

Götz, S. & Granger, S. (2024). Introduction: Learner corpus research for pedagogical purposes: An overview and some research perspectives. In Götz, S. & Granger, S. (eds) (2024). *Learner corpus research for pedagogical purposes*. Special issue of the *International Journal of Learner Corpus Research* 10(1), 1-38.  
<https://www.jbe-platform.com/content/journals/10.1075/ijlcr.00039.got>

## INTRODUCTION

### **Learner corpus research for pedagogical purposes: An overview and some research perspectives**

Sandra Götz and Sylviane Granger  
Philipps University Marburg / University of Louvain

#### **1. Introduction**

For some time now, native corpora along with corpus linguistic tools and methods have been utilized in crafting learning and teaching materials, as well as in the creation of data-driven learning (DDL) scenarios, broadly defined as “any utilization of language corpora by second or foreign language learners” (Boulton, 2012: 263). However, the influence of learner corpora on language pedagogy (encompassing teaching, learning, testing and assessment) is only just beginning to emerge. In 2009 Granger stated that “there is undeniably very little evidence of fully-fledged up-and-running applications” (p. 24). Although the field of learner corpus research is growing rapidly, “pedagogical ‘implications’ are [still] much more numerous than ‘applications’” (Granger, 2015: 487). This is genuinely surprising, since learner corpus research and language pedagogy, despite having their own research agendas, share several commonalities: (1) they investigate the same type of data (learner language), (2) they pose the same types of research questions (e.g. finding out which variables and learning processes affect language learning or which specific difficulties learners face) and (3) they have similar objectives in terms of study implications (e.g. to better understand and improve language learning). Despite these common objectives, the most significant impact of learner corpus research for pedagogical purposes has been observed in the realms of materials development (see Section 2) and language testing and assessment (see Section 3). In the context of the classroom itself, even the studies presented in a previous volume on a similar topic (Götz & Mukherjee, 2019) offer surprisingly few specific suggestions on how the findings from learner-corpus-based (LC-based) research studies can translate into meaningful implications and applications for language pedagogy. It seems particularly striking that there are significantly fewer empirical studies testing the effectiveness of learner corpora in the classroom than examining native corpora. A recent survey conducted by Boulton and Vyatkina (2021) on the types of corpora utilized in DDL scenarios suggests that relatively little attention has been paid to the role of learner corpora in DDL. Only 4% of the studies investigated (21 out of 489) incorporate learner corpora. Lusta et al.’s review (2023) of studies focused on DDL in the classroom does not mention learner corpora at all. Furthermore, there has not been a systematic survey of the techniques and effectiveness of learner-corpus-based DDL (see, however, Section 4).

Learner corpora used in teaching and testing applications can be situated along a series of dimensions. Granger (2009: 24-26) distinguishes between corpora for *delayed pedagogical use* (DPU) and corpora for *immediate pedagogical use* (IPU). DPU corpora

are not used directly by the learners who have produced the data. They are compiled by academics or publishers with a view to designing pedagogical tools and methods which will benefit similar-type learners. IPU corpora, on the other hand, are collected by teachers as part of their normal classroom activities and the learners are simultaneously producers and users of the corpus data. Both types of corpus have their advantages and disadvantages. DPU corpora are larger and the findings are therefore more generalizable, but the data they contain may not be directly relevant for learners. IPU corpora tend to be much smaller but are arguably more motivating as students work on their own productions. The notions of global and local learner corpora are related to this first dimension. Gilquin (2015: 15) defines *global* learner corpora as corpora that are “part of large-scale projects and [being] collected among learners who are subjects providing data for inclusion in the corpus”. *Local* learner corpora, on the other hand, “are typically collected by teachers among their students”. They are usually produced and used by the same students but can also “come from the same population (typically, the same school/university)” (ibid.). Another dimension distinguishes between *generic* and *L1-specific* learner corpora. The former target all learners, whatever their mother tongue background; they have a wider user base and are therefore quite naturally favoured by publishers. The latter target exclusively learners from a specific L1; they tend to be collected by teachers who often work in mono-L1 contexts and are keen to design materials tailored to L1-specific needs. Another key dimension differentiates *direct* and *indirect* approaches. Going back to Leech’s (1997) taxonomy for native corpora, we classify the development of LC-based reference tools under the heading of indirect corpus applications, as opposed to direct applications, in which learners consult the learner corpus data themselves. Finally, a more general dimension contrasts *weak* and *strong* approaches to learner corpus use. In a weak approach, learner corpora are used as a mere repository of language to be used to check or document intuition-based interlanguage features. A strong approach has a much greater heuristic potential as the learner corpus is truly taken as the starting point for the analysis; in Tognini-Bonelli’s (2001) terminology, it is ‘corpus-driven’.

Over 20 years ago, Granger (2003: 542) enthusiastically suggested that “[l]earner corpus research opens up exciting pedagogical perspectives in a wide range of areas of English language teaching (ELT) pedagogy: materials design, syllabus design, language testing, and classroom methodology”. From our perspective, these primary areas of application have persisted to the present day. In this introduction to our special issue, we aim to assess how these different domains have been impacted by LC-based research and to outline future research directions. To this end, we will begin by providing an overview of recent advances in using learner corpora for materials design (Section 2), language testing and assessment (Section 3) and classroom practice (Section 4). We will conclude by highlighting some avenues for future research (Section 5) and providing an outline and brief summaries of the studies featured in this special issue (Section 6).

## 2. Learner corpus data for materials design

Incorporating insights derived from LC-based research yields manifold benefits in crafting pedagogical materials. As highlighted by Walsh (2010), learner corpora represent “arguably, one of the most useful ways to help learners understand their own problems

and develop new insights into their interlanguage system” (p. 233). This section aims to delve deeper into the ways in which learning materials can be enriched by integrating findings from LC-based research, with a specific emphasis on reference tools and teaching materials.

## 2.1 Reference tools

Undoubtedly, one of the most common applications of learner corpus data for pedagogical purposes lies in the realm of reference tool development, encompassing resources such as pedagogical grammars and learners’ dictionaries. Both fall into the category of indirect corpus applications. In this domain, learner corpora are utilized for delayed pedagogical purposes, indicating a time lapse between data collection and implementation, where the learners benefiting from these tools may not be the same individuals whose data populate the corpora (Granger, 2009). Within the sphere of reference tool development, learner corpora offer advantages in such indirect applications across three key levels: (1) selection, (2) description and (3) sequencing, which we will outline in the following discussion.

Firstly, LC-based studies can uncover instances of learners’ over- or underuse, as well as misuse, of various phonological, morphological, lexical, grammatical or lexico-grammatical features. Both quantitative and qualitative characterization of these instances can aid reference tool developers in selecting which phenomena to emphasize more prominently in learning materials, while also providing insights into those that may warrant less emphasis. Thorough analyses of such findings are invaluable in enabling developers to provide detailed descriptions of the features under scrutiny. Moreover, analysing error-tagged corpora adds significant value, as it allows researchers to adopt a bottom-up approach to identifying learners’ errors (Nicholls, 2003). With the expansion of learner corpora in terms of size, diversity and metadata, advanced statistical techniques facilitate increasingly nuanced analyses, enabling researchers to offer comprehensive descriptions of common stumbling blocks encountered by learners of a foreign language. Concurrently, these techniques enable predictions regarding the influence of learners’ native language, proficiency level, text type and other pertinent contextual variables on their performance (Götz, 2019; Granger, 2015; Mukherjee & Götz, 2015). Findings from such multivariate analyses, in turn, can guide reference tool developers in identifying linguistic features relevant to all learners of a specific foreign language (e.g. phrasal verbs in English as a foreign language). Additionally, they can assist materials developers in identifying features which it is particularly beneficial to include when tailoring materials to learners with specific native languages or proficiency levels, or even when utilizing certain text types or genres (e.g. addressing the confusion between simple past and present perfect tense for learners of English whose native language uses these tenses in different ways, such as German).

When it comes to the sequencing of linguistic features within reference tools, i.e. the order in which certain linguistic variables are dealt with, LC-based studies can also be of enormous help to materials developers. Here, the analysis of truly longitudinal or even ‘quasi-longitudinal’ (Granger, 2002: 11) learner corpora can be useful as a way of making materials developers aware of features that learners typically acquire at different proficiency levels. This information can be used to check the order of variables in their

materials against the findings from LC-based analyses that attest developmental stages empirically. If need be, materials developers can update the order of appearance of certain features accordingly (see also Section 3).

In the area of reference tool development, learner corpora have made a notable contribution, particularly in the domain of pedagogical lexicography (cf. Granger, 2003, 2015; Granger et al., 2007). For instance, the most renowned learner's dictionaries have been significantly influenced by analyses based on learner corpora, enabling compilers to incorporate valuable annotations. These annotations include sections dedicated to 'common learner errors' or 'get it right' boxes, which highlight particularly error-prone categories identified through learner-corpus-based studies. Prominent examples include the *Longman Dictionary of Contemporary English* (2014), the *Cambridge Advanced Learner's Dictionary* (2013), and the *Macmillan English Dictionary for Advanced Learners* (2007) (cf. Granger et al., 2007: 255). Although some comprehensive analyses of such annotations have suggested that they may not always align perfectly with learners' proficiency levels (De Cock & Granger, 2005), learner corpora undoubtedly add immense value to reference tools across various dimensions.

Learner grammars have not been impacted to the same extent by learner corpora. Similarly to learner's dictionaries, learner's grammars informed by LC-based studies draw primarily on error analyses that uncover common learner errors. These analyses have enabled grammar developers to augment specific grammatical categories prone to errors with error warnings. However, these warnings often bear a resemblance to those found in intuitive grammars, possibly due to a 'weak' approach adopted in LC-based analyses (cf. Granger, 2015). Notable examples of learner grammars, all derived from analyses of the *Cambridge English Corpus*, include the *Cambridge Grammar of English* (CGE; Carter & McCarthy, 2006), *English Grammar Today* (Carter et al., 2011) and the *Grammar and Beyond* series. Despite the enhancements made to learner grammars thus far, there remains significant potential for fully leveraging the pedagogical value that LC-based analyses can offer, as highlighted by Granger (2015: 492): "If we want to go beyond the traditionally recognized errors, a strong, i.e., truly corpus-driven, approach is necessary". Such an approach would not only unveil linguistic features not already identified as error-prone, but also shed light on the influence of learners' native languages, proficiency levels and other contextual variables.

## 2.2 Teaching materials

### 2.2.1 Textbooks

Learner corpora hold significant promise for textbook development. Firstly, they can uncover error-prone categories that many learners are likely to struggle with, encompassing both universal challenges and those specific to learners' native languages, necessitating special attention in textbooks (O'Dell, 2005). While textbooks often address universally recognized errors derived from learner corpus analyses, there are also instances where textbooks incorporate insights from analyses of language-specific errors, as seen in the *English in Mind* series, which features 'Get it Right!' sections focusing on typical Italian learner errors (Granger, 2015: 495). However, the potential of LC-based studies extends beyond error identification, as emphasized by Lorenz (1999: 5): "The

strength of learner corpus analysis lies in the detection of patterns, not errors”. In addition to identifying typical learner errors, such studies are equally valuable for textbook developers in uncovering instances of learners’ over- or underuse of certain language features, which can directly influence the frequency of particular constructions in textbooks. For instance, learner corpus studies have highlighted tendencies among English learners across various native languages to intensify adjectives in a less varied manner than native speakers, often relying excessively on certain intensifiers such as *very* or *really* while neglecting others (e.g. *particularly*), even at advanced proficiency levels (Granger, 1998; Lorenz, 1999; Schweinberger, 2020). In response to such findings, textbook developers could incorporate specialized exercises for practising more native-like adjective intensification and introduce a broader range of adjective intensifiers into dialogues. Furthermore, learner corpus analyses have identified constructions frequently used by learners that, while not necessarily incorrect, are considered unidiomatic (e.g. Martinez-Garcia & Wulff, 2012; Palacios-Martínez & Martínez-Insua, 2006). Including such features in textbooks can not only enhance learners’ language proficiency, but also minimize potential misunderstandings, thereby warranting their inclusion in instructional materials. Beyond error analysis, learner corpus data can serve to illustrate speech or text samples from peers at similar proficiency levels who have completed similar assignments (Staples et al., 2023), offering learners a more relatable and attainable target than native speaker data, thus potentially enhancing motivation.

Despite the wealth of pedagogically relevant features unearthed by LC-based research, their integration into textbooks by developers has been surprisingly limited. Even the number of textbooks drawing upon findings from native corpora is rather low, with notable exceptions such as the *Touchstone* series (McCarthy, 2014), which is primarily based on analyses of the *Cambridge International Corpus*, and *Academic Interactions* (Feak et al., 2009), which incorporates analyses of both native and non-native (albeit expert) speech data. Granger (2015: 495) suggests that this limited uptake may stem from the scattered nature of findings from LC-based research, making them less accessible to materials designers. Moreover, even in instances where collaborations between LC-based researchers and textbook developers occur (still relatively rarely; cf. Gablasova & Bottini, 2022), developers often hesitate to deviate from their established practices on the basis of a handful of research findings. However, recent large-scale LC-based projects such as *The English Profile* project, which prioritize language-pedagogical implications, along with websites promoting exercise taxonomies and concrete examples of how learner corpora enhance teaching materials, offer hope for change. Furthermore, LC-based researchers are increasingly encouraged to publish their findings in more accessible formats, focusing on aspects that should take precedence in teaching materials (Granger, 2013), potentially paving the way for LC-based findings to become integral to textbook development.

### 2.2.2 Local learner corpora for creating situated teaching materials

Local learner corpora are typically compiled from the productions of a specific learner group, making them immediately applicable for pedagogical purposes within that group. These corpora often consist of data derived from students’ assignments in class, projects, exams and similar sources (Chambers, 2015). Beyond the opportunity for students to engage with their own data in the classroom (see Section 4), teachers can also leverage

these data to create learning materials tailored to the specific needs of their learners. The advantage of utilizing local learner corpora in crafting customized teaching materials lies in their inherent pedagogical relevance: learners are likely to be highly motivated when they work with materials based on their own or their peers' data. Not only are such materials perfectly suited to the learners' proficiency level, but they also facilitate form-focused instruction effectively (e.g. Schmidt, 1990), enhancing the likelihood of successful second language acquisition (cf. Walsh, 2010). Furthermore, employing local learner corpora enables teachers to gain precise insights into the strengths and weaknesses of their learners, allowing them to personalize teaching priorities accordingly.

While gathering written student output can be relatively straightforward using electronic submission of assignments, compiling spoken learner corpora poses a somewhat greater challenge (Gablasova & Bottini, 2022). In the case of the latter, Pérez-Paredes (2003) offers a valuable suggestion: collaborative transcription of learners' spoken output can provide additional benefits to the learning process. This approach encourages learners to become more aware of potential issues in their own output as they transcribe their data.

Once more, despite the considerable pedagogical promise that local learner corpora hold for the development of contextually relevant teaching materials, their adoption by teachers remains constrained, notwithstanding the conducting of several learner-corpus-based classroom experiments (see Section 4). The obvious obstacles holding teachers back from integrating local learner corpora into their practice include time constraints and a dearth of corpus linguistic training in teacher education programmes at universities. In our assessment, incorporating (learner-)corpus-based pedagogy as a mandatory component of university teacher training is imperative in order to address this shortfall.

### 2.2.3 *Computer-assisted language learning (CALL) programmes*

The domain of CALL, encompassing technology-supported language teaching, has experienced significant growth. Its tools offer substantial benefits to learners seeking to acquire a foreign language independently, outside the confines of a traditional classroom setting. Recent intelligent CALL programmes (ICALL; cf. Meurers, 2015) reflect the vision articulated by Granger et al. (2007):

One particularly promising avenue for future progress in the field [of CALL] comes from the field of Natural Language Processing (NLP). Combining use of NLP techniques such as POS-tagging, parsing, lemmatization and other processing and error-tagged corpora, makes it possible to design a wide range of resources, ranging from highly sophisticated automatic error detection and correction systems to simpler, but potentially very useful web-based interfaces to authentic errors and their corrections. (Granger et al., 2007: 253)

While direct implementation of LC-based (I)CALL programmes remains relatively uncommon, several noteworthy examples merit attention. Hegelheimer (2006) and Hegelheimer and Fisher (2006) present compelling cases in this regard. Their studies underlie the *Internet Writing Resource for the Innovative Teaching of English* (iWrite), an online interface constructed from a corpus of learner texts enhanced with XML

annotations. This resource, tailored to intermediate non-native speakers (NNS) of English, enables users to explore a fully error-tagged learner corpus, facilitating targeted searches for specific errors or sorting by error types presented in context alongside the correct target hypothesis. Through engagement with this tool, learners enhance their grammatical awareness and proficiency in rectifying errors in their writing (Hegelheimer, 2006). Similarly, Granger et al. (2007) developed a comparable tool for French as a foreign language (FLL): the *EXample eXtractor Engine for LANguage Teaching* (Exxelant). This searchable online interface, based on a fully lemmatized and POS-tagged corpus, comprises over 20,000 errors, serving as a valuable resource for learners and prospective FLL instructors alike. While these tools primarily address writing skills, Campillos Llanos (2012) devised a searchable interface based on a spoken Spanish learner corpus, fully error-tagged and encompassing learners from diverse L1 backgrounds. Furthermore, automated error correction tools have proved beneficial for language learners (Umezawa et al., 2013). One case in point is the *Cool English* interface, which is a grammatical error correction system based on large learner corpora (Lo et al., 2018). Despite the obvious utility of and interest in these programs, the majority of them are, regrettably, not publicly accessible.

Indirect applications are more commonly encountered in (I)CALL contexts. Here, learner corpora are typically scrutinized to identify the most frequent or persistent errors, which are then integrated into CALL programs. Noteworthy examples include *GrammarTalk* (Chuang & Nesi, 2007), an e-learning grammar resource that targets two prevalent error types identified in the corpus analysed (articles and prepositions). Similarly, the *ESL Writing Tutor* (Cowan et al., 2003, 2014) focuses on addressing a selection of frequently occurring errors among Korean learners of English. Broader in scope are applications such as *Check My Words* and *Mark My Words* (Milton, 2006), both drawing on analyses of a Chinese learner corpus encompassing around 100 lexico-grammatical or stylistic errors. These tools enable learners to assess their language usage against online resources while writing (*Check My Words*) and afford teachers the same capability while grading (*Mark My Words*). Apart from analysing errors, another very versatile CALL interface, where learners can work with their own texts, is the *Compleat Lexical Tutor*,<sup>1</sup> which includes (among various other DDL activities) a tool where learners can upload their own texts and have them analysed, e.g. for their similarity to native speaker texts on the basis of word frequencies, as well as in terms of their use of collocations, transition phrases and frequently used idioms.

Even more sophisticated and intricate ICALL programs are engineered to provide automatic feedback on the accuracy and style of learners' texts. It is worth noting, however, that while such automatic feedback generators may be highly beneficial, achieving a satisfactory level of precision in their implementation remains challenging (Meurers, 2015; Wanner et al., 2013). That said, one ICALL program that has demonstrated considerable success is *E-Tutor*. Developed for German learners in conjunction with face-to-face instruction (Heift, 2001; Heift & Caws, 2013), this tool utilizes learners' input to generate feedback based on linguistic data analysis. There is little doubt that Artificial Intelligence (AI) will further augment such ICALL applications

---

<sup>1</sup> <http://www.lex tutor.ca> (last accessed on 2 February 2024).

in the future, making them into even more adaptive and personalized language learning environments (Schmidt & Strasser, 2022).

As Falk and Götz (2016) point out, maybe the most significant contributions of learner data to (I)CALL programs are made by language learning applications such as *Duolingo*.<sup>2</sup> Although we cannot consider the data that are entered by users to be learner corpora, maybe these kinds of ‘big data’ can give us an insight into how they might be combined with the use of learner corpora for (I)CALL applications in the future. The data language learners contribute probably form one of the principal innovations in the field of ICALL, since each touch on a screen and each key stroke a language learner makes on an app is stored and analysed by researchers behind the app. The huge effects that such analyses can have on language learning are explained by the CEO of Duolingo as follows:

For example, right now we’re teaching adjectives after plurals, so let’s try teaching adjectives first. We pick a subset of 50,000 users and see if they learn better – if they come back more often and make fewer mistakes. We can see the statistics and if they’re positive, switch all users to the new method. We then give the community making the courses access to the same information so they can run experiments to work out how to do better. (Luis von Ahn, CEO of Duolingo in Rutkin, 2013)

This concept, known as “crowdlearning” (Llorente, Morant & Garrigos-Simon, 2012), posits the advantage that a vast number of learners bring together diverse skills and amalgamate their knowledge. While such adaptable methods of tailoring the learning environment to suit the majority of learners offer numerous benefits, von Ahn (2013) concedes that “it’s going to take some years to figure out what will actually work. [...] But I do think we’re in the process of figuring out what the next wave of education will look like”. Despite the promise of such approaches, Jordan (2015) highlights the risk of solely focusing on big data while “ignoring decades of SLA research on didactic principles of meaningful grammatical progression”. Here, future research endeavours evaluating the merits and drawbacks of strictly data-driven learning approaches will determine whether such criticisms are justified.

### 3. Learner corpus data for language testing and assessment

Learner corpora have significantly impacted language testing and assessment (LTA) (cf. Barker et al., 2015; Callies & Götz, 2015a, b). Callies et al. (2014: 76) provide a helpful categorization of how learner corpora are applied in LTA, distinguishing between corpus-informed, corpus-based and corpus-driven approaches. In corpus-informed approaches, learner corpora offer practical insights into learners’ language usage at specific proficiency levels, aiding in test development or validating human ratings. Corpus-based approaches view the corpus as a valuable resource for linguistic research, allowing researchers to test existing hypotheses about learner language or to differentiate between proficiency levels. Finally, corpus-driven approaches involve no predetermined hypotheses before conducting corpus analyses, focusing instead on computer techniques for data extraction and evaluation. These three types of application will be further explored in the discussion that follows.

---

<sup>2</sup> <https://www.duolingo.com/> (last accessed on 2 February 2024).

### 3.1 Test development

One domain where LC can be usefully employed is test development. A notable early example illustrating how LC-informed research can shape test creation comes from Hargreaves (2000). He identified collocations used by learners across various proficiency levels in the *Cambridge Learner Corpus* (CLC) and devised a novel test format focusing on collocational knowledge. Usami (2013) offers an intriguing proposal on making use of Japanese LC data to enhance distractors in multiple-choice questions, citing examples of inaccuracies identified through LC-based error analyses. In a more recent investigation, Gyllstad and Snoder (2021) delve into the phraseological repertoires of L1 Swedish and L1 Italian learners' written English, drawing on data sourced from the third edition of the *International Corpus of Learner English* (ICLEv3; Granger et al., 2020). Their findings on V+N collocations open avenues for various test development applications: they illustrate how commonly occurring collocational errors can serve as distractors in receptive multiple-choice item formats. Furthermore, they propose strategies such as avoiding the most obvious high-frequency collocates of a frequent noun in a test, opting instead for lower-frequency yet conventionalized verbs with higher mutual information (MI) scores (Gyllstad & Snoder, 2021). In tests evaluating productive collocational knowledge, they suggest rewarding collocates with higher MI scores with correspondingly higher test scores.

While these methods hold significant promise, they are still relatively underutilized and have not escaped criticism entirely. Common criticisms often revolve around concerns regarding the representativeness of learner corpora (Gyllstad & Snoder, 2021) or the potential for circularity when learner output is used to inform test development assessing the use of the very same feature (cf. Hulstijn, 2010).

### 3.2 Learner corpora to inform descriptor scales

Proficiency testing, which assesses a learner's proficiency level, is the most commonly employed form of LTA (Barker et al., 2015), thus warranting focus in this section. Typically, a learner's language abilities are evaluated by measuring specific language skills – listening, speaking, reading or writing – on the basis of their performance in a standardized language test. The learner's test performance is typically measured against a predefined framework, such as the *Common European Framework of Reference* (CEFR; Council of Europe, 2017), in which trained raters gauge what learners 'can do' at various proficiency levels. In this realm of testing, LC-based research has found frequent application, with learner corpus analyses increasingly informing descriptor scales. This is typically achieved by scrutinizing specific language features, such as lexical or grammatical aspects like tense/aspect, and furnishing detailed accounts of how these features evolve across learner groups with increasing proficiency (e.g. Díez-Bedmar, 2015; Dumont, 2017; Gilquin, 2022; Götz, 2015; Leńko-Szymańska, 2015; Thewissen, 2012, 2013; Wisniewski, 2017). Insights gleaned from such analyses can refine existing descriptor scales such as the CEFR, enriching them with more nuanced and substantive information regarding learners' actual capabilities at different proficiency levels. Hence, LC-based research helps to distinguish between different proficiency levels by identifying "criterial features" (Hawkins & Buttery, 2010; Hawkins & Filipović, 2012), thereby

augmenting “CEFR’s functional characterization of the different levels” with grammatical and lexical nuances (Hawkins & Filipović, 2012: 5). Analysis of truly longitudinal corpora, enriched with error tagging and/or information on learners’ proficiency levels, would also prove highly valuable for such endeavours.

One of the most influential and expansive projects in this domain is *The English Profile* project (Salamoura, 2008). The overarching objective of this project is to categorize, organize and compare learners’ authentic output across various proficiency levels in order to refine the CEFR levels. Leveraging the substantial size of the learner corpus employed – the *Cambridge Learner Corpus* (CLC), a 55-million-word corpus comprising over 250,000 scripts from Cambridge English exams spanning all proficiency levels and sourced from over 130 countries –, the refined levels can be referenced with significantly greater certainty than anything preceding this project. Consequently, it holds immense potential for informing test writers about learners’ expected performance and suitable learning content at specific proficiency levels, helping researchers to provide a comprehensive and exceedingly detailed description of learners’ progression (Götz, 2022). Furthermore, the project has significant implications for materials developers, offering insights into content suitable for learning at particular proficiency levels. All materials derived from the project are freely accessible via an online interface, serving as valuable resources for researchers, language teachers and test developers alike. Nonetheless, it should be noted that this project is not without its drawbacks. As Granger (2015: 506) aptly observes, the corpus “is linked to a well-defined syllabus, which may bring about some degree of circularity: the learners master a particular language feature at a particular level because it is part of the syllabus at that level”. Swan (2005: 96) similarly cautions against basing standard language descriptors solely on analyses from a single examining body.

While the original version of the CEFR was intended to have universal applicability across all languages, the Council of Europe has encouraged the development of Reference Level Descriptors (RLDs) tailored to national and regional languages, aimed at customizing the descriptors for local contexts and objectives. This recommendation has prompted initiatives such as the CEFR-J project, which created a Japanese version of the CEFR. This adapted version of the CEFR reflects the specific characteristics and requirements of Japanese learners of English (Negishi et al., 2013). In the CEFR-J, a meticulous analysis of learner corpora formed the foundation for detailed elaboration of the Japanese version of the CEFR, encompassing the unique traits of Japanese learners of English as they progress through various proficiency levels. A parallel effort is underway with the *FineDesc* project,<sup>3</sup> currently developing a Spanish version of the CEFR based on a Spanish EFL learner corpus.

In general terms, we firmly contend that learner corpora have the potential to render the assessment of L2 proficiency based on descriptor scales more transparent and consistent, and ultimately tailored to specific L1 backgrounds.

### 3.3 Automatic assessment of test and learner data

---

<sup>3</sup> <http://web.ujaen.es/investiga/finedesc/> (last accessed 20 February 2024).

With significant advances in natural language processing (NLP) and machine learning, automatic assessments of learners' test data are increasingly reliable, particularly in the case of written data (Kyle, 2021a, b; Marchand & Atsuku, 2015; Yannakoudakis et al., 2011). Many prominent language tests, such as the *Test of English as a Foreign Language* (TOEFL) or the *Scholastic Aptitude Test* (SAT), incorporate constructed-response formats (i.e. students are required to formulate their own answers) or essays, making automatic essay scoring systems invaluable resources in language testing. According to a survey by Higgins et al. (2015: 589-590), notable scoring systems include the *Intelligent Essay Assessor* (IEA; Landauer et al., 2003), the *e-rater* (Attali & Burstein, 2006), the *Project Essay Grade* (Page, 2003), and *IntelliMetric* (Rudner et al., 2006). These systems utilize learner corpus data for calibration, training on samples of authentic, pre-scored learner essay data across various proficiency levels. The tools are then evaluated on separate validation samples, also pre-scored by humans. The number of learner corpus files utilized for this process typically ranges from 400 to 1,200. The models developed for automatic scoring systems are based on the training set and tested on the validation set until satisfactory agreement is reached. Developers assess agreement using various criteria, including correlation measures and standardized mean score differences, comparing evaluations between human raters and automatic systems. The criteria for automatic scoring are largely informed by contrastive corpus analyses of learner versus native data, encompassing aspects such as word selection, errors and grammatical patterns (Kyle, 2021a). In essence:

by applying corpus linguistic techniques to learner corpora, it is possible to identify the features that characterize learner language, distinguishing between features shared by several learner populations and those that are typical of one particular learner group, thereby gaining a clearer picture of the distance that still separates the learner's language from the targeted 'norm'. (Granger et al., 2007: 253)

Error-tagged learner corpora thus play a crucial role in informing automatic assessment tools. Thorough error analyses provide valuable insights that can predict the quality of learner texts (e.g. Bestgen & Granger, 2011) and aid in the development of NLP algorithms designed to detect common learner errors (Higgins et al., 2015: 590).

Automated scoring of spoken test data remains relatively underdeveloped compared with its counterpart in written assessment. Evaluating speech presents more complexities and nuances than assessing written text. While certain aspects such as fluency or pronunciation can already be measured semi-automatically, others pose significant challenges for automated assessment (cf. Higgins et al., 2015). The training process for automatic scoring systems for spoken language follows a similar pattern to that of written assessments. They are trained using human-scored, error-tagged learner data and then tested on another set of similarly tagged data. However, it is important to note that automated scoring in speech assessment is not as accurate as human scoring, necessitating cautious consideration when applied (Zechner et al., 2009).

The focus of attention in automatic scoring systems of spoken test data has largely been on computer-assisted pronunciation training (CAPT) systems. These systems offer non-native English speakers targeted feedback on their pronunciation errors (Higgins et

al., 2015: 595). Often integrated into language learning applications such as *Babbel*,<sup>4</sup> CAPT systems are developed using transcribed and phonetically annotated learner corpora. Particularly valuable are corpora that are error-tagged for pronunciation errors, as demonstrated in a study by Wang and Lee (2012).

Recent advances in AI-informed tools offer significant potential for language learners, particularly in the realm of automatic grammar- and spell-checking, exemplified by platforms such as *Grammarly*<sup>5</sup> or *Language Tool*.<sup>6</sup> Although these tools are now commonly integrated into various writing software platforms, e.g. *Word*®, mainstream options typically rely on errors made by native speakers (Leacock et al., 2015) and are found to be more beneficial for them than for language learners (Rimrot & Heift, 2008). Nevertheless, leveraging error-annotated learner corpora has enabled considerable progress and technical enhancements in the development of automatic error-detection systems (Han et al., 2010; Rozovskaya & Roth, 2010; Tetrault et al., 2014) and spell-checkers (Mitton & Okada, 2007). Despite these promising advances, the integration of such tools into L1-specific writing software programs appears to be a formidable challenge that requires further exploration and development.

#### 4. Learner-corpus-based classroom practice

Direct applications of learner corpus data in the classroom are much less common than indirect applications: “the dominant trend in learner corpus research still remains analysis of learner data rather than empirical studies of the use of learner data in the learning and teaching context” (Chambers, 2015: 456). In direct approaches “learners explore corpus data either through activities designed by their teachers or independently by getting unmediated access to the data therein (Fuster-Márquez & Gregory-Signes, 2018: 166.). These two types of corpus exploration are sometimes referred to as direct and indirect use (see e.g. Yoon & Jo 2014: 9 and Samoudi & Modirghamene 2022: 701). To avoid any confusion with the more traditional use of the terms ‘direct’ and ‘indirect’, we prefer to use the terms ‘mediated’ and ‘unmediated’ direct use of LC data. In both cases the LC-based activities fall within the confines of DDL. The theoretical underpinnings are essentially the same as those of DDL activities relying on native/expert corpora (inductive approach, focus on form, noticing, consciousness-raising) but involve some specificities due to the nature of the data (smaller datasets, developmental perspective, focus on errors, etc.).

While learner corpus data have had a much smaller impact on DDL-based classroom practice than native corpus data, an informal survey of the literature reveals that LC-based studies are more numerous than one would expect. The list of 33 studies provided in the Appendix bears testimony to the growing activity in the field. It must be acknowledged, however, that direct use of learner corpus data has not caught on in a major way over the years, one of the possible reasons being that teachers are reluctant to expose learners to erroneous output. But for Chambers (2015: 462), learner data is “a crucial part of a pedagogic corpus, with the condition that it is recommended to give the

---

<sup>4</sup> <http://www.babbel.com> (last accessed on 24 January 2024).

<sup>5</sup> <http://www.grammarly.com> (last accessed on 24 January 2024).

<sup>6</sup> <https://languagetool.org/> (last accessed on 24 January 2024).

learners access to the erroneous data only in conditions carefully prepared by the teacher”. Similarly, for Mendikoetxea et al. (2010: 184), exposing learners to errors produced by themselves or other learners is not a problem “as long as the errors are appropriately signaled”. In addition, several authors insist on the need to complement learner data with native data and to include follow-up exercises “to stabilise correct usage” (Nesselhauf, 2004: 140).

In this section we first describe some of the key design features of LC use. We then provide a survey of LC-based pedagogical approaches before reporting on the efficiency and students’ perception of LC-based activities.

#### 4.1 Design features of LC use

As rightly pointed out by Forti (2023a, n.p.), “[t]he potential of a learner corpus in DDL activity development is inextricably linked to its design”. As we have pointed out above, one key design feature is related to the pedagogical use to which the learner corpus is to be put, which can be immediate or delayed (Granger, 2009: 24-26). The two types are evenly represented in the LC-based literature: some authors (e.g. Ackerley, 2013, Novikov & Vinokurova, 2022) rely on DPU corpora, others (e.g. Seidlhofer, 2002; Yoon & Jo, 2014) on IPU corpora. In some studies, both types of data are used. Fuster-Márquez and Gregory-Signes (2018: 165), for example, complemented their IPU corpus with a larger DPU corpus to ensure that the selected language features were sufficiently relevant for the whole group.

Another important contrast is that between synchronic and longitudinal approaches. Only the latter aims to track the development of learners’ proficiency. Authors who have adopted a synchronic approach include Hartle (2023), who has learners focus on their own writing in order to improve their use of collocations, and Charles (2018), who uses a similar approach to enhance doctoral students’ general editing skills. In many cases, however, studies are longitudinal: they compare learner use before and after a corpus-based DDL intervention (Ackerley, 2015; Moon & Oh, 2018; Pérez Cañado & Díez Bedmar, 2006). Much rarer is the use of pseudo-longitudinal learner data, i.e. data that are developmental but not longitudinal, as the data were not produced by the same learners but by different learners at different proficiency levels (Forti, 2023b).

A third important design feature relates to the inclusion or not of a reference corpus. The majority of LC-based activities involve a reference corpus in addition to the learner corpus. In most cases, the reference corpus is a large native and/or expert corpus, such as the *British National Corpus* (BNC; Rankin & Schiftner, 2011) or the *Corpus of Contemporary American English* (COCA) for English (Orenha-Ottaiano, 2012). As these corpora are generic and therefore arguably less attuned to the needs of students targeting a particular discipline (e.g. chemistry or law), some researchers prefer to use discipline-specific reference corpora which are more relevant to learners’ needs (Charles, 2018; Cotos, 2014). Some rarer studies rely on corpora of novice native writing, such as the *Louvain Corpus of Native English Essays* (LOCNESS) (Ackerley, 2013, 2015) or graded readers (Moon & Oh, 2018), regarded as more realistic baselines. Staples et al. (2023) go one step further and suggest using learner texts as models for other language learners, but

acknowledge “the value of bringing in models (e.g. expert or L1 corpora) from outside the learner’s immediate context”.

A fourth feature concerns the format in which the CL data are presented to learners. Concordances are very popular and prove to be a particularly effective way of identifying lexico-grammatical patterns. However, several studies show that it is possible, even necessary, to use whole sentences (Fuster-Márquez & Gregory-Signes, 2018), longer excerpts or even whole texts (Seidlhofer, 2002) for a range of other features. As is clearly apparent from the title of her article (“Corpus-assisted editing for doctoral students: More than just concordancing”), Charles (2018: 24) argues that “corpus tools have much more to offer the student-editor than just the concordancing of lexicogrammatical items” and demonstrates the value of designing exercises that rely on tools such as clusters, word lists and keyword lists.

#### 4.2 Survey of LC-based pedagogical approaches

LC-based activities aim to expose student groups to learner corpus data with a view to fostering their inductive skills and raising their awareness of specific linguistic items, as well as general patterns of use or genre conventions. Whether their objective is for learners to notice the gap between their use of the language and the targeted language variety or to heighten their awareness of the errors they tend to produce, studies that report on LC-based DDL activities all emphasize the reflective nature of these activities. For example, Hartle (2023) focuses on learners’ lexical difficulties, in particular those involving collocations, but insists that the learning design did not focus exclusively on error correction but also “combined elements of awareness raising, reflection and analysis” (p. 18).

Our survey of LC-based studies brings out some common trends. First, the DDL data are usually mediated, i.e. the teacher selects a number of linguistic items or patterns that they deem worthy of focus in the classroom. Although the selection is sometimes made on the basis of the teacher’s intuition, it is usually based on prior investigation of a learner corpus as compared with a reference corpus (Belz & Vyatkina, 2008). Unmediated activities, which put learners in direct contact with the LC data without prior intervention by the teacher, are much rarer. One notable exception is Seidlhofer (2002), who has learners carry out two intertextual tasks (summarizing a text and providing their personal reaction to it) and lets them explore the texts in class. The activity is clearly student-led as the learners were asked “to make a note of all the questions that came to their mind while looking at their joint product” (p. 225). Second, paper-based LC-based activities, often in the form of printouts of edited concordance lines, tend to be favoured over hands-on ones. When hands-on activities are integrated into the teaching design, they usually involve native rather than learner data. Third, as is generally the case in DDL, the linguistic focus is predominantly lexical or lexico-grammatical (Rankin & Schiftner, 2011; Todd, 2001), but grammar and discourse features are also targeted (Hsieh & Liou, 2008; Pérez Cañado & Díez Bedmar, 2006).

The LC-based activities designed by teachers are extremely varied and involve a series of steps which are described in great detail in some studies, thereby facilitating replication by other teachers (see e.g. Belz & Vyatkina, 2008 and Hartle, 2023). In spite of their diversity, LC-based activities can be seen as falling into two broad categories:

comparison of LC data and NS/expert data (Section 4.2.1) and error identification and/or correction (Section 4.2.2).

#### *4.2.1 Comparison of LC and native/expert data*

Many LC-based studies involve a comparison between learner corpus data and a reference corpus, usually a corpus of native/expert productions. The earliest examples of this type of activity are provided by Tribble and Jones (1997 [1990]) and Granger and Tribble (1998). The idea is to use concordances drawn up from learner corpora “in parallel with native concordances to make learners aware of grammatical, lexical or discourse features which distinguish their interlanguage from the target norm” (Granger & Tribble, 1998: 201). These paired concordances are intended to raise learners’ consciousness of error-prone linguistic items or structures, i.e. those that they tend to misuse, such as learners’ use of the *to*-infinitive after the verb *accept* and the adjective *possibility* (Granger & Tribble, 1998: 202-203). In addition, as shown by Millar and Lehtinen (2008), use of word list and keyword analysis can help identify language features that are over- or underused by learners, and these features can also be the focus of DDL activities. Moon and Oh (2018), for example, present Korean EFL learners with NS and NNS edited concordance lines with a view to raising their awareness of overgenerated *BE*. The idea that underlies this contrastive approach is that, in line with Schmidt’s (1990, 1993) Noticing Hypothesis, noticing the gap between learners’ productions and the targeted baseline will help learners convert input into intake. Most of the time, however, the comparison between learner and reference data is not carried out by the learners themselves but by the teachers, who use this method to design DDL exercises tailored to learners’ attested needs (Ackerley, 2013).

#### *4.2.2 Error identification and/or correction*

Among the most popular types of LC-based DDL activities are those that are focused on errors, the objective being to get learners to identify and/or correct errors extracted from learner corpus data. Error-oriented instruction has often been criticized for representing a defective view of language learning (Aston, 2008). However, many teachers view error-related activities in a much more positive light. For Zaki (2021: 195-196), for example,

making errors in language learning is an integral constituent of the learning process. Hence, errors are more of an opportunity rather than a hindrance. For pedagogues who support self-correction approaches in the language classroom, errors are an opportunity to develop the learner’s autonomy.

Identification and correction of errors is one particularly effective way of improving learners’ accuracy, a worthwhile objective especially in contexts where accuracy is highly valued or even required: “While in other contexts error only tends to matter as long as it impedes communication, correctness and accuracy are highly valued in an academic context such as the one we are describing” (Mendikoetxea et al., 2010: 185). However, some researchers, aware that it is important to “focus on success in learner discourse, not merely on failure’ (Aston, 2008: 348), include both correct and incorrect usage in their DDL exercises (Forti, 2023b; Hartle, 2023; Klimova et al., 2021).

Two types of error-oriented DDL exercises can be distinguished according to whether learners are asked:

- 1) to identify and correct errors in LC corpus sentences and/or excerpts selected by the teacher (Fuster-Márquez & Gregori-Signes, 2018; Pérez Cañado & Díez Bedmar, 2006; Todd, 2001).
- 2) to correct errors identified by the teacher and presented in LC corpus sentences and/or excerpts (Klimova et al., 2021; Mendikoetxea et al., 2010; Zaki, 2021)

For the correction part of the activity learners are either asked to draw up concordances of the targeted items in a reference corpus or to use paper-based concordances prepared by the teacher and, on that basis, to try and induce patterns of correct usage.

To design error-focused DDL exercises, teachers sometimes make use of error-coded learner corpora (Klimova et al., 2021; Mendikoetxea et al., 2010; Zaki, 2021). The advantage of using annotated data is that it is easier for teachers to design exercises that target specific problematic areas (e.g. articles, prepositions or verb morphology).

#### 4.3 Efficiency and perception of LC-based activities

Whether LC-based data should be used in language instruction depends on the answers to two crucial questions: Does LC-based DDL work and how is it perceived by the learners?

##### 4.3.1 *Effectiveness*

There are several methods that can be used to assess whether the LC-based intervention is effective or not. One frequently used method consists in comparing learners' performance before and after a corpus intervention. A survey of the literature reveals that few studies provide objective measures of effectiveness. Those that do provide generally positive results but point to a few mitigating factors. For one thing, results vary depending on the linguistic phenomenon investigated. This is demonstrated by Pérez Cañado and Díez Bedmar's (2006) study that shows significant improvement between pre- and post-tests for the categories of articles, prepositions and verbal complementation, but not for verb tenses. Similarly, Samoudi and Modirghamene's (2022) study also produced mixed results: it showed that learners were able to write more accurate and fluent essays after the corpus instruction but the results demonstrated no improvement in syntactic complexity. Another determining factor is the learners' proficiency level. While Zaki (2021: 206) obtained highly positive results in an error-correction activity, he noted that high-intermediate-level students performed better than those at the low-intermediate level. Some studies revealed no difference in learners' performance before and after the corpus intervention. For example, Gaskell and Cobb (2004) noted no improvement in error rates after 15 weeks of treatment. They attribute this disappointing result to the complexity of the linguistic phenomena focused on in the DDL activities and the brevity of the time frame: "if we are going to follow the development of complex grammatical patterns in our learners, then one school term is insufficient" (p. 315).

As it only relies on one group of learners who received the same type of instruction, the method described above cannot reveal how the LC-based intervention compares vis-

à-vis other types of instruction. To establish this, it is necessary to have recourse to another method, one that includes a control group. Gaskell and Cobb (2004) list the integration of a control group as one of the desiderata for the future and describe the control group that could be included within the framework of their study as follows: “This would be an equivalent class working on the same error categories but using some means of feedback requiring the same time and effort that did not involving [sic] concordancing” (p. 316). In spite of the undeniable benefits that can be derived from a control group, the reality is that few studies include one. This is due to the practical difficulty of setting up dual teaching experiments: duplicating classes or splitting a class into two groups receiving different pedagogical treatments is impracticable in the majority of real-life language teaching contexts. Teachers have to work with “the constraints of the instructional context, which are often inevitable in instruction-embedded research” (Cotos, 2014: 206). A few studies, however, do include a control group. For example, Moon and Oh (2018: 59) compared groups of Korean EFL learners who received traditional grammar learning to DDL groups and found that the latter outperformed the former with respect to the target feature (overgenerated *BE*). A similar comparison was carried out by Klimova et al. (2021). The results showed improvement in the two groups of students but failed to reveal any significant difference between the two instruction modes. A different type of comparison was carried out by Cotos (2014). She compared a group of students who were exposed to DDL activities that only included native speaker data to a group of students whose DDL activities involved both native and learner corpus data. The results showed that both groups improved, but the group of students who were exposed to data that also contained their own writing showed more significant improvement in their use of adverbials than the group who were only exposed to native data. She interpreted these findings “as suggestive that combining learner and native-speaker data is a feasible and effective practice, which can be readily integrated in DDL-based instruction with positive impact” (p. 202).

#### 4.3.2 Perception

As pointed out by Fuster-Márquez and Gregory-Signes (2018: 173), “[a]n important ingredient of any methodology is that it wins the acceptance of learners”. For Gaskell and Cobb (2004: 315), this is even more important than actual language gains: “Our purpose in this project was not to discover a miracle cure for sentence errors but to see whether learners could enjoy and use networked concordancing as a learning tool, and this seems to have been borne out”. Similarly to what is reported on DDL activities based on native corpora, practically all the LC-based studies in our survey that gave an account of learners’ perception report positive appreciation. This is especially the case when learners are exposed to their own writing. Cotos (2014: 217), for example, observes that enabling learners to analyse their local corpus “created a motivating learning environment where the corpus-based tasks were perceived as a personally relevant learning experience”. This is echoed by Fuster-Márquez and Gregory-Signes (2018: 174), who note that learners were especially curious about the specific errors they and their peers produced and were enthusiastic about the activities as they “gave them a chance to escape from the routine of textbook activities”. This is echoed by Mendikoetxea et al. (2010: 184), who observe that “[s]tudents seem to find error correction exercises motivating and fun”. In the same vein, Novikov and Vinokurova (2022: 62) report that the activities were “generally well-

received by students: such language puzzles are more exciting for them than simple explanations of grammatical rules”.

Although teachers’ reports are undeniably worthwhile, more objective reports can be obtained by asking learners to fill in a course appreciation questionnaire. This method is rarely used but there are some exceptions and, as is the case with the more informal reports, the survey results are quite positive. For example, Klimova et al.’s (2021: 120) survey confirms “the pedagogical value of the DDL approach in terms of the motivation and attitude of both teachers and students”. Similarly, Moon and Oh (2017: 58) find that the vast majority of the learners – 92% (n= 87) – showed positive perceptions”.

## **5. Discussion and outlook**

While Flowerdew’s (2012) observation that learner corpora are “‘the poor relation’ as far as pedagogic applications are concerned” is still true today, the present overview shows that they have already had an impact on a wide range of pedagogical tools and methods. While their impact may not be as pronounced as that of native corpora, it is nonetheless tangible. With the present overview, we hope to have shown how undeniably beneficial LC-based research can be for language pedagogy.

In general terms, we can assert that LC-based research has primarily impacted the development of reference tools, as some of the best-known learner dictionaries are informed by LC-based research. By offering valuable insights into learners’ language learning processes (including learners’ strengths and weaknesses), LC-based research has facilitated the creation of more precise, pertinent and tailored resources to aid language learning. Dictionaries have experienced greater benefits than grammars, with the impact on the latter still being relatively constrained. Nevertheless, we believe we have demonstrated the considerable potential as regards the selection, description and sequencing of grammatical phenomena by the application of a truly corpus-driven approach, which we anticipate will be embraced by grammar developers in the future.

While the greatest number of well-documented applications can be reported in the development of reference tools, we consider it unfortunate that we are not witnessing the same degree of uptake in the development of teaching materials. This is probably because LC-based findings are not widely accessible and are predominantly derived from small corpora and/or specific learner groups. This is a problem, since textbooks tend to be generic, i.e. expected to cater for all learners at a particular proficiency level. We hope that future large-scale projects (as modelled by *The English Profile* project) will persuade textbook developers to capitalize more on the advantages offered by LC-based research. From our perspective, this situation may only improve with significantly stronger collaborations between LC-based researchers and textbook developers.

Our literature review has also uncovered a scarcity of systematic reports regarding the utilization of local learner corpora for developing situated teaching materials. Although we recognize that teachers may regularly implement these practices in their classrooms without documenting them in articles, it is more likely that many teachers do not. This could be attributed to time constraints or a lack of corpus literacy among teachers. Therefore, we strongly advocate the inclusion of corpus linguistics, DDL and

LC-based pedagogy as mandatory subjects in teacher education programmes at universities.

Although LC-based (I)CALL programs are still the exception rather than the rule, they are a particularly valuable type of resource as they allow learners to get immediate feedback on their productions and foster autonomous learning. The combination of LC-based research with CALL and NLP techniques is a particularly promising avenue for the future, once the accuracy of the tools increases (cf. Kyle, 2021a). We recognize the greatest potential for development in these applications in their capacity to automatically detect errors in learners' texts and to offer tailored feedback that addresses specific areas of weakness. Additionally, with the aid of adaptive learning systems, ICALL programs can assess learner corpora and customize content and activities. This leads us to anticipate that it will become increasingly feasible to provide more personalized learning experiences. Since NLP techniques can also be used to create models of individual learners based on their interactions with CALL systems and performance on language tasks, in combination with LC-based research, such data can be used to inform instructional strategies and resource selection tailored to each learner. Although the overall results of using NLP tools on learner corpora are quite encouraging, we echo Kyle (2021a: 13) in his plea that "[i]t is hoped that future researchers will continue to explore the affordances of various NLP tools and their weaknesses in a variety of contexts" in LC-based research.

In the field of language testing and assessment, we have also seen many promising advances, especially, again, in connection with NLP technology. If automatic assessment algorithms become better and eventually reach satisfying accuracy, the role played by LC might well change from giving direct input to LTA to mainly being used to train NLP tools, although there is currently a dearth of large annotated L2 corpora (Kyle, 2021a), especially for languages other than English. While for some areas automatic scoring works quite well (e.g. pronunciation), there are other pedagogically relevant areas that are notoriously difficult to assess automatically, such as quality of topic development, pragmatics or stylistics, where research still has a lot of room for improvement, especially for languages other than English.

Classroom practice is clearly an area that has huge potential for future LC-based research. In nearly every study we have examined, it has become evident that there is a significant motivational factor associated with using local data, which is even more pronounced when learners work with their own data. The vast majority of the studies also highlighted the learners' favourable responses to LC-based DDL activities. The extensive array of potential DDL activities is clearly apparent from the diverse range of pedagogical approaches that are facilitated by learner corpus data. From this perspective, the studies reviewed in Section 4 serve as a highly valuable source of inspiration for teachers. However, we need to acknowledge that LC-based DDL requires a great deal of input from teachers, as Chambers (2015: 455) rightly points out:

the time and effort involved in mastering the technical skills necessary to create and analyse corpora and in creating local learner corpora and preparing exercises using the NS and NNS data [...] suggest that giving learners a positive experience can be very demanding on the teacher.

Although we see the numbers of relevant studies increasing, it must be acknowledged that there are as yet few truly conclusive results in terms of generalizability. This may be for several reasons: one being that we cannot expect the same degree of rigour in study design and statistical robustness when real-life classroom data are involved, given small classroom sizes, different objectives within a learner group in a classroom, varying proficiency levels within classrooms, difficulties of including a control group, etc. Additionally, in the context of a real-life classroom, it seems unrealistic (or even undesirable) to anticipate achieving generalizability of one's findings beyond the local setting. This appears self-evident, given that numerous studies utilize locally collected data from distinct learner groups, resulting in very group-specific learning needs, objectives, L1 backgrounds and proficiency levels.

Moving out of the everyday classroom in order to progress towards more generalizable research in LC-based classroom practice, enhanced collaborations between teachers and LC-based DDL researchers are imperative. These collaborations should entail joint undertaking of classroom-action research projects with improved designs and advanced methods. For such research-oriented contexts, we very much agree with Pérez Cañado and Díez Bedmar (2006) that future research testing the usefulness of LC-based DDL would ideally include experimental and control groups, follow a truly longitudinal set-up, conduct pre- and post-tests (including delayed post-tests), and even be paired with qualitative questionnaires. Moreover, for a quantitative assessment of the advantages of LC-based pedagogy, empirical studies should extend over a longer period. This is essential because challenging linguistic features typically require more than a few learning sessions for acquisition by learners (Luo & Zhou, 2017; Ortega, 2003; Storch, 2009). Also, as Novikov and Vinokurova (2022: 61) critically remark: “some of the linguistic features that are challenging for students might not be appropriate or possible to address with DDL”. Future research will therefore have to differentiate clearly between the linguistic features that can be beneficially taught and learned using one of the following teaching methods: DDL using learner corpora, DDL using native corpora, DDL using both native and learner corpora and ‘traditional’ teaching (cf. Moon & Oh, 2017).

If more rigorous research turns out to show that learners benefit from learning with LC-based DDL, the teaching of corpus didactics – i.e. the techniques of how best to teach and learn with (learner) corpora – needs to become obligatory in university teacher education in order to empower teachers to be more confident in using this (seemingly difficult) tool in their everyday teaching.

## **6. Outline of the special issue**

As outlined in this introduction, the scope of LC-based applications in language pedagogy is vast. Our survey of the state-of-the-art in the field reveals that the most significant research gap and the highest potential for research lies within LC-based classroom practice. This prompted us to devote the special issue to original research studies that examine the integration of learner corpora into classroom settings and empirically assess whether they offer advantages over other teaching materials and assessment tools. Consistent with Crosthwaite's (2012) recommendation that “[r]ather than wasting such a golden opportunity (...) we should make the call for teachers and researchers alike to

utilize the technology”, we have gathered several studies that empirically investigate the potential pedagogical advantages and limitations of incorporating learner corpora. All the studies featured in our special issue are thoroughly researched, employing suitable analytical methods and, where feasible, statistical measures. Furthermore, they offer concrete applications for language pedagogy, mainly encompassing the field of language teaching. One of the studies featured in this special issue focuses on concrete applications of LC-based research to testing and assessment. The papers cover a diverse range of L1s and L2s, as well as various modes and genres, thereby making this special issue both multifaceted and cohesive in its approach.

In the initial paper, Pauliina Peltonen and Pekka Lintunen present the findings of a teaching experiment where they employed learner-corpus-based awareness-raising activities to assess their impact on L2 fluency development in an L2 English pronunciation course at a university. The study involved eighteen Finnish university students of English, who were divided into experimental and control groups. Both groups provided speech samples at the beginning and end of the L2 English pronunciation course, which were subsequently analysed using a mixed methods approach. While no significant quantitative differences between the groups were found, qualitative analyses revealed some positive effects on individual learning trajectories concerning stalling mechanisms.

In the second study, conducted by Virginie André, Alex Boulton, Maud Ciekanski and Clara Cousinard, the researchers delve into the utility of L2 corpora and multimodal corpora for DDL using the FLEURON online platform. Their investigation focuses on the interaction of nineteen international learners of French as a foreign language with the platform in a DDL approach at the University of Lorraine. The analysis underscores how L2 corpora in particular can enhance learners’ awareness of intricate phenomena related to conversational narratives by stimulating their meta-cognitive strategies during their time abroad. Additionally, the study unveils the potential for integrating an L2 component into the array of resources available for teaching and learning spoken interaction.

The next study, by Cédric Sarré, Cédric Bruderemann and Muriel Grosbois, examines the varying effects of different noticing activities on the development of grammatical accuracy in written productions of EFL learners. Leveraging an error-tagged learner corpus, the authors explore the influence of diverse types of noticing activities through practical experiments utilizing learner corpus data. Employing a pre-test/post-test quasi-experimental design with three experimental groups receiving different treatments and one control group, they analyse error frequencies at both group and individual level, together with the proportions of learners producing errors on three specific error types (articles, verb tense, verb agreement). Their findings indicate that accuracy in the use of articles and verb agreement is more readily enhanced through the comparison of learner output with native data, while for verb tenses, the effectiveness of a more conventional form of corrective feedback appears to be greater. Conversely, the utilization of online machine translation tools does not seem to positively impact learners’ accuracy development.

The fourth study, conducted by Shelley Staples, Nina Conrad, Anh Dang and Hui Wang, delves into the influence of corpus-based teaching in a series of second language (L2) genre-based writing courses. The study focuses on activities implemented in one L2

writing course between the initial and final drafts of two major assignments (a Literacy Narrative and a Genre Analysis). The materials were developed using the *Corpus and Repository of Writing* to complement an existing genre-based curriculum. Their findings, which indicate modest positive changes among students who engaged with the corpus-based materials, provide a foundation for future investigations in this area.

In the final study, Dana Gablasova, Luke Harding, Vaclav Brezina and Jamie Dunlea investigated the occurrence of adverbial, adjectival and verbal epistemic stance markers across proficiency levels and task types in the *British Council-Lancaster Aptis Corpus*. Their aim was to examine these patterns in computer-mediated speaking tests without live human interlocutors. Their findings revealed that epistemic stance was prevalent in the output of test takers and was influenced by both proficiency level and task type. These results indicate that computer-mediated speaking tests prompt the use of stance expressions by test-takers in a manner comparable to tests with human assessors. The study carries pertinent implications for examiner training and test preparation.

The special issue concludes with a position paper presented by Luciana Forti. She argues that proficiency-rated learner corpora can provide valuable linguistic data to suit a range of proficiency levels, especially when the individual texts of a corpus are proficiency-rated. In support of this, she demonstrates how such learner corpora can provide a model of target language through error-free learner productions, as well as erroneous productions, which can provide a basis for enhancing language knowledge and awareness.

With this introductory article, our aim is not only to pique the interest of researchers and language practitioners in exploring the subject further, but also to inspire numerous researchers to engage in more research collaborations with teachers within the fascinating realm of LC-based language pedagogy. The studies collected in this special issue will hopefully make an initial valuable contribution to laying the groundwork for this joint (ad)venture.

## References

- Ackerley, K. (2013). A comparison of learner and native speaker writing in online self-presentations: Pedagogical applications. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of Learner Corpus Research: Looking back, moving ahead* (pp. 1–10). Presses universitaires de Louvain.
- Ackerley, K. (2015). Short-term effects of students' exploration of corpora: A longitudinal study of pre- and post-modification of noun phrases in learner English. In E. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in learner corpus linguistics: Research and applications for Foreign Language Teaching and assessment* (pp. 199–218). Peter Lang.
- Aston, G. (2008). It's only human.... In A. Martelli, & V. Pulcini (Eds.), *Investigating English with corpora: Studies in honour of Maria Teresa Prat* (pp. 343–354). Monza.
- Attali, Y., & Burstein, J. (2006). Automatic essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment*, 4(3).  
<https://files.eric.ed.gov/fulltext/EJ1110978.pdf>

- Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of Learner Corpus Research* (pp. 511–533). Cambridge University Press.
- Belz, J., & Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning & Technology*, 12(3), 33–52.
- Bestgen, Y., & Granger, S. (2011). Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2/3), 235–252.
- Boulton, A. (2012). Hands-on /hands-off: Alternative approaches to data-driven learning. In J. Thomas & A. Boulton (Eds.), *Input, process, and product: Developments in teaching and language corpora* (pp. 152–168). Masaryk University Press.
- Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology*, 25(3), 66–89.
- Callies, M., & Götz, S. (Eds.). (2015a). *Learner corpora in language testing and assessment*. John Benjamins. <https://doi.org/10.1075/scl.70>
- Callies, M., & Götz, S. (2015b). Learner corpora in language testing and assessment: Prospects and challenges. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 1–9). John Benjamins. <https://doi.org/10.1075/scl.70.001int>
- Callies, M., Díez-Bedmar, B., & Zaytseva, E. (2014). Using learner corpora for testing and assessment: Current practice and future challenges. In E. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in Learner Corpus Linguistics: Research and applications for foreign language teaching and assessment* (pp. 21–35). Peter Lang.
- Campillos Llanos, L. (2012). Designing a search interface for a Spanish learner oral corpus: The end-user's evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey (pp. 241–248). [http://www.lrec-conf.org/proceedings/lrec2012/pdf/574\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/574_Paper.pdf)
- Carter, R., & McCarthy, M. (2006). *Cambridge Grammar Today: A comprehensive guide*. Cambridge University Press.
- Carter, R., McCarthy, M., Mark, G., & O'Keeffe, A. (2011). *The English Grammar Today*. Cambridge University Press.
- Cambridge Learner Corpus: A typology of exercises. (2011). <https://www.cambridge.org/sketch/help/about/Using%20the%20Learner%20Corpus%20to%20create%20exercises.pdf>
- Cambridge Advanced Learner's Dictionary* (4<sup>th</sup> ed.). (2013). Cambridge University Press.
- Chambers, A. (2015). The learner corpus as a pedagogic corpus. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge handbook of Learner Corpus Research* (pp. 445–464). Cambridge University Press.
- Charles, M. (2018). Corpus-assisted editing for doctoral students: More than just concordancing. *Journal of English for Academic Purposes*, 36, 15–25. <https://doi.org/10.1016/j.jeap.2018.08.003>

- Chuang, F.-Y., & Nesi, H. (2007). Grammar Talk: Developing computer-based materials for Chinese EAP students. In O. Alexander (Ed.), *New approaches to materials development for language learning* (pp. 315–330). Peter Lang.
- Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2), 202–224.
- Council of Europe (2017). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Cowan, R., Choi, E. E., & Kim, D. H. (2003). Four questions for error diagnosis and correction in CALL. *CALICO Journal*, 20(3), 451–463.
- Cowan, R., Choo, J., & Lee, G. S. (2014). ICALL for improving Korean L2 writers' ability to edit grammatical errors. *Language Learning and Technology*, 18(3), 193–207.
- Crosthwaite, P. (2012). Learner corpus linguistics in the EFL classroom. *PASAA Journal*, 44, 133–147.
- De Cock, S., & Granger, S. (2005). Computer learner corpora and monolingual learners' dictionaries: The perfect match. *Lexicographica*, 20, 72–86.
- Díez-Bedmar, B. (2015). Article use and criterial features in Spanish EFL writing: A pilot study from CEFR A2 to B2 levels. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 163–190). John Benjamins. <https://doi.org/10.1075/scl.70.07die>
- Dumont, A. (2017). The contribution of learner corpora to the substantiation of fluency levels. In P. De Haan, R. D. Vries, & S. Van Vuuren (Eds.), *Language, learners and levels: Progression and variation* (pp. 281–308). Presses universitaires de Louvain.
- Falk, S., & Götz, S. (2016). Interactivity in mobile learning applications: A case study based on Duolingo. In T. Zeyer, S. Stuhlmann, & R. D. Jones (Eds.), *Interaktivität beim Fremdsprachenlehren und -lernen mit digitalen Medien. Hit oder Hype?* (pp. 237–258). Narr.
- Feak, C. B., Reinhard, S. M., & Rohlck, T. N. (2009). *Academic interactions: Communicating on campus*. Michigan University Press.
- Flowerdew, L. (2012). *Corpora and language education*. Palgrave Macmillan.
- Forti, L. (2023a). *The learner corpus advantage in data-driven learning activity design*. Poster presented at EUROCALL 2023: CALL for all languages, 15-18 August 2023, Reykjavik, University of Iceland.
- Forti, L. (2023b). *Learner corpora and the design of data-driven learning activities*. EUROCALL 2023: CALL for all Languages, 15-18 August 2023, Reykjavik, University of Iceland.
- Fuster-Márquez, M., & Gregori-Signes, C. (2018). Learning from learners: A non-standard direct approach to the teaching of writing skills in EFL in a university context. *Innovation in Language Learning and Teaching*, 12(2), 164–176.
- Gablasova, D., & Bottini, R. (2022). Spoken learner corpora for language teaching. In R. Jablonkai, & E. Csomay (Eds.), *Routledge handbook of corpora in English Language Teaching and learning* (pp. 296–310). Routledge.
- Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, 32(3), 301–319.
- 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. (2022). Written learner corpora to inform teaching. In R. R. Jablonkai, & E. Csomay (Eds.), *Routledge handbook of corpora in English Language Teaching and Learning* (pp. 281–295). London: Routledge.
- Götz, S. (2015). Tense and aspect errors in spoken learner language: Implications for language testing and assessment. In M. Callies, & S. Götz (Eds.), *Learner corpora in language testing and assessment* (pp. 191–215). John Benjamins. <https://doi.org/10.1075/scl.70.08got>
- Götz, S. (2019). Filled pauses across proficiency levels, L1s and learning context variables: A multivariate exploration of the Trinity Lancaster Corpus Sample. *International Journal of Learner Corpus Research*, 5(2), 159–180.
- Götz, S. (2022). Learner corpora to inform testing and assessment. In R. R. Jablonkai, & E. Csomay (Eds.), *The Routledge handbook of corpora and English Language Teaching and learning* (pp. 311–326). Routledge.
- Götz, S., & Mukherjee, J. (Eds.). (2019): *Learner corpora and language teaching*. John Benjamins.
- Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145–160). Oxford university press.
- Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & J. Petch-Tyson (Eds.), *Computer learner corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 3–33). John Benjamins.
- Granger, S. (2003). The *International Corpus of Learner English*: A new resource for foreign language learning and teaching and Second Language Acquisition research. *TESOL Quarterly*, 37(3), 538. <https://doi.org/10.2307/3588404>
- Granger, S. (2009). The contribution of learner corpora to Second Language Acquisition and Foreign Language Teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13–33). Amsterdam: John Benjamins.
- Granger, S. (2013). The passive in learner English: Corpus insights and implications for pedagogical grammar. In I. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (pp. 5–15). Kobe University: Kobe.
- Granger, S. (2015). 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. 485–510). Cambridge University Press.
- Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English*. Version 3. Presses universitaires de Louvain.
- Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL*, 19(3), 252–268. <https://doi.org/10.1017/S0958344007000237>
- Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199–209). Addison Wesley Longman.
- Gyllstad, H., & Snoder, P. (2021). Exploring learner corpus data for language testing and assessment purposes: The case of verb + noun collocations. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora* (pp. 49–71). Blue Ridge Summit: Multilingual Matters. <https://doi.org/10.21832/9781788924863-004>

- Han, N.-R., Tetreault, J. Lee, S. H., & Ha, J.Y. (2010). Using an error-annotated learner corpus to develop an ESL/EFL error correction system. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Malta, (pp. 763–770). [http://www.lrec-conf.org/proceedings/lrec2010/pdf/821\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/821_Paper.pdf)
- Hargreaves, P. (2000). How important is collocation in testing the learner's language proficiency? In M. Lewis (Ed.), *Teaching collocation – Further developments in the lexical approach* (pp. 205–223). Language Teaching Publications.
- Hartle, S. (2023). From learner corpus to data-driven learning (DDL): Improving lexical usage in academic writing. *EuroAmerican Journal of Applied Linguistics and Languages*, 10(2), 9–31.
- Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(1), 1–23. <https://doi.org/10.1017/S204153621000010>
- Hawkins, J. A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge University Press.
- Hegelheimer, V. (2006). Helping ESL writers through a multimodal, corpus-based, online Grammar Resource. *CALICO Journal*, 24(1), 5–31. <https://doi.org/10.1558/cj.v24i1.5-31>
- Hegelheimer, V., & Fisher, D. (2006). Grammar, writing, and technology: A sample technology-supported approach to teaching grammar and improving writing for ESL learners. *CALICO Journal*, 23(2), 257–279. <https://doi.org/10.1558/cj.v23i2.257-279>
- Heift, T. (2001). Error-specific and individualized feedback in a web-based language tutoring system: Do they read it? *ReCALL*, 13(2), 129–142.
- Heift, T., & Caws, C. (2013). Recycling learner data to construct data-driven learning tools: Practical applications for the learner, the instructor, and the teacher. In *Proceedings of the 6<sup>th</sup> edition of the ICT for Language Learning Conference*. Florence, Italy. [https://conference.pixel-online.net/conferences/ICT4LL2013/common/download/Paper\\_pdf/009-ITL02-FP-Heift-ICT2013.pdf](https://conference.pixel-online.net/conferences/ICT4LL2013/common/download/Paper_pdf/009-ITL02-FP-Heift-ICT2013.pdf)
- Higgins, D., Ramineni, C., & Zechner, K. (2015). Learner corpora and automatic scoring. In S. Granger, G. Gilquin, & F. Meunier (Eds.) *The Cambridge handbook of Learner Corpus Research* (pp. 587–604). Cambridge University Press.
- Hsieh, W. M., & Liou, H. C. (2008). A case study of corpus-informed online academic writing for EFL graduate students. *CALICO Journal*, 26(1), 28–47.
- Hulstijn, J. H. (2010). Linking L2 proficiency to L2 acquisition: Opportunities and challenges of profiling research. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research*. (Eurosia monographs series; No. 1). European Second Language Association (pp. 233–238). <http://eurosia.org/monographs/EM01/233-238Hulstijn.pdf>
- Jordan, G. (2015). Geoff Jordan vs. Duolingo. *ELT Jam*. <https://learnjam.com/geoff-jordan-vs-duolingo/>
- Klimova, M. A., Smilga, V. K., & Overnikova, D. A. (2021). Using an error-annotated learner corpus (REALEC) in DDL lessons. *Proceedings of the International Conference Corpus Linguistics 2021* (pp. 112–121), St. Petersburg.

[https://www.hse.ru/data/2021/09/22/1474091483/kli%20Corpora-2021\\_paper\\_23%20\(1\).pdf](https://www.hse.ru/data/2021/09/22/1474091483/kli%20Corpora-2021_paper_23%20(1).pdf)

- Kyle, K. (2021a). Introduction: Natural language processing for Learner Corpus Research. *International Journal of Learner Corpus Research*, 7(1), 1–16.
- Kyle, K. (Ed.). (2021b). *International Journal of Learner Corpus Research*, 7(1): *Special issue on Natural Language Processing for Learner Corpus Research*. John Benjamins.
- Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis, & J. Burnstein (Eds.), *Automated essay scoring: A Cross-disciplinary perspective* (pp. 87–112). Erlbaum.
- Leacock, C., Chodrow, M., & Tetrault, J. (2015). Automatic grammar- and spell-checking for language learners. In S. Granger, G. Gilquin, & F. Meunier (Eds.) *The Cambridge handbook of Learner Corpus Research* (pp. 567–586). Cambridge University Press.
- Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1-23). Longman.
- Leńko-Szymańska, A. (2015) The English Vocabulary Profile as a benchmark for assigning levels to learner corpus data. In M. Callies, & S. Götz (Eds.) *Learner corpora in Language Testing and Assessment* (pp. 115–140). John Benjamins.
- Llorente, R., Morant, M., & Garrigos-Simon, F. J. (2012). Crowdlearning, applying social collaboration in higher education. *Proceedings of new perspectives in science education*. <http://conference.pixel-online.net/NPSE/files/npse/ed0004/FP/1550-SERA935-FP-NPSE4.pdf>
- Lo, Y.-C., Chen, J.-J., Yang, C.-Y., & Chang, J. S. (2018). Cool English: A grammatical error correction system based on large learner corpora. In *System Demonstrations: Proceedings from the 27<sup>th</sup> international conference on computational linguistics* (pp.82-85). Santa Fe <https://aclanthology.org/C18-2018.pdf>
- Longman Dictionary of Contemporary English*. Sixth edition (2014). Longman.
- Lorenz, G. R. (1999). *Adjective intensification – Learners versus native speakers: A corpus study of argumentative writing*. Rodopi.
- Luo, Q., & Zhou, J. (2017). Data-driven learning in second language writing class: A survey of empirical studies. *International Journal of Emerging Technologies in Learning*, 12(3), 182–196. <https://doi.org/10.3991/ijet.v12i03.6523>
- Lusta, A., Demirel , O., & Mohammadzadeh, B. (2023). Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Heliyon*, 9, e22731. <https://www.sciencedirect.com/science/article/pii/S2405844023099395/pdf?md5=8c1dadf8f06fa539105abcf8b6774aff&pid=1-s2.0-S2405844023099395-main.pdf>
- Macmillan English dictionary for advanced learners*. Second edition. (2007). Macmillan Education.
- Marchand, T., & Akutsu, S. (2015). First steps in assigning proficiency to texts in a learner corpus of computer-mediated communication. In M. Callies, & S. Götz (Eds.), *Learner corpora in Language Testing and Assessment* (pp. 85–112). John Benjamins. <https://doi.org/10.1075/sci.70.04mar>
- Martinez-Garcia, M. T., & Wulff, S. (2012). Not wrong, yet not quit right: Spanish ESL students’ use of gerundial and infinitival complementation. *International Journal of Applied Linguistics*, 22(2), 225–244.

- McCarthy, M. (2014). Putting the CEFR to good use: Designing grammars based on learner-corpus evidence. *Language Teaching*, 49(1), 99–115. <http://dx.doi.org/10.1017/S0261444813000189>
- Mendikoetxea, A., Murcia-Biels, S., & Rollinson, P. (2010). Focus on errors: Learner corpora and pedagogical tools. In M. Campoy-Cubillo, B. Bellés-Fortuño, & L. Gea-Valor (Eds.), *Corpus-based approaches to English Language Teaching* (pp. 180–194). Continuum.
- Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge handbook of Learner Corpus Research* (pp. 537–566). Cambridge University Press.
- Millar, N., & Lehtinen, B. (2008). DIY local learner corpora: Bridging gaps between theory and practice, *The JALT CALL Journal*, 4(2), 61–72.
- Milton, J. (2006). Resource-rich web-based feedback: Helping learners become independent writers. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 123–137). Cambridge University Press.
- Mitton, R., & Okada, T. (2007). The adaptation of an English spellchecker for Japanese writers. *Symposium for Second Language Writing*. Nagoya, Japan. <https://eprints.bbk.ac.uk/id/eprint/592/3/592.pdf>
- Moon, S., & Oh, S. (2018). Unlearning overgenerated *BE* through data-driven learning in the secondary EFL classroom. *ReCALL*, 30(1), 48–67.
- Mukherjee, J., & Götz, S. (2015). Learner corpora and learning context. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge handbook of Learner Corpus Research* (pp. 423–442). Cambridge University Press.
- Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E. D. Galaczi, & C. J. Weir (Eds.), *Exploring language frameworks* (pp. 137–165). Cambridge University Press.
- Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–52). John Benjamins.
- Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference* (pp. 572–581). UCRE Technical Paper 16. Lancaster University, <https://ucrel.lancs.ac.uk/publications/cl2003/papers/nicholls.pdf>
- Novikov, A., & Vinokurova, V. (2022). Learner corpus as a medium for tasks. In S. V. Nuss, & M. W. Whitehead (Eds.), *Task-based instruction for teaching Russian as a foreign language* (pp. 47–64). Routledge.
- O'Dell, F. (2005). How the *Cambridge Learner Corpus* helps with materials writing. *Humanising Language Teaching*, 7(1).
- Orenha-Ottaiano, A. (2012). English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. *Acta Scientiarum. Language and Culture Maringá*, 34(2), 241–251.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518.
- Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis, & J. Burnstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Erlbaum.

- Palacios-Martínez, I., & Martínez-Insua, A. (2006). Connecting linguistic description and language teaching: Native and learner use of existential *there*. *International Journal of Applied Linguistics*, 16(2), 213–231.
- Pérez Cañado, M. L., & Díez Bedmar, M. B. (2006). Data-driven learning and awareness-raising: An effective tandem to improve grammar in written composition? *International Journal of Emerging Technologies in Learning (iJET)*, 1(3) <https://doi.org/10.3991/ijet.v1i3.55>
- Pérez-Paredes, P. (2003). Integrating networked learner oral corpora into foreign language instruction. In S. Granger, & S. Petch-Tyson (Eds.), *Extending the scope of corpus-based research: New applications, new challenges* (pp. 249–261). Rodopi.
- Rankin, T., & Schiftnr, B. (2011). Marginal prepositions in learner English: Applying local corpus data. *International Journal of Corpus Linguistics*, 16(3), 412–434.
- Rimrott, A., & Heift, T. (2008). Evaluating automatic detection of misspellings in German. *Language Learning and Technology*, 12(3), 73–92.
- Rozovskaya, A., & Roth, D. (2010). Training paradigms for correcting errors in grammar and usage in human language technologies. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 154–162). Los Angeles: The Association for Computational Linguistics.
- Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric<sup>SM</sup> essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4). <https://files.eric.ed.gov/fulltext/EJ843850.pdf>
- Salamoura, A. (2008). Aligning English Profile research data to the CEFR. *Cambridge ESOL: Research Notes*, 33, 5–7.
- Samoudi, N., & Modirghamene, S. (2022). Concordancing in writing pedagogy and CAF measures of writing. *International Review of Applied Linguistics in Language Teaching*, 60(3), 699–722.
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129–158.
- Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13, 206–226.
- Schmidt, T., & Strasser, T. (2022). Artificial intelligence in foreign language learning and teaching. *Anglistik*, 33, 165–184. <https://doi.org/10.33675/ANGL/2022/1/14>
- Schweinberger, M. (2020). How Learner Corpus Research can inform language learning and teaching: An analysis of adjective amplification among L1 and L2 English speakers. *Australian Review of Applied Linguistics*, 43(2), 196–218.
- Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 213–234). John Benjamins.
- Staples, S., Dang, A., & Wang, H. (2023). Learner corpora in corpus-informed instruction: Moving toward an asset- and genre-based model. *TESOL Quarterly*, early view. <https://onlinelibrary.wiley.com/doi/epdf/10.1002/tesq.3293> <https://doi.org/10.1002/tesq.3293>
- Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, 18(2), 103–118.
- Swan, M. (2005). *Practical English usage*. 3rd edition. Oxford University Press.

- Tetrault, J., Chodrow, M., & Madnani, N. (2014). Bucking the trend: Improved evaluation and annotation practices for ESL error detection systems. *Language Resources and Evaluation*, 48(1), 5–31.
- Thewissen, J. (2012). *Accuracy across proficiency levels: Insights from an error-tagged EFL learner corpus* [Unpublished doctoral dissertation]. Université catholique de Louvain.
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(1), 77–101. <https://doi.org/10.1111/j.1540-4781.2012.01422.x>
- Todd, R. W. (2001). Induction from self-selected concordances and self-correction. *System*, 29(1), 91–102.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.
- Tribble, C., & Jones, G. (1997[1990]). *Concordances in the classroom*. Athelstand Publications.
- Umezawa, J., Mizuno, J., Okazaki, N., & Inui, K. (2013). Evidence in automatic error correction improves learners' English skill. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing. CICLing 2013. Lecture notes in computer science*, vol 7817. Springer. [https://doi.org/10.1007/978-3-642-37256-8\\_46](https://doi.org/10.1007/978-3-642-37256-8_46)
- Usami, H. (2013). Using a learner corpus to improve distractors in multiple choice grammar questions. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of Learner Corpus Research: Looking back, moving ahead. Proceedings of the first learner corpus research conference (LCR 2011)* (pp. 455–462). Presses universitaires de Louvain.
- von Ahn, L. (2013). Duolingo: learn a language for free while helping to translate the web. In J. Kim, J. Nichols, & P. Szekely (Eds.), *Proceedings of the 2013 international conference on intelligent user interfaces* (pp. 1–2). NCA.
- Walsh, S. (2010). What features of spoken and written corpora can be exploited in creating language teaching materials and syllabuses? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 333–344). Routledge.
- Wang, Y.-B., & Lee, L.-S. (2012). Improved approaches of modelling and detecting error patterns with empirical analysis for computer-aided pronunciation training. In *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5049–5052), Kyoto, Japan.
- Wanner, L., Ramos, M. A., Vincze, L., Nazar, R., Ferraro, G. Mosqueira, E., & Prieto, S. (2013). Annotation of collocations in a learner corpus for building a learning environment. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead* (pp. 493–503). Presses universitaires de Louvain.
- Wisniewski, K. (2017). The empirical validity of the Common European Framework of Reference scales. An exemplary study for the vocabulary and fluency scales in a language testing context. *Applied Linguistics*, 39(6), 933–959. <https://doi.org/eres.qnl.qa/10.1093/applin/amw057>
- Yannakoudakis, H., Briscoe, D., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 180–189), Portland, Oregon. <https://aclanthology.org/P11-1019.pdf>

- Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology*, 18(1), 96–117.
- Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10), 883–895.
- Zaki, M. (2021). Self-correction through corpus-based tasks: Improving writing skills of Arabic learners. *International Journal of Applied Linguistics*, 31(2), 193–210.

#### **Appendix: Select list of publications on LC-based classroom practice**

1. Ackerley, K. (2013). A comparison of learner and native speaker writing in online self-presentations: Pedagogical applications. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research: Looking back, moving ahead* (pp. 1–10). Presses universitaires de Louvain.
2. Ackerley, K. (2015). Short-term effects of students' exploration of corpora: A longitudinal study of pre- and post-modification of noun phrases in learner English. In Castello, E., Ackerley, K. & Coccetta, F. (Eds.), *Studies in learner corpus linguistics: Research and applications for Foreign Language Teaching and assessment* (pp. 199–218). Peter Lang.
3. Belz, J. A., & Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction, *Language Learning and Technology*, 12(3), 33–52.
4. Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, 19(3), 307–328.
5. Charles, M. (2018). Corpus-assisted editing for doctoral students: More than just concordancing. *Journal of English for Academic Purposes*, 36, 15–25. <https://doi.org/10.1016/j.jeap.2018.08.003>
6. Chitez, M., & Bercuci, L. (2019). Data-driven learning in ESP university settings in Romania: Multiple corpus consultation approaches for academic writing support. In F. Meunier, J. Van de Vyver, L. Bradley, & S. Thouësny (Eds.), *CALL and complexity – Short papers from EUROCALL 2019* (pp. 75–81). <https://doi.org/10.14705/2019.38.989>
7. Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, 26(2), 202–224. <https://doi.org/10.1017/S0958344014000019>
8. Crosthwaite, P., & Steeples, P. (2022). Data-driven learning with younger learners: Exploring corpus-assisted development of the passive voice for science writing with female secondary school students. *Computer Assisted Language Learning*. 10.1080/09588221.2022.2068615
9. Eriksson, A. (2012). Pedagogical perspectives on bundles: Teaching bundles to doctoral students of biochemistry. In J. Thomas, & A. Boulton (Eds.), *Input, process and products. Developments in teaching and language corpora* (pp. 195–211). Masaryk University Press.
10. Forti, L. (2017) Data-driven learning and the acquisition of Italian collocations: From design to student evaluation. In K. Borthwick, L. Bradley, & S. Thouësny (Eds.), *CALL in a climate of change: Adapting to turbulent global conditions – short papers from EUROCALL 2017* (pp.110–115). Research-publishing.net. <https://files.eric.ed.gov/fulltext/ED578177.pdf>

11. Fuster-Márquez, M., & Gregori-Signes, C. (2018). Learning from learners: A non-standard direct approach to the teaching of writing skills in EFL in a university context. *Innovation in Language Learning and Teaching*, 12(2), 164–176.
12. Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors?. *System*, 32(3), 301–319.
13. Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199–209). Addison Wesley Longman.
14. Gut, U. (2006). Learner speech corpora in language teaching. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (p. 69–86). Peter Lang.
15. Hartle, S. (2023). From learner corpus to data-driven learning (DDL): Improving lexical usage in academic writing. *Euro-American Journal of Applied Linguistics and Languages*, 10(2), 9–31.
16. Hsieh, W. M., & Liou, H. C. (2008). A case study of corpus-informed online academic writing for EFL graduate students. *CALICO Journal*, 26(1), 28–47.
17. Klimova, M. A., Smilga, V. K., & Overnikova, D. A. (2021). Using an error-annotated learner corpus (REALEC) in DDL lessons. *Proceedings of the International Conference Corpus Linguistics 2021* (pp. 112–121), July 1–3, 2021, St. Petersburg.
18. Mendikoetxea, A., Murcia-Biels, S., & Rollinson, P. (2010). Focus on errors: Learner corpora and pedagogical tools. In M. C. Campoy-Cubillo, B. Bellés-Fortuño, & L. Gea-Valor (Eds.), *Corpus-based approaches to English Language Teaching* (pp. 180–194). Continuum.
19. Millar, N., & Lehtinen, B. (2008). DIY local learner corpora: Bridging gaps between theory and practice', *The JALT CALL Journal*, 4(2), 61–72.
20. Moon, S., & Oh, S. (2018). Unlearning overgenerated *BE* through data-driven learning in the secondary EFL classroom. *ReCALL*, 30(1), 48–67.
21. Mukherjee, J., & Rohrbach, J.-M. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann, & G. Marko (Eds.), *Planing, gluing and painting corpora: Inside the applied corpus linguist's workshop* (pp. 205–232). Peter Lang.
22. Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–152). John Benjamins.
23. Novikov, A., & Vinokurova, V. (2022). Learner corpus as a medium for tasks. In S.V. Nuss, & W. Whitehead Martelle (Eds.), *Task-based instruction for teaching Russian as a Foreign Language* (pp. 47–64). Routledge.
24. Orenha-Ottaiano, A. (2012), English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. *Acta Scientiarum. Language and Culture Maringá*, 34(2), 241–251.
25. Papp, S. (2007). Inductive learning and self-correction with the use of learner and reference corpora. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 207–220). Rodopi.
26. Pérez Cañado, M. L., & Díez Bedmar, M. B. (2006). Data-driven learning and awareness-raising: An effective tandem to improve grammar in written

- composition? *iJET International Journal of Emerging Technologies in Learning*, 1(3). <https://doi.org/10.3991/ijet.v1i3.55>
27. Rankin, T., & Schiffner, B. (2011). Marginal prepositions in learner English: Applying local corpus data, *International Journal of Corpus Linguistics*, 16(3), 412–434.
  28. Samoudi, N., & Modirghamene, S. (2022). Concordancing in writing pedagogy and CAF measures of writing. *International Review of Applied Linguistics in Language Teaching*, 60(3), 699–722.
  29. Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 213–234). John Benjamins.
  30. Todd, R. W. (2001). Induction from self-selected concordances and self-correction. *System*, 29(1), 91–102.
  31. Tribble, C., & Jones, G. (1997[1990]). *Concordances in the classroom*. Athelstand Publications.
  32. Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology*, 18(1), 96–117.
  33. Zaki, M. (2021). Self-correction through corpus-based tasks: Improving writing skills of Arabic learners. *International Journal of Applied Linguistics*, 31(2), 193–210.