Learner Corpus Research and Second Language Acquisition: an attempt at bridging the gap

Bert Le Bruyn and Magali Paquot

1 Learner Corpora and Their Origins

Learner corpora are traditionally defined as 'systematic collections of authentic, continuous and contextualized language use (spoken or written) by L2 learners stored in electronic format' (Callies & Paquot 2015). With this characterization, it appears very clearly that learner corpus researchers have always been interested in exploring the output of the more open-ended types of contextualized production tasks assigned to L2 learners (e.g. Granger 2008; Tracy-Ventura & Myles 2015). The term 'learners' here refers to Foreign and/or Second Language learners rather than to learners acquiring their native language (L1).

The original idea behind the development of major learner corpora like the *International Corpus of Learner English* (ICLE; Granger 1993) was to enrich existing corpus collections with learner varieties and 'pass on the advances made in computerized corpus linguistics to applied linguistics', with a focus on the domains of language learning and teaching (Granger 1993). This situates the origins of Learner Corpus Research (LCR) outside the domain of theory-driven Second Language Acquisition (SLA) research.

2 From (Applied) Linguistics to Theory-Driven SLA

The fact that learner corpora systematically record learner language could have made them into a relevant resource for theory-driven SLA – much in line with the role CHILDES corpora have played in the L1 acquisition literature (MacWhinney 2000). However, a cursory look at publications in major SLA journals from the 1990s till now shows that this expectation is not borne out: learner corpus methodology has been more the exception than the rule. Looking back at the beginnings of LCR, Bell, Collins, and Marsden (in this volume) attribute LCR's lack of popularity within SLA research to its

https://doi.org/10.1017/9781108674577.002 Published online by Cambridge University Press

'preoccupation for coding errors, L1 transfer errors, and deviations from a target-like norm' at a time when SLA research took pride in having moved beyond surface generalizations.

Over the years, LCR and theory-driven SLA have started to interact. In *The Cambridge Handbook of Learner Corpus Research* (Granger et al. 2015), Myles shows that LCR can interact with a range of SLA theories, citing examples like Rankin's work on the Interface Hypothesis (Rankin 2009) and Bonilla's work on Processability Theory (Bonilla 2015). She also highlights, however, the fact that SLA studies require different corpora to answer the questions they raise. For example, whereas the ICLE is a good choice for checking certain predictions of the Interface Hypothesis, its focus on essays by advanced learners would make it irrelevant for testing predictions of Processability Theory.

Myles's (2015) conclusion on the interaction between LCR and SLA is a careful one. She notes punctual excursions from one field to the other but no systematic collaboration: 'second language researchers have been rather slow in taking advantage of learner corpora and their associated computerised methodologies …, and LCR is not always fully informed by SLA research'. The present volume sets out to give an updated evaluation and showcases the continued efforts of the two communities to grow closer to each other, this volume being one of them.

3 LCR Meets SLA

The initiative for this volume dates back to a workshop the two editors organized in March 2017. The goal was to bring together researchers interested in doing SLA research through LCR. The call that was launched invited contributions that would:

(1) address SLA research questions with the help of learner corpus data and corpus linguistic techniques;
(2) test key constructs in SLA theory on the basis of learner corpus data;
(3) carry out a learner corpus study and compare results with findings from previous SLA studies that relied on more experimental techniques;
(4) adopt a mixed method approach that involves learner corpus data as one of the data types used;
(5) revisit previous LCR studies in the light of SLA theory.

The submissions we received were mostly of types (1) to (3). The present volume brings together the majority of the papers that were

https://doi.org/10.1017/9781108674577.002 Published online by Cambridge University Press

presented at the workshop. Each paper was reviewed by a specialist in SLA and a specialist in LCR. We know this was a challenge both for the reviewers and the authors. We wholeheartedly thank all of them for their efforts.

In the remainder of this introduction we give a brief description of the different research contributions. We leave the evaluation of the interaction between SLA and LCR to two eminent scholars, one from the field of LCR, the other from the field of SLA. Sylviane Granger and Florence Myles gracefully agreed to write commentary chapters. These conclude the volume.

4 Overview of Chapters

The origins of this volume make it clear that it is by no means intended as a survey of all current SLA-oriented work in LCR (see Tracy-Ventura & Paquot, in press). However, we believe the volume provides a fair impression of how the fields of LCR and SLA are currently interacting with each other. It showcases in an exemplary way how SLA theory can be used today to inform learner corpus analyses and how learner corpus findings can be used to inform SLA theory. This cross-fertilization has arguably started to take place with the broadening of learner data types collected in LCR. The first learner corpora consisted of written texts – typically argumentative essays – produced by intermediate to advanced English as a Foreign Language (EFL) learners in higher education. However, the chapters in this volume show that many different types of learner corpora are now being compiled, e.g. spoken learner corpora, corpora of foreign languages other than English, cross-sectional corpora representing various proficiency levels, and longitudinal learner corpora. Similarly, learner corpus researchers have always sought to describe learner data and complement them with text and learner variables such as L1 background, age, or time spent studying English at school. More recently, however, efforts have been made to widen the scope of recorded metadata and include variables that are key to the SLA research agenda (e.g. external measurement of proficiency, answers to a motivation test, language aptitude scores) (see Möller 2017). The chapters in this volume further bear witness to the ever more sophisticated ways researchers succeed in making learner corpora speak, from data extraction to statistical analysis. A final strength of current learner-corpus-based research highlighted by the volume is its inclusive character: all theoretical frameworks are welcome, data are shared, and bridges are built, e.g. to the neighbouring discipline of World Englishes.

We have refrained from organizing the volume into clearly delimited parts but found it practical for the purposes of this overview to highlight some common themes in the chapters that appear together.

4.1 Universal Tendencies and Cross-Linguistic Influence

The first three chapters focus on classical themes in SLA, namely the role of cross-linguistic influence, the role of universal tendencies, and the interaction between the two.

Tania Ionin and María Belén Díez-Bedmar bring together their SLA and LCR expertise to compare the L2 article acquisition of Spanish and Russian B1 and B2 learners of English. The empirical basis of their investigation consists of 200 texts from the *Cambridge Learner Corpus*. The results confirm the generalization that learners with a native article language (like Spanish) are quicker in picking up and applying relevant semantic insights when acquiring a second article language than learners with an article-less L1 (like Russian). The results, however, do not confirm the role specificity has played in uncovering universal tendencies in SLA research (Ionin et al. 2004; Ionin et al. 2009).

Valentin Werner, Robert Fuchs, and Sandra Götz investigate the role of cross-linguistic influence and universal mechanisms in the acquisition of the perfect/simple past alternation in English. Cantonese and German learner data are drawn from the LINDSEI corpus (*Louvain International Database of Spoken English Interlanguage*) and the ICLE corpus. Native comparison data come from the LOCNEC and LOCNESS corpora (*Louvain Corpus of Native English Conversation* and *Louvain Corpus of Native English Essays*). The authors follow the MuPDAR approach (Multifactorial Prediction and Deviation Analysis with Regressions) (Gries & Deshors 2014; Gries & Bernaisch 2016; Wulff and Gries, this volume): they first determine the way different factors interact in native speaker production and then see how these factors play out in learner production. The results show that there are quantitative differences between Cantonese and German learners, the latter being closer overall to native production. The authors, however, also find that it is difficult to pinpoint where qualitative differences would reside that can be attributed unambiguously to L1 influence. They conclude – citing Bardovi-Harlig – that cross-linguistic influence, if at play, is to be found 'in the details rather than in the larger picture'.

Lea Meriläinen approaches SLA and LCR from the adjacent domain of World Englishes. What makes World Englishes corpora different from most LCR corpora is that they lack metadata on individual

https://doi.org/10.1017/9781108674577.002 Published online by Cambridge University Press

speakers' L1s. The focus in this field has consequently been on universal tendencies rather than cross-linguistic influence, and detailed comparisons between linguistic systems have been less central in this field than in SLA or LCR. The author presents two case studies based on learner data from the ICLE corpus and the MEC corpus (*Corpus of Matriculation Examination*) and on native data from the LOCNESS corpus. The learners come from a number of different L1 backgrounds: Finnish, German, Swedish, Chinese, and Japanese. The case studies focus on embedded inversions ('I don't know what are you asking me') and preposition omission ('She went town'). Results tentatively show that embedded inversions and preposition omission are found in the production of learners of all backgrounds but are most frequent when learners have L1s that differ in the relevant respects from the L2.

4.2 Proficiency and Time

The issue of proficiency extends beyond research focusing on the interplay between cross-linguistic influence and universal tendencies. Pinning down what it means to become more proficient over time is one of the core issues of SLA research. One way of approaching proficiency is through the constructs of *complexity*, *accuracy*, and *fluency*, CAF for short (see e.g. Housen & Kuiken 2009 for a classic reference). The rationale behind these is that learners who become more proficient will start using more complex language, will become more accurate, and will become more fluent in their production. A variety of automated measures have been proposed to operationalize the constructs of complexity and fluency. This does not hold for accuracy, though. Its operationalizations typically require laborious hand-coding and – depending on the dimension of accuracy the researcher is interested in – have led to varying degrees of reliability. Charlene Polio and Hyung-Jo Yoon, building on previous work by Bestgen and Granger (2014), introduce a number of automated measures of accuracy drawn from usage-based theories of SLA. They explore whether the construct of accuracy can be measured by checking if the bi- and trigrams occurring in learner texts also occur in large corpora like COCA (*Corpus of Contemporary American English*) and by assessing how likely the words composing them are to occur together in this corpus. Polio and Yoon's learner data are taken from three corpora: the Michigan State University (MSU) corpus (descriptive essays), the genre corpus (argumentative essays), and Yoon's dissertation pilot data (narrative essays). Their results show that bi- and trigrams that do not occur in COCA are in general

considered erroneous by human judges. They also find that their automated measures of accuracy typically pattern with classical accuracy measures – as desired – and not with complexity measures. Finally, their results show that up to half of the variance of hand-coded error counts is accounted for by their automated measures, a welcome result that warrants further exploration.

Magali Paquot, Hubert Naets, and Stefan Gries's contribution focuses on phraseological complexity and looks into the effects of proficiency and time spent on learning English on the sophistication of verb + object co-occurrences. Interestingly, they assess phraseological complexity in much the same way as Polio and Yoon assess accuracy, namely by checking how likely words making up word combinations found in learner texts are to be found together in a large reference corpus. There are two crucial differences, though. First, Paquot et al. look not at bi- and trigrams but at words in a syntactic dependency relation, namely verbs and their objects. Second, they make use of mean mutual information (MI) scores to examine the phraseological complexity of foreign language texts; this association measure has been shown to promote relatively less frequent and more semantically complex word pairs, thus tapping into the sophistication of word combinations (cf. Paquot 2019).The learner data Paquot et al. use are taken from the LONGDALE corpus (*Longitudinal Database of Learner English*). Two important assets of this corpus are that it comes with up to three measure points per learner – spread over the three years of their undergraduate degrees – and that each learner text comes with an Oxford Quick Placement Test score. The reference corpus is ENCOW14 AX, a large web corpus. Results show that phraseological complexity improves with proficiency and that institutional status (i.e. the year the students are in) does not hold any predictive power. Next to establishing phraseological complexity as a good way to probe higher levels of proficiency, the crucial point the chapter makes is that time and institutional training do not have an effect on the development of phraseological complexity per se; what matters more is foreign language proficiency and whether learners rate higher on an external test from one year to the next.

The issue of time is also central to the next two chapters. Nicole Tracy-Ventura, Amanda Huensch, and Rosamond Mitchell look into forces of retention and attrition of the effects of institutional training, with a focus on lexical diversity. Data are taken from the original LANGSNAP corpus (*Languages and Social Networks Abroad Project*) that looked into the effects of a nine-month stay abroad in France or Spain by English students (with data collection before, during, and up to a year after the stay). These data are supplemented

with those from a follow-up study with data collection four years after the stay abroad. The results show that lexical diversity continues to improve post-instruction in oral but not in written tasks and that peak attainment and language exposure/use contribute significantly to predicting gains, with peak attainment explaining most variance. The results are relevant for the Dynamic Model of Multilingualism (Herdina & Jessner 2002) and the Neurolinguistic Theory of Bilingualism (Paradis 1993, 2004).

Marjolein Verspoor, Wander Lowie, and Martijn Wieling report on a longitudinal study of Dutch high school students learning English. What makes the learner corpus they use exceptional is the high density of the data collection, with one measurement per week for a period of 23 weeks. The authors find that the holistic scores of the last two learner texts are significantly higher than those of the first two, suggesting an increase in proficiency. They also find that syntactic development (operationalized as mean length of T-unit or MLTU) and lexical development (operationalized as Guiraud) show an upward trend at the group level with a clear but small decrease at the end. Crucially, however, when the authors zoom in on the syntactic and lexical development of individual learners, they find that no learner follows the same path or evolves in a linear fashion. The only generalization that does seem to hold across learners is that gains in lexical development precede gains in syntactic development. Verspoor et al. interpret these results from a Complex Dynamics Systems point of view in which individual trajectories take centre stage.

4.3 Corpus Analysis and Development

The remaining two chapters focus on corpus analysis and development. The first responds to Verspoor et al.'s call for attention for individual learners. Stefanie Wulff and Stefan Gries further develop the MuPDAR approach that we also found in the contribution by Werner et al. On the basis of a case study of the *s*- and *of*-genitive alternation, they show how the approach provides a comprehensive picture of native speaker behaviour and allows for a fine-grained analysis of the aspects of a target structure the learner commands or still has difficulties with. They show how results can be obtained not only at the group level but also at the individual level, a move that is both relevant from a theoretical perspective and is also shown to lead to a better analysis of the data.

Philippa Bell, Laura Collins, and Emma Marsden offer a comprehensive overview of design choices they are facing in the development of a learner corpus looking into L2 development in foreign language

classes in primary and secondary education. Their goal is both applied and theoretical. On the applied side, the main question is how and what to teach to optimize acquisition. On the theoretical side, they are interested in finding out how language is acquired in limited-exposure contexts. The authors report on and discuss design choices they piloted. These include the language of the instruction of the tasks, the tasks themselves, and transcription/coding issues.

Acknowledgements

We would like to thank the authors and reviewers for all their hard work on the volume. Special thanks go to Lex Tavenier for his in-house editorial support as well as to Cambridge University Press for their long-term support for this volume. We also gratefully acknowledge the financial support of the F.R.S.-FNRS, NWO, and the department of Language, Culture and Communication at Utrecht University.

References

Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26, 28–41.

Bonilla, C. L. (2015). From number agreement to the subjunctive: Evidence for Processability Theory in L2 Spanish. *Second Language Research* 31(1), 53–74.

Callies, M. & Paquot, M. (2015). Learner Corpus Research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research* 1(1), 1–6.

Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan, & N, Oostdijk (eds.), *English Language Corpora: Design, Analysis and Exploitation*, 57–69. Amsterdam & Atlanta, GA: Rodopi.

(2008). *Learner Corpora*. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, 259–275. Berlin & New York, NY: Walter de Gruyter.

Granger, S., Gilquin G., & Meunier, F. (eds.) (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Gries, S. Th. & Bernaisch, T. (2016). Exploring epicentres empirically: Focus on South Asian Englishes. *English World-Wide* 37(1), 1–25.

Gries, S. Th. & Deshors, S. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora* 9(1), 109–136.

Herdina, P. & Jessner, U. (2002). *A Dynamic Model of Multilingualism: Changing the Psycholinguistic Perspective*. Clevedon: Multilingual Matters.

Housen, A. & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4), 461–473.

Ionin, T., Ko, H., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition* 12(1), 3–69.

Ionin, T., Zubizarreta, M., & Philippov, V. (2009). Acquisition of article semantics by child and adult L2-English learners. *Bilingualism: Language and Cognition* 12(3), 337–361. https://doi.org/10.1017/S1366728909990149

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Möller, V. (2017). A statistical analysis of learner corpus data, experimental data and individual differences: Monofactorial vs. multifactorial approaches. In P. de Haan, S. van Vuuren, & R. de Vries (eds.), *Language, Learners and Levels: Progression and Variation*, 409–439. Louvain-la-Neuve: Presses universitaires de Louvain.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (eds.). *The Cambridge Handbook of Learner Corpus Research*, 309–332. Cambridge: Cambridge University Press.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research* 35(1), 121–145. https://doi.org/10.1177/0267658317694221

Paradis, M. (1993). Linguistic, psycholinguistic, and neurolinguistic aspects of 'interference' in bilingual speakers: The Activation Threshold Hypothesis. *International Journal of Psycholinguistics*, 9(2), 133–145.

(2004). *A Neurolinguistic Theory of Bilingualism*. Amsterdam: John Benjamins.

Rankin, T. (2009). Verb second in advanced L2 English: A learner corpus study. In *Proceedings of the 10th Generative Approaches to Second Language Acquisition Conference*, 46–59. Somerville, MA: Cascadilla Proceedings Project.

Tracy-Ventura, N. & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research* 1(1), 58–95.

Tracy-Ventura, N. & Paquot, M. (eds). (in press). *The Routledge Handbook of Second Language Acquisition and Corpora*. London: Routledge.