Editorial

Learner Corpus Research
An interdisciplinary field on the move

The compilation and accessibility of computer corpora and software tools has revolutionized linguistic analysis in the last two decades. Corpora and corpus linguistic tools and methods are also increasingly used in the study of second and foreign language (L2) learning. Learner corpora, generally defined as systematic collections of authentic, continuous and contextualized language use (spoken or written) by L2 learners stored in electronic format, are a special type of empirical data used by scholars in a variety of disciplines. One of the first learner corpora compiled is probably the European Science Foundation (ESF) Second Language Database which was collected in the 1980s under the direction of Wolfgang Klein and Clive Purdue (Klein & Purdue 1997, Perdue 1993). The ESF project focused on the structure and success of naturalistic Second Language Acquisition (SLA) in forty adult immigrant workers living in five different European countries as they acquired one of five L2s (Dutch, English, French, German and Swedish) over a period of two and a half years. The resulting dataset amounted to a dense and longitudinal corpus of 18,000 pages of L2 oral transcriptions that is available today on the Talkbank project website1 (Perdue 1993: 108). Learner Corpus Research (LCR) as a field, however, only emerged at the turn of the 1990s from the then developing field of Corpus Linguistics when academics and publishing houses, simultaneously but independently, started to realize the considerable potential of large computerized datasets of learner production to describe learner language and/or develop new pedagogical tools and methods that target language learners’ specific needs. Indeed, the few LCR articles published in the first half of the 1990s all promoted the potential benefits of authentic language use for foreign language learning and teaching (e.g. Granger 1993, Granger 1994, Granger & Meunier 1994, Milton & Chowdhury 1994, Milton & Tsang 1993). Drawing a parallel with the revolutionary impact that the use of language corpora had had on dictionary-making, Granger (1994) wrote:

I have no doubt that the investigation of computerized learner corpora may well be able to achieve the kind of spectacular results which we have witnessed in lexicography, opening up new avenues of research and giving rise to a new generation of grammars, dictionaries, vocabulary books and language software programs developed with the difficulties of the learner in mind. (Granger 1994: 29)

In 1990, Granger initiated the International Corpus of Learner English (ICLE) project to complement the International Corpus of English (ICE) with a corpus of learner English while at the same time providing a more solid empirical basis for SLA research and the development of material for teaching English as a Foreign Language (EFL) (Granger 1996a). The second edition of the ICLE contains 6,085 argumentative and literary essays written by higher-intermediate to advanced EFL learners from sixteen mother tongue backgrounds (Granger et al. 2009). Two other major learner corpus compilation projects were launched at the same time: the Longman Learners’ Corpus, a ten million word corpus of essays and exam scripts submitted by EFL learners all over the world, and the Hong Kong University of Science and Technology (HKUST) Learner Corpus, which currently amounts to 25 million words and consists of different academic text types written in English by Chinese undergraduate students.

Although restricted to EFL at the outset, LCR is now being undertaken on many different languages, creating a diverse and rapidly expanding international network of researchers. This is evidenced by the number and variety of learner corpus compilation projects listed on the ‘Learner Corpora around the World’ webpage maintained by the Centre for English Corpus Linguistics (CECL) at Louvain-la-Neuve, Belgium. While there is still a strong focus on foreign language learning and teaching in LCR, the use of learner corpora has diversified considerably. Learner corpus data are now being used in various disciplines including SLA research, Language Testing and Assessment and Pedagogical Lexicography. They also constitute the raw material for a wide range of Natural Language Processing (NLP) based tasks (e.g. training of part-of-speech taggers and parsers, automatic error detection and correction, automatic exercise generation) and are used in tools as varied as essay scoring systems and intelligent Computer Assisted Language Learning (CALL) programs. Recently, they have also been at the forefront in a new field of research, i.e. native language identification, which aims to automatically identify the first language of a language user on the basis of that person’s production of the target language.


LCR has developed rapidly over the last twenty years and now has its own international academic association, the Learner Corpus Association, officially founded in 2013, which coordinates the organization of a biennial international conference. The field will also have its own handbook in the near future (Granger et al. forthcoming). In view of these recent developments, the publication of an international peer-reviewed journal dedicated to LCR naturally presented itself as the next step in the maturation of the field. We believe that the *International Journal of Learner Corpus Research* fills a gap because so far, LCR studies had no privileged high-quality publication outlet. The journal is accordingly meant to be a forum for researchers who collect, annotate, and analyze computer learner corpora and/or use them to investigate topics in SLA research and linguistic theory in general, inform foreign language teaching, develop learner corpus-informed tools (e.g. courseware, proficiency tests, dictionaries, and automatic spell- and grammar-checkers), or conduct NLP tasks. It is our ambition that the journal showcases the multidisciplinary and broad scope of practice that characterizes the field and features original research covering methodological, theoretical and applied work in any area of LCR. We will publish two issues per year and welcome original research articles, shorter research notes and book reviews. We also encourage the publication of special issues with articles commissioned by guest editors.

It is a great pleasure and honor to open the first issue of the IJLCR with an article written by Sylviane Granger, one of the pioneers and main driving forces behind LCR. In 1996, she proposed Contrastive Interlanguage Analysis (CIA) as a systematic method to analyze and compare learner corpus data (Granger 1996b). Since then, CIA has been used extensively in LCR but it has also been criticized on several fronts. Almost twenty years later, Granger discusses the major criticisms the method has faced since its introduction and presents a revised model, CIA², which, she argues, explicitly acknowledges the important role played by interlanguage variation and is thought to be more closely in alignment with current theory and practice in L2 learning and teaching.

The issue then features four research articles that bear testimony to the broadening of LCR described above and point to important future developments in the field. Ursula Maden-Weinberger’s study of the German subjunctive mood by learners of L2 German could be described as a classic CIA study comparing learner data to native speaker data, and learner data at different levels of proficiency. Its major strength lies in the author’s attempt to move beyond a descriptive study and explain learners’ difficulties by drawing on connectionist theories of SLA.

The following two contributions both critically examine the standard concept of a learner corpus, although in different ways. Nicole Tracy-Ventura and Florence

---

Myles make use of three communicative tasks (a guided interview, a picture-based narrative and a historical figure description) and an experimental task to investigate the impact of task variability on the acquisition of tense and aspect in L2 spoken Spanish. They report significant differences in the emergence and accurate use of past tense morphology across tasks and show that the less controlled tasks triggered few instances of tense-aspect morphological contrasts theoretically relevant for SLA research. As a result, they emphasize the importance of taking task variability into account in LCR and call for a broadening of the definition of what counts as a learner corpus to provide sufficiently rich and representative data to study learner language.

Quite differently, Theodora Alexopoulou, Jeroen Geertzen, Anna Korhonen and Detmar Meurers explore the opportunities offered by big learner data for SLA research. Like big data, big learner data have tremendous potential, but also present serious challenges. Unlike traditional learner corpora, for example, big learner data such as the *Education First Cambridge Open Language Database* come with very few details about the learners and Alexopoulou et al. have to rely on the learners’ nationality to approximate their L1 background. The authors also discuss a number of methodological issues that arise from the use of big educational data as a learner corpus and show how NLP is essential for the automated processing of the noisy and often unpredictable properties of such datasets. As a case study, they follow the developmental trajectory of relative clauses, constructions that necessitate deep syntactic analysis.

Stefan Th. Gries and Sandra Deshors’s study adds to the growing body of research that investigates the similarities and differences between EFL and English as a Second Language (ESL, i.e. indigenized post-colonial varieties of English). The authors report the results of a multifactorial analysis of seventeen lexical verbs in the dative alternation in speech and writing of German and French EFL learners and ESL speakers from Hong Kong, India, and Singapore, addressing the theoretical question of whether EFL and ESL represent discrete variety types or a continuum. Gries and Deshors’s study is particularly important in that it constitutes yet further compelling evidence that more rigorous methodological approaches and statistical techniques are urgently needed in LCR.

The first issue also includes an interview we conducted with another LCR pioneer, Yukio Tono. In this interview, Tono reflects on LCR and its major developments since the publication of *Learner English on Computer* (Granger 1998), the first book dedicated to LCR and now a classic in the field. He shares with us his views on a number of issues in LCR including its relationship with SLA, its being criticized for paying too much attention to errors and deficiencies, and its relatively limited impact on pedagogical material development. Our inaugural issue closes with reviews of three recently published books, two research
monographs and an edited volume, that bear witness to the expanding research agenda in LCR.

We would like to express our sincere gratitude to a number of people who played a big part in the publication of the IJLCR. We are very grateful to Gaëtanelle Gilquin, Sylviane Granger and Fanny Meunier who were instrumental in setting up the journal project. We most particularly thank Sylviane Granger, who will act as consulting editor to the journal, for her catching enthusiasm and expert guidance. Many thanks are also due to Kees Vaes for his interest in and support of the project from the beginning; Sandra Götz, the reviews editor for IJLCR, for her dedication to the journal; our board members for their support and willingness to share their expertise with us; the many colleagues who acted as reviewers for the present issue; and last but not least Maike Rocker from the University of Bremen, who acts as editorial assistant for the IJLCR.

Marcus Callies & Magali Paquot
IJLCR General Editors

References
