (pp. 1-10). John Benjamins Publishing Company. https://doi.org/10.1075/scl.12.02sin

Tang, Ramona (2012). The Issues and Challenges Facing ESL/EFL Academic Writers in Higher Education Contexts: An Overview. In *Academic writing in a second or foreign language: issues and challenges facing ESL/EFL academic writers in higher education contexts* (pp. 1–20). Continuum Publishing. https://doi.org/10.5040/9781472541543

Timmis, Ivor (2008). The lexical approach is dead: long live the lexical dimension! *Modern English Teacher, 17*(3), 5–10.

Van Lier, Leo (2008). Agency in the classroom. In Lantolf James P. & Poehner, Matthew (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 163–186). Equinox. https://doi.org/10.1558/equinox.29307

Whittle, Clayton, Tiwari, Sonia, Yan, Shulong, & Williams, Jeff (2020). Emergency remote teaching environment: a conceptual framework for responsive online teaching in crises. *Information and Learning Science.* https://doi.org/10.1108/ILS-04-2020-0099

Widdowson, Henry G. (1991). The description and prescription of language. In J. Alatis (Ed.), *Georgetown University Round Table on Languages and Linguistics* (pp. 11–24). Georgetown University Press. https://onlinebooks.library.upenn.edu/webbin/book/lookupid?key=olbp67603

Appendix

Overview of collocation findings grouped according to higher and lower levels

The collocations are reported with some context, for ease of interpretation. Some errors, related to other language areas, such as the wrong word form choice, have been left in the reported examples.

Higher levels: (90-95% marks on pre-test)

Coll2: 85 instances, which equates to 39,516.5 per million tokens

The first 21 occurrences of general effective collocation in higher-level participants including lexical phrases
It deals mostly
The change of people's behaviour
The story of the Noughties tells us where…
Various speakers give their opinions about the issue
They all point out the fact that…
You no longer need to be young to act young
As long as you can afford it
Adults no longer act like adults but rather like…
Massive change
The only way of communication was…
(The television)that not everybody could afford
Breathing fresh air
Time has changed
The meeting space
(In particular) through the most famous social network in the world.
Reachable everywhere and everytime.
A fixed rendezvous
They can't control what their children and their friends are doing.
A friend of mine
Mix reality and their life with what is happening
Act two different roles

Coll1: 16, which equates to 7,438.4 per million tokens

All 16 occurrences of general ineffective collocation in higher-level participants including lexical phrases
It deals on the change
Gains a lot of profit
Many have possibilities
The desire of (break the rules)
You can have Facebook (on a mobile phone)
(Our parents) follow the mould
'Kidulthood' is made up by two nouns.
Globalisation has much influenced…
(Every aspect of our life) synonym with…
Stylish young garment
Thoroughly discussed
Taking part of the discussion
Do this tendency (to a form of 'perpetual childhood')
(Possess) a consistent amount…
Amount to money

Lower levels (60-63% marks on pre-test)

Coll2: 77 instances, which equates to 27,827.97 per million tokens

The first 21 occurrences of general effective collocation in lower-level participants including lexical phrases
What interested me in particular
The idea of a new adult lifestyle
The current century
From different points of view
Everyday lives
They are obviously influenced by
Let's analyse these aspects in more details
Significant changes
Similar to the nineties
Without any kind of real need
View that period as
A new "Golden Age"
A terrible economic or world crisis
Seem to be unable to
Solve these problems
Conscious of the difficulties
Tend to create
A parallel world
Convinced that limits don't exist
Is due to the fact that
Attach too much importance

Coll1: 29, which equates to 10,480.66 per million tokens

The first 21 occurrences of general ineffective collocation in lower-level participants (60-63 marks) including lexical phrases
(Society has) deeply changed
Reason of (many social changes)
Guarantee to (their children)
The same of (young people)
Have a reality
An equal reality
Doing operations (with reference to medical procedures)
Live the present time
Live (well) our middle age
Deep changes
Ambiental problems
Thinking to the crisis
Go on retire
Discuss about (the phenomenon)
Spend money in (something)
Make some experiences
Give a bad example
different than the past
Convinced that limits don't exist
Is due to the fact that
The principal activities

Sharon Hartle, Università degli Studi di Verona
sharon.hartle@univr.it

EN **Sharon Hartle** is an Associate Professor in the Department of Foreign Languages and Literatures at Verona University. She is specialized in English Language Teaching (ELT) pedagogy and didactics and works specifically in the field of English for Specific Purposes (ESP). She has worked for years in the field of e-learning with a particular focus on multimedia lesson design for ELT in Blended Learning contexts. Her research interests also extend to include English Language Assessment, English Medium Instruction (EMI) and she is currently researching inclusive, accessible foreign language learning.

ES **Sharon Hartle** es profesora asociada en el Departamento de Lenguas y Literaturas Extranjeras de la Universidad de Verona. Está especializada en pedagogía y didáctica de la enseñanza de la lengua inglesa (ELT) y trabaja específicamente en el campo del Inglés con Fines Académicos (ESP). Ha trabajado durante años en el campo del e-learning con especial atención al diseño de materiales multimedia para ELT en contextos de Blended Learning. Sus intereses en el campo de la investigación se extienden también a la evaluación de la lengua inglesa y la enseñanza del inglés como lengua extranjera (EMI – English Medium Instruction). Actualmente investiga sobre el aprendizaje de lenguas extranjeras accesible e inclusivo.

IT **Sharon Hartle** è professoressa associata presso il Dipartimento di Lingue e Letterature Straniere dell'Università degli Studi di Verona. È specializzata nella pedagogia e didattica dell'insegnamento della lingua inglese (ELT) e lavora in particolare nel campo dell'inglese per scopi specifici (ESP). Da alcuni anni lavora nel campo dell'e-learning con un particolare interesse per lo sviluppo di materiali multimediali per l'ELT in contesti di Blended Learning. I suoi interessi di ricerca comprendono anche la valutazione della lingua inglese, l'istruzione in inglese (English Medium Instruction - EMI) ed è attualmente impegnata nella ricerca sull'apprendimento inclusivo e accessibile delle lingue straniere.