Learner Corpora in Foreign Language Education

Sylviane Granger

Abstract

Analyzing learner language is a key component of second and foreign language education research and serves two main purposes: it helps researchers gain a better understanding of the mechanisms of second language acquisition (SLA) and it is a useful source of data for practitioners who are keen to design teaching and learning tools that target learners' attested difficulties. The learner corpus (LC) is a new resource that is currently bringing learner language back into focus and is enjoying growing interest from the language education community at large. It first emerged as a branch of corpus linguistics in the late 1980s but is only now beginning to attract significant attention from L2 theoreticians and practitioners. This chapter aims to highlight the relevance of learner corpora to the field of language education. The next section gives an overview of the main defining features of this new resource and some of the dimensions along which they can be classified. The section "Work in Progress" is devoted to methods of analysis: contrastive interlanguage analysis and automated analysis. "Problems and Difficulties: Pedagogical Applications" presents some of the main pedagogical applications of learner corpus research, and the final section suggests some possible avenues for future research.

Keywords

Corpus linguistics • Learner corpus • Learner corpora • Second language acquisition • Contrastive interlanguage analysis • Automated analysis

This chapter is an updated version of that included in the 2008 edition of the encyclopedia.

S. Granger (✉)
Université catholique de Louvain, Louvain-La-Neuve, Belgium
e-mail: sylviane.granger@uclouvain.be

S.L. Thorne, S. May (eds.), *Language, Education and Technology*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02237-6_33

Contents

Introduction

Analyzing learner language is a key component of second and foreign language education research and serves two main purposes: it helps researchers gain a better understanding of the mechanisms of second language acquisition (SLA) and it is a useful source of data for practitioners who are keen to design teaching and learning tools that target learners' attested difficulties.

Learner data types can be ranged along a continuum that reflects the degree of control exerted on language production. According to Ellis and Barkhuizen (2005), the less constrained types of production should be favored, since "they demonstrate how learners use the L2 when they are primarily engaged in message construction," unlike experimental data, which must be treated with circumspection, as it may contain artificial interlanguage forms. Researchers have traditionally shied away from the more natural data types, however, opting instead for experimentally elicited samples precisely because they are more constrained. This allows for tighter control of the many variables affecting learner output, thereby facilitating interpretation of the results. In addition, as it is difficult to subject a large number of learners to experimentation, SLA research has tended to be based on a relatively narrow empirical foundation, which raises questions about the generalizability of the results. Looking at the situation from a more pedagogical perspective, Mark (1998) deplores the relative lack of focus on the description of learner language, which contrasts sharply with the increased attention devoted to other aspects of mainstream language teaching, such as learner variables (motivation, learning styles, etc.) and the description of the target language.

The learner corpus (LC) is a new resource that is currently bringing learner language back into focus and is enjoying growing interest from the language education community at large. It first emerged as a branch of corpus linguistics in the late 1980s but is only now beginning to attract significant attention from L2 theoreticians and practitioners. This chapter aims to highlight the relevance of learner corpora to the field of language education. The next section gives an overview of the main defining features of this new resource and some of the dimensions along which they can be classified. Section "Work in Progress" is devoted to methods of analysis: contrastive interlanguage analysis and automated analysis. Section "Problems and Difficulties: Pedagogical Applications" presents some of the main pedagogical applications of learner corpus research, and the final section suggests some possible avenues for future research.

Major Contributions

Learner corpora are electronic collections of natural or near-natural foreign or second language learner texts assembled according to explicit design criteria. Several aspects of this definition require clarification. The term *near-natural* is used to highlight the "need for data that reflects as closely as possible 'natural' language use (i.e., language that is situationally and interactionally authentic) while recognizing that the limitations facing the collection of such data often obligate researchers to resort to clinically elicited data (for example, by using pedagogic tasks)" (Ellis and Barkhuizen 2005, p. 7). In principle, learner corpora can contain data from both *foreign language* (*FL*) *learners*, who learn a language in a country where they have little exposure outside the classroom (e.g., learning English in Germany or Japan), and *second language* (*SL*) *learners*, who acquire a language in a country where that language is the predominant language of communication (e.g., learning English in the United States). The term *texts* highlights the fact that learner corpora contain continuous stretches of oral or written discourse rather than decontextualized sentences. This makes it possible to study a much wider range of interlanguage features than in previous SLA studies, which have tended to focus on more local features like grammatical morphemes. The requirement of *explicit design criteria* stems from the necessity to control the wide range of variables that affect learner language. As can be seen in Table 1, which lists the criteria governing the collection of the *International Corpus of Learner English* (*ICLE*) (Granger et al. 2009), some of these variables pertain to the language situation or task, while others relate to the learner.

It is this requirement that makes learner corpus collection such a laborious undertaking and yet it is a crucial requirement: the usefulness of a learner corpus

Table 1 *ICLE* design criteria

Learner variables	Task variables
Age	Medium
Learning context	Field
Proficiency level	Genre
Gender	Length
Mother tongue background	Topic
Region	Timing
Knowledge of other foreign languages	Exam
Amount of L2 exposure	Use of reference tools

is directly proportional to the care that has been taken in designing it, and compromising the design stage inevitably leads to less solid results. If the variables are recorded and stored in a database, they can be used to compile homogeneous subcorpora. The interface of the *ICLE* makes it possible, for instance, to study gender differences, topic effects, the influence of timing, even to compare FL learners who have never spent any time in an English-speaking country with those who have done so for extended periods of time.

Learner corpora can be classified on the basis of the following features:

– *Target languages*: while English still has the lion's share, learner corpus collection is now active in a wide range of languages (Dutch, French, German, Italian, Norwegian, Spanish, and Swedish, inter alia) (for a survey, see the "Learner corpora around the world" webpage on the Louvain website: http://www.uclouvain.be/en-cecl-lcworld.html). Most learner corpora cover only one target language, the MERLIN corpus (Abel et al. 2013) being a notable exception in this respect. Bilingual learner corpora like the German-English *Telekorp* corpus (Belz and Vyatkina 2005) are a promising development resulting from the growing use of telecollaborative communication in language education.
– *Mother tongue backgrounds*: learner corpora can contain data from learners of one and the same mother tongue background or from several mother tongue backgrounds. The latter are necessary if the purpose of the data collection is to produce generic pedagogical tools such as monolingual learners' dictionaries (see Section "Problems and Difficulties: Pedagogical Applications"). Most academic learner corpora contain data from only one language background, for example, Japanese learners of English in the case of the *NICT JLE Corpus* (Izumi et al. 2004), Chinese learners of English for the *Chinese Learner English Corpus* (Gui and Yang 2002), or Swedish learners of French for the *Interfra Corpus* (Bartning and Schlyter 2004). The *International Corpus of Learner English*, which covers 16 different mother tongue backgrounds, is a notable exception in this regard.
– *Medium*: corpora of learner writing were the first to be collected and are still the dominant type today. The supremacy of written corpora is primarily due to the difficulty of collecting and transcribing learner oral data. In spite of this difficulty, some oral learner corpora have been compiled. These include the *College English Learners' Spoken English Corpus*, which contains data from Chinese learners of English (Yang and Wei 2005), and the *Louvain International Database of Spoken English Interlanguage,* which contains data from learners with 11 different mother tongue backgrounds (cf. Gilquin et al. 2010). A new type, the multimodal (or multimedia) learner corpus, which contains learners' texts linked to audio-video recordings, is a recent and welcome addition that enables analysts to investigate nonverbal as well as verbal aspects of communication (Reder et al. 2003; Hashimoto and Takeuchi 2012).
– *Genre*: while some genres are well represented in current learner corpora, particularly essay writing and informal interviews, many are hardly covered at all, which makes it difficult to assess the influence of task on learner production. The *NICT JLE Corpus* (Izumi et al. 2004), which comprises three types of tasks –

picture description, role-playing, and story-telling, is exceptional in this respect. The collection of large multitask learner corpora is clearly one of the major desiderata for the future.

– *Time of collection*: learner corpora can be collected at a single point in time or at successive points over a period of time. Only the latter, which are much more difficult to collect and are therefore in the minority, allow for longitudinal studies of learner language and are a rich resource for describing stages of acquisition (for L2 French, see Bartning and Schlyter 2004).
– *Pedagogical use*: corpora for delayed pedagogical use sample a given learner population and are used to produce pedagogical tools that will subsequently benefit similar-type learners. The vast majority of learner corpora collected to date have been of this type. More recently, however, learner corpus collection has begun to be integrated into normal classroom activities: learner data is collected from a given learner population to inform pedagogical activities that involve, in the first instance, those same learners, while also allowing for subsequent use with similar-type learners. Learner corpora for immediate pedagogical use thus involve learners as both producers and users of the data.

Learner corpora differ in their degree of accessibility. Many are unfortunately not available outside the arena where they have been collected. However, a growing number are available for scientific research and/or can be consulted online.

Work in Progress

Contrastive Interlanguage Analysis

A learner corpus is a solid empirical base from which to uncover the linguistic features that characterize the interlanguage of foreign and second language learners at different stages of proficiency and/or in a range of language situations. The method that has mainly been used for that purpose is contrastive interlanguage analysis (CIA) (Granger 1996, 2015a). Unlike classic contrastive analysis, which compares different languages, CIA compares varieties of one and the same language and involves the following two types of comparison:

1. Comparisons of corpora of learner language and native (or expert) reference language
2. Comparisons of corpora representing different varieties of learner language

The first plays an important role in revealing or uncovering the distinguishing features of learner language, while the second makes it possible to assess the degree of generalizability of interlanguage features across learner populations and language situations. The latter type has never come in for any criticism from SLA specialists, unlike the former, which has been criticized for being guilty of the “comparative fallacy” (Bley-Vroman 1983), i.e., for comparing learner language to a native speaker norm and thus failing to analyze interlanguage in its own right. Although it is important to stress the need to view interlanguage on its own terms, there are several arguments that can be invoked in defense of native/learner comparisons. First, the native speaker norm that is used in learner corpus studies is explicit and corpus-based (Mukherjee 2005) rather than implicit and intuition-based, as has usually been the case in SLA studies. Second, there is not just *one* reference corpus but several to choose from. In the case of English, for instance, analysts can choose between the many geographical varieties of English covered in the *International Corpus of English* (http://www.ucl.ac.uk/english-usage/projects/ice.htm), several of which are available in electronic format, or may opt for a corpus of expert L2 user data instead (Seidlhofer 2004). From a pedagogical point of view, comparisons of learner data to a native or expert reference corpus is even more obvious, as they help teachers identify the lexical, grammatical, and discourse features that differentiate learners' production from the targeted norm and may therefore be usefully integrated into the teaching program.

Automated Analysis

One important feature that distinguishes learner corpus data from traditional learner data is the fact that the texts are stored in electronic format. Once computerized, learner data can be examined with a variety of software tools that can radically change the way foreign/second language researchers set about analyzing learner language. Some degree of automation is arguably essential, as several learner corpora contain millions rather than hundreds or thousands of words. Automation contributes to a better analysis of learner language in three main ways: (1) it makes it possible to quantify learner language; (2) it helps discover interlanguage patterns of use; and (3) it makes it possible to enrich learner data with a wide range of linguistic annotations.

Frequency

One of the major contributions of automation is that it brings forth a wealth of quantitative information on learner language that had hitherto been unavailable. Text retrieval software tools like *WordSmith Tools* (*WST*) (Scott 2012) or *Antconc* (Anthony 2014) are language-independent programs that enable researchers to count and sort lexical items in text samples automatically. Using these tools, researchers have immediate access to frequency lists of all the single words or sequences of words in their corpora. One particularly useful function in *WST* allows researchers to compare these lists, highlight the significant differences between them, and draw up lists of words that display a significantly higher or lower frequency of use in learner data. This option plays an important role in identifying cases of over- and under-representation that, as already pointed by Levenston in

Table 2 Sample of significantly over- and underused lexical verbs in *ICLE*

Overused verbs	Underused verbs
Think	*Describe*
Get	*Occur*
Dream	*Note*
Want	*Suggest*
Watch	*Require*
Live	*Contain*
Ban	*Obtain*
Learn	*Identify*
Pay	*Involve*
Like	*Assume*
Go	*Derive*
Buy	*Follow*
Need	*Include*
Smoke	*Record*
Spend	*Determine*

1971, characterize learner language just as much as downright errors, especially at the more advanced proficiency levels. For example, Granger and Paquot (2009) used *WST* to compare the top 100 lexical verbs in the 3.7 million-word *ICLE corpus* of writing by higher intermediate to advanced EFL learners and a comparable native academic corpus (*ACAD*). As Table 2 indicates, the comparison shows that EFL learners tend to significantly overuse some lexical verbs and underuse others.

While some of the overused verbs are topic-dependent (e.g., *dream, ban,* or *smoke*), many are indicative of students' over-reliance on high-frequency verbs that are more typical of conversation than academic writing (e.g., *think*, *get*, or *want*). The underused verbs, however, are typical EAP verbs that merit focused pedagogical attention.

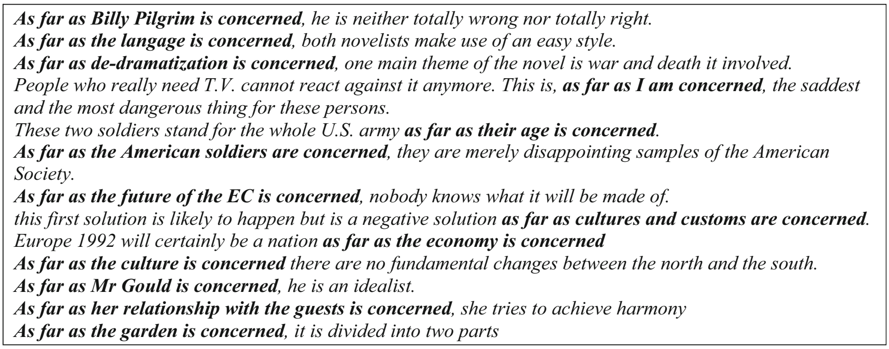
Patterns of Use

The quantitative benefits of computerized learner data should not obscure the equally impressive qualitative insights afforded by computer-aided methods. Corpus methods are very powerful heuristic devices for uncovering words' preferred lexical and grammatical company. The concordancing function in text retrieval software tools enables researchers to extract all occurrences of a given lexical item (single word or phrase) in a corpus and sort them in a variety of ways, thereby allowing typical patterns to emerge. Table 3 highlights some of the striking differences that emerge from the concordance of the word *as* in a corpus of essays written by native American-English students (*LOCNESS*) and EFL learners with Spanish, French, and German mother tongue backgrounds (*ICLE*).

While the figures reveal some degree of commonality between the three learner groups, such as the tendency to overuse *as far as* and underuse *as well as* and *as*

Table 3 Patterning of the word *as* in native and learner corpora (relative frequency per 200,000 words)

Patterning of *as*	*LOCNESS*	*ICLE*-SP	*ICLE*-FR	*ICLE*-GE
as a conclusion	0	16.3	34.5	0
as far as	6.7	14.2	95.2	34.4
as far as X is concerned	1.3	11.2	87.9	15
as well as	108.2	34.6	46	61.9
as long as	57.4	2	16.7	23.8



***As far as Billy Pilgrim is concerned**, he is neither totally wrong nor totally right.*

***As far as the langage is concerned**, both novelists make use of an easy style.*

***As far as de-dramatization is concerned**, one main theme of the novel is war and death it involved.*

*People who really need T.V. cannot react against it anymore. This is, **as far as I am concerned**, the saddest and the most dangerous thing for these persons.*

*These two soldiers stand for the whole U.S. army **as far as their age is concerned**.*

***As far as the American soldiers are concerned**, they are merely disappointing samples of the American Society.*

***As far as the future of the EC is concerned**, nobody knows what it will be made of.*

*this first solution is likely to happen but is a negative solution **as far as cultures and customs are concerned**.*

*Europe 1992 will certainly be a nation **as far as the economy is concerned***

***As far as the culture is concerned** there are no fundamental changes between the north and the south.*

***As far as Mr Gould is concerned**, he is an idealist.*

***As far as her relationship with the guests is concerned**, she tries to achieve harmony*

***As far as the garden is concerned**, it is divided into two parts*

Fig. 1 Concordance excerpt of *as far as x is concerned* in *ICLE*-FR

long as, they also highlight varying patterns of use, such as overuse of *as a conclusion* by Spanish- and French-speaking but not German-speaking learners. As evidenced by several recent studies (e.g., Paquot 2013), this variability is often the result of transfer from the learners' mother tongue. For example, the striking predilection of French-speaking learners for the phrase *as far as x is concerned*, which emerges clearly from the concordance excerpt in Fig. 1, is modeled on the French phrase *en ce qui concerne*. Most of the examples show students' difficulty in introducing topics and could serve as useful prompts for rewriting exercises.

Typical collocations, i.e., pairs of words that have a strong tendency to co-occur within a few words of each other, can be extracted fully automatically using statistical association measures. Durrant and Schmitt (2009) employ this method to highlight differences in the patterning of adjective/noun + noun combinations in learner and native writing. Clusters, i.e., recurrent contiguous sequences of two or more words, can also easily be extracted from learner corpora. Applying this method to a corpus of EFL speech and a comparable native speaker corpus, De Cock (2004) shows that EFL learners significantly underuse discourse markers like *you know* or *I mean* and vagueness markers like *sort of* or *and things* and therefore prove to be lacking in routinized ways of interacting and building rapport with their interlocutors and weaving into their speech the right amount of imprecision and vagueness, both typical features of informal interactions.

Annotation

A learner corpus can also be annotated. In corpus linguistics terms, “annotation” refers to “the practice of adding interpretative (especially linguistic) information to an existing corpus of spoken and/or written language by some kind of coding attached to, or interspersed with, the electronic representation of the language material” (Leech 1993, p. 275). In learner corpus terms, this means that any information about the learner samples that the researcher wants to code can be inserted into the text. In a learner corpus, it is therefore not only words that are contextualized but also information about the words.

Although there is, in principle, no limit to the type of annotation that can be used to enrich a learner corpus, two types are by far the most common: morpho-syntactic annotation and error annotation. Part-of-speech (POS) taggers automatically attach a tag to each word in a corpus, indicating its word-class membership. These programs are particularly useful, as they help disambiguate the many words that belong to more than one part of speech. Only a POS-tagged learner corpus would allow researchers to attribute the over- or underuse of the word *to* to differences in frequency of use of the infinitive particle *to* or the preposition *to*. It is important to bear in mind, however, that morpho-syntactic annotation programs – whether lemmatisers, POS taggers, or parsers – have been trained on the basis of native-speaker corpora, and there is no guarantee that they will perform as accurately on learner data. While the success rate of POS taggers has been found to be quite good with advanced learner data, it has proved to be very sensitive to morpho-syntactic and orthographic errors (Van Rooy and Schäfer 2003), and the success rate will therefore tend to decrease as the number of these errors increases. To counter this weakness, a number of researchers prefer to use CHILDES (MacWhinney 1999), a suite of software tools that gives them a high degree of flexibility in the annotating process. Initially designed for L1 acquisition research, it was subsequently adapted for L2 data analysis (Myles and Mitchell 2004).

Although error analysis has fallen into disfavor in SLA, it remains a crucial aspect of learner language and one that in fact still lies at the heart of many SLA studies, hidden under labels such as negative transfer, fossilization, corrective feedback, measures of linguistic accuracy, and developmental sequences. Two methods are used in learner corpus research to chart attested learner errors: computer-aided detection and error annotation. In the former, it is the analyst who chooses the linguistic items on which to focus, using his/her intuition, pedagogical experience, or previous SLA studies. Once selected, the linguistic forms can be searched automatically in the learner corpus, then counted and sorted as described in section “Patterns of Use”. The study of overpassivization errors by Cowan et al. (2003) is a good illustration of this method. The problem is that this method presupposes that one knows what errors to look for, which is far from always being the case.

The only method that can ensure comprehensive error detection is error annotation, which is enjoying growing popularity, in spite of its difficulty and time-costliness, and several systems have now been developed (for a survey, see Díaz-Negrillo and Fernández- Domínguez 2006). In most of these, the error is coded for error type (number, gender, tense, etc.), word category (noun, verb, etc.), and in some cases, error domain (spelling, grammar, lexis, etc.). When applied to a learner corpus that has been carefully compiled on the basis of strict design criteria (mother tongue background, level of proficiency, etc.), error annotation is a valuable resource that makes it possible to tailor pedagogical materials to the needs of a given learner population (cf. Granger 2003). However, error annotation will always contain an element of subjectivity, as the very notion of error is far from clear-cut. As rightly pointed out by Milton and Chowdhury (1994, p. 129), "Tagging a learner corpus allows us, at least and at most, to systematize our intuitions." To cater for errors that can have more than one interpretation, some systems allow for the inclusion of several target hypotheses (Lüdeling and Hirschmann 2015). Whatever the system used, it is essential that annotators be provided with a comprehensive error-tagging manual and undergo rigorous training. It is also important to bear in mind that error annotation is a very time-consuming, hence costly, process. Limitations in manpower and/or budget may lead researchers to tag only part of their corpus or to limit the tagging to some specific error categories (morphological errors, preposition errors, article errors, etc.).

Problems and Difficulties: Pedagogical Applications

Among the many pedagogical applications that could potentially benefit from learner-corpus-informed insights, only a few can boast a number of concrete achievements: pedagogical lexicography, courseware, and language assessment.

The field in which advances have been quickest is **pedagogical lexicography**. Monolingual learners' dictionaries like the *Macmillan English Dictionary for Advanced Learners* (2007), the *Longman Dictionary of Contemporary English* (2014), and the *Cambridge Advanced Learner's Dictionary* (2013) contain error notes based on learner corpora, which are intended to help learners avoid common mistakes. These notes offer clear added value for dictionary users, as they draw their attention to very frequent errors, which in the case of advanced learners have often become fossilized (*accept* + infinitive, *persons* instead of *people*, *news* + plural, etc.). Although the selection of the errors is not always optimal (cf. De Cock and Granger 2005), this is a major first step that will undoubtedly be followed by others. While learner corpus data has begun to have a marked impact on EFL dictionaries, it has yet to find its way into EFL grammars. This is less surprising in light of the fact that even native corpus data was only integrated into grammars as recently as 1999, with the publication of the very first corpus-based grammar of English, *the Longman Grammar of Spoken and Written English* (Biber et al. 1999). However, it seems both inevitable and highly desirable that learner corpus data will become an essential component of grammar design in years to come. Pedagogical grammars would clearly benefit from corpus-attested information on the difficulty of grammatical categories and structures for learners in general or some L1-specific learner population. Recent initiatives such as the English Grammar Profile project (Harrison 2015) hold great promise in this regard.

While there may still be relatively little LC-informed **courseware** on the market, a fair number of teachers have used learner corpora to develop their own in-house teaching materials, which share a number of characteristics: (1) they tend to be based on learner corpora for immediate pedagogical use; (2) they are often L1-specific rather than generic; (3) they are designed with a clear teaching objective in a well-defined teaching context; and (4) they tend to be electronic rather than paper tools. This latter characteristic results from the fact that new technologies – web-based platforms, CALL authoring tools, e-mail – have brought the design of electronic pedagogical material within the reach of any computer-literate teacher/researcher and provide an ideal platform for the production and use of learner corpus data. The web-based writing environment of Wible et al. (2001) is the perfect example of a tool that facilitates the generation, annotation, and pedagogical exploitation of learner corpora. The environment contains a learner interface, where learners write their essays, send them to their teacher over the Internet, and revise them when they have been corrected by the teacher, as well as a teacher interface, where teachers correct the essays using their favorite comments (comma splice, article use, etc.) stored in a personal comment bank. This environment is extremely attractive both for learners, who get immediate feedback on their writing and have access to lists of errors they are prone to produce, and for teachers, who gradually and effortlessly build a large database of learner data from which they can draw to develop targeted exercises. Other researchers are using data resulting from computer-mediated written communication (Kung 2004; Belz and Vyatkina 2005) or oral tasks (Kindt and Wright 2001). Some pedagogical tools target LC-attested errors typical of a particular learner population. Chuang and Nesi (2007), for example, have developed *GrammarTalk*, an electronic resource focused on two of the most error-prone areas for Chinese learners, viz. articles and prepositions.

A third field in which “research from learner corpora has much to offer” (Purpura 2004, p. 272) is **language assessment.** When carefully analyzed, learner corpora can help practitioners select and rank testing material at a particular proficiency level (Barker et al. 2015). Combined with natural language processing techniques, they can also be used to draw up automatic profiles of learner proficiency. The *Direkt Profil* analyzer, for example, provides a grammatical profile for L2 French and can be used to assess learners’ grammatical level (Granfeldt et al. 2005). Learner corpora are also increasingly being used to develop and fine-tune **automated scoring** systems (Higgins et al. 2015).

All these applications show the tremendous potential of learner corpus data to inform pedagogical tools and methods. At this stage, however, LC-informed materials are still the exception rather than the rule, and there is scope for the development of a much wider range of applications in future.

Future Directions

Although learner corpora have not yet achieved a major breakthrough in the educational sector (Granger 2015b), the buzzing activity in the field and the number of learner-corpus-informed reference and teaching tools that have already been produced or are currently being designed are a clear indication that they are here to stay. Efforts in the future should be directed towards collecting data representing a wider range of target languages and sampling more diversified learner populations in a wider range of language situations and tasks. Over and above data collection, the focus should be on interpreting the data in the light of SLA theory and incorporating the results into innovative pedagogical applications. Prime among these are electronic applications and, in particular, web-based environments that allow researchers to collect and exploit learner data within the same environment and customize instructional content to meet the needs of differentiated learner populations.

Cross-References

▶ Data-Driven Learning and Language Pedagogy

References

Abel, A., Nicolas, L., Hana, J., Štindlová, B., Bykh, S., & Meurers, D. (2013). A trilingual learner corpus illustrating European reference levels. In *Learner corpus research conference 2013 – Book of abstracts,* Bergen, pp. 3–5.

Anthony, L. (2014). *AntConc,* Tokyo, Waseda University. Available at www.laurenceanthony.net/

Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, F. Meunier, & G. Gilquin (Eds.), *The Cambridge handbook of learner corpus research* (pp. 511–533). Cambridge: Cambridge University Press.

Bartning, I., & Schlyter, S. (2004). Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies, 14*, 181–199.

Belz, J. A., & Vyatkina, N. (2005). Learner corpus research and the development of L2 pragmatic competence in networked intercultural language study: The case of German modal particles. *Canadian Modern Language Review/Revue Canadienne des Langues Vivantes, 62*(1), 17–48.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning, 33*, 1–17.

Cambridge Advanced Learner's Dictionary: Fourth Edition. (2013). Cambridge: Cambridge University Press

Chuang, F.-Y., & Nesi, H. (2007). GrammarTalk: Developing computer-based materials for the Chinese EAP student. In O. Alexander (Ed.), *Proceedings of the joint conference of BALEAP (British Association of Lecturers in English for Academic Purposes) and SATEFL (The Scottish Association for the Teaching of English as a Foreign Language) on new approaches to materials development for language learning* (pp. 315–330). Bern: Peter Lang.

Cowan, R., Choi, H. E., & Kim, D. H. (2003). Four questions for error diagnosis and correction in CALL. *CALICO Journal, 20*(3), 451–463.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL), New Series, 2*, 225–246.

De Cock, S., & Granger, S. (2005). Computer learner corpora and monolingual learners dictionaries: The perfect match. *Lexicographica, 20*, 72–86.

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada, 19*, 83–102.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47*(2), 157–177.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Gilquin, G., De Cock, S., & Granger, S. (2010). *Louvain international database of spoken English interlanguage.* Louvain-la-Neuve: Presses universitaires de Louvain. Available from http://www.i6doc.com/fr/collections/cdlindsei/

Granfeldt, J., Nugues, P., Persson, E., Persson, L., Kostadinov, F., Agren, M., & Schlyter, S. (2005). Direkt Profil: A system for evaluating texts of second language learners of French based on developmental sequences. In *Proceedings of the second workshop on building educational applications using natural language processing*, *43rd annual meeting of the association of computational linguistics*, pp. 53–60. Ann Arbor, MI. Available from http://ask.lub.lu.se/archive/00021213/01/acl2005_banlp/acl2005_banlp.pdf

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast. Text-based cross-linguistic studies* (Lund studies in English, Vol. 88, pp. 37–51). Lund: Lund University Press.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO, 20*(3), 465–480.

Granger, S. (2015a). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24.

Granger, S. (2015b). The contribution of learner corpora to reference and instructional materials. In S. Granger, F. Meunier, & G. Gilquin (Eds.), *The Cambridge handbook of learner corpus research* (pp. 485–510). Cambridge: Cambridge University Press.

Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic writing. At the interface of corpus and discourse* (pp. 193–214). London: Continuum.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *The international corpus of learner english. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain. Available from http://www.i6doc.com/fr/collections/cdicle/

Gui, S., & Yang, H. (2002). *Chinese learner English corpus*. Shanghai: Shanghai Foreign Language Education Press.

Harrison, J. (2015). The English grammar profile. In J. Harrison & F. Barker (Eds.), *English profile in practice* (pp. 28–48). Cambridge: Cambridge University Press.

Hashimoto, K., & Takeuchi, K. (2012). Prototypical design of learner support materials based on the analysis of non-verbal elements in presentation. In T. Watanabe, J. Watada, N. Takahashi, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent interactive multimedia: Systems and services. Proceedings of the 5th international conference on intelligent interactive multimedia systems and services (IIMSS 2012)* (pp. 531–540). Heidelberg: Springer.

Higgins, D., Ramineni, C., & Zechner, K. (2015). Learner corpora and automated scoring. In S. Granger, F. Meunier, & G. Gilquin (Eds.), *The Cambridge handbook of learner corpus research* (pp. 587–604). Cambridge: Cambridge University Press.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). SST speech corpus of Japanese learners' English and automatic detection of learners' errors. *ICAME Journal, 28*, 31–48.

Kindt D., & Wright, M. (2001). Integrating language learning and teaching with the construction of computer learner corpora. *Academia: Literature and Language*. Available from http://www.nufs.ac.jp/~kindt/media/corpora.pdf

Kung, S.-C. (2004). Synchronous electronic discussions in an EFL reading class. *ELT Journal, 58*(2), 164–173.

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing, 8*(4), 275–281.

Levenston, E. A. (1971). Over-indulgence and under-representation – Aspects of mother-tongue interference. In G. Nickel (Ed.), *Papers in contrastive linguistics*. Cambridge: Cambridge University Press.

Longman Dictionary of Contemporary English: Sixth Edition. (2014). Pearson: Harlow.

Lüdeling, A., & Hirschmann, H. (2015). Error annotations systems. In S. Granger, F. Meunier, & G. Gilquin (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135–157). Cambridge: Cambridge University Press.

Macmillan English Dictionary for Advanced Learners: Second Edition. (2007). Oxford: Macmillan Education.

MacWhinney, B. (1999). The CHILDES system. In *Handbook of child language acquisition* (pp. 457–494). San Diego: Academic.

Mark, K. L. (1998). The significance of learner corpus data in relation to the problems of language teaching. *Bulletin of General Education, 312*, 77–90.

Milton, J., & Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. In L. Flowerdew & A. K. Tong (Eds.), *Entering text* (pp. 127–143). Hong Kong: The Hong Kong University of Science and Technology.

Mukherjee, J. (2005). The native speaker is alive and kicking – Linguistic and language-pedagogical perspectives. *Anglistik, 16*(2), 7–23.

Myles, F., & Mitchell, R. (2004). Using information technology to support empirical SLA research. *Journal of Applied Linguistics, 1*(2), 169–196.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics, 18*(3), 391–417.

Purpura, J. (2004). *Assessing grammar.* Cambridge: Cambridge University Press.

Reder, S., Harris, K., & Setzler, K. (2003). The multimedia adult ESL learner corpus. *TESOL Quarterly, 37*(3), 546–557.

Scott, M. (2012). *WordSmith Tools*. Liverpool: Lexical Analysis Software.

Seidlhofer, B. (2004). Research perspectives on teaching English as a Lingua Franca. *Annual Review of Applied Linguistics, 24*, 209–239.

Van Rooy B., & Schäfer L. (2003). Automatic POS tagging of a learner corpus: The influence of learner error on tagger accuracy. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the corpus linguistics 2003 conference,* UCREL, Lancaster University, pp. 835–844.

Wible, D., Kuo, C.-H., Chien, F.-Y., Liu, A., & Tsao, N.-L. (2001). A web-based EFL writing environment: Integrating information for learners, teachers, and researchers. *Computers and Education, 37*, 297–315.

Yang, H., & Wei, N. (2005). *College English learners' spoken English corpus*. Shanghai: Shanghai Foreign Language Education Press.