Language learning research at the intersection of experimental, computational and corpus-based approaches

Patrick Rebuschat[1,2], Detmar Meurers[2,3], and Tony McEnery[1,4]

[1] Department of Linguistics and English Language, Lancaster University; [2] LEAD Graduate School and Research Network, University of Tübingen; [3] Seminar für Sprachwissenschaft, University of Tübingen; [4] ESRC Centre for Corpus Approaches to Social Science

Keywords: Psycholinguistics; corpus linguistics; computational linguistics; natural language processing; first language acquisition; second language acquisition; multimethod approaches; interdisciplinary research

Acknowledgements: Our research was supported by the Economic and Social Research Council, UK (grant ES/K002155/1) and by the LEAD Graduate School & Research Network (grant DFG-GSC1028), a project of the Excellence Initiative of the German federal and state governments.

Correspondence: Correspondence concerning this article should be addressed to Patrick Rebuschat, Department of Linguistics and English Language, Lancaster University, Lancaster LA1 4YL, United Kingdom, E-mail: p.rebuschat@lancaster.ac.uk.

Language acquisition occupies a central place in the study of human cognition, and research on how we learn language can be found across many disciplines, from developmental psychology and linguistics to education, philosophy and neuroscience. It is a very challenging topic to investigate given that the learning target in first and second language acquisition is highly complex, and part of the challenge consists in identifying how different domains of language are acquired to form a fully functioning system of usage (Ellis, this volume). Correspondingly, the evidence about language use and language learning is generally shaped by many factors, including the characteristics of the task in which the language is produced (Alexopoulou et al, this volume). The challenge is further complicated by the fact that language acquisition is affected by individual learner characteristics. Individual differences are particularly well-studied for second language acquisition, where it is clear that factors such as native language, type of instruction, and motivation affect learning rate and ultimate attainment (Ushioda & Dörnyei, 2012; Williams, 2012). But recent research indicates that there is also considerable individual variation in child language development (see Rowland, 2013). To develop an understanding of language acquisition, we need to take into account these individual differences (MacWhinney, this volume).

Despite these and other challenges, the past decades have witnessed significant progress in our understanding of how children and adults learn languages. The conceptual and empirical progress arguably is fueled by an increasing range of methods and approaches that are being used to study language acquisition (see Hoff, 2011; Mackey & Gass, 2012). For example, experimental approaches using artificial or natural languages have made it possible to investigate how changes across exposure conditions such as input frequency, instruction type, or prior knowledge affect learning in rigorously controlled environments. Learner corpora are growing in

size and task types covered, with increasingly rich annotation supporting detailed analyses employing sophisticated statistical methods. Digital learning environments integrating computational methods hold the promise of supporting the systematic exploration of learning mechanisms in authentic teaching and learning, providing new sources of evidence on the roles played by the linguistic environment, interaction, and feedback in learning. The investigation of a complex phenomenon like language acquisition can significantly benefit from insights, tools, and methods from many disciplines, yet it is still relatively rare to find studies that combine multiple approaches.

The research described in Monaghan and Mattock (2012), Ellis, Römer and O'Donnell (2016), and Christiansen and Chater (2016) transparently illustrates the potential of multimethod approaches to language. For example, Monaghan and Mattock's (2012) investigation of word learning is an excellent illustration of how corpus research can connect with experimental research. Monaghan and Mattock first conducted corpus analyses of child-directed speech. They then used the information derived from these analyses to construct an artificial language that is based on natural language statistics. On this basis, they investigated the acquisition of nouns and verbs by adult learners in an artificial language experiment. While artificial language research is occasionally criticized for its limited ecological validity, the use of distributional information from natural language corpora in the artificial language construction mitigates some of this criticism (see also Monaghan & Rowland, this volume). Another impressive example of multimethod research is Ellis, Römer, and O'Donnell (2016). Ellis et al. investigate the acquisition, processing and use of Verb-Argument Constructions (VACs), and their monograph contains series of behavioral experiments, large-scale corpus analyses supported by Natural Language Processing (NLP) techniques, and several computational simulations (connectionist

and agent-based). The result of this systematic multimethod exploration is a significant, in-depth understanding of how we learn, process and use VACs – and a research model for others to follow suit. Finally, Christiansen and Chater's (2016) theoretical framework for understanding language acquisition, evolution, and processing is the direct result of multimethod research and would not be possible without the insights the authors gained from working at the intersection of experimental, computational and corpus-based approaches for more than two decades.

The question of how to promote multidisciplinary research across methodological boundaries has been central to the work of the three editors of this volume. A series of review articles aiming to connect research areas and introduce methodologies exemplify this (e.g., Meurers, 2012, 2015; Meurers & Dickinson, this volume; Rebuschat, 2013). One of the editors, Tony McEnery, directs the ESRC Centre for Corpus Approaches to Social Sciences (CASS, http://cass.lancs.ac.uk) at Lancaster University, whose primary objective is to enable colleagues in other, non-linguistic disciplines to utilize the corpus approach. The two other editors are part of Tübingen's unique LEAD Graduate School and Research Network, which brings together over 130 scientists from Education, Psychology, Linguistics, Neuroscience, Informatics, Sociology, and Economics to investigate learning and educational achievement.[1] The LEAD initiative includes an interdisciplinary research and training program for doctoral students and postdocs, which is funded by Germany's Excellence Initiative. In the same spirit, we have enjoyed organizing numerous symposia, workshops, summer schools, and conferences, and we have edited several books and special journal issues with the specific aim of bringing together leading researchers from different disciplines whose paths would normally not cross (e.g., Andringa & Rebuschat, 2015; Meurers, 2009; Monaghan and Rebuschat, in prep; Rebuschat,

[1] For more information on the LEAD Graduate School & Research Network, please see http://www.lead.uni-tuebingen.de

2015; Rebuschat, Rohrmeier, Hawkins, & Cross, 2012; Rebuschat & Williams, 2012). The present volume is part of this ongoing effort.

This volume

This volume was inspired by a symposium on "Connecting data and theory: Corpora and second language research", which was jointly organized by the editors and took place in Lancaster, UK, on July 19, 2015. The symposium was jointly funded by the Language Learning Roundtable Grant Program and by the ESRC Centre for Corpus Approaches to Social Science (CASS). The objective was to establish a dialogue between experts on second language acquisition, corpora, and computational analysis methods. This dialogue can significantly enrich the empirical basis of second language research but, to date, collaborations across these fields are still rare. The symposium aimed at directly addressing this shortcoming. There were three sessions, each approaching the symposium topic from a distinct research area. Nick Ellis and Brian MacWhinney provided the view from cognitive psychology, Detmar Meurers and Markus Dickinson the view from computational linguistics, and Anke Lüdeling and Sylviane Granger the view from corpus linguistics. The symposium concluded with a general discussion.

The discussion and feedback were both very positive and lively, and when the opportunity arose to produce a volume of *Currents in Language Learning*, we readily agreed to do so. Five presentations of the symposium provided the basis for four expanded and updated chapters (Ellis; Lüdeling et al.; MacWhinney; Meurers & Dickinson). Additional chapters were written by colleagues who attended the symposium and made thoughtful contributions (Alexopoulou et al.; Gablasova et al; Monaghan & Rowland; Ziegler et al.). Based on the symposium discussions, we decided to expand the scope for the special issue in two areas. We

solicited a manuscript that would contribute a language testing angle (Wisniewski) and broadened the topic to language learning in general, given the long and fruitful tradition of using corpora, NLP tools and computational modeling in child language research.

As a result, the third volume of the *Currents in Language Learning* series brings together leading researchers in cognitive psychology, computational linguistics, corpus linguistics, developmental psychology, and linguistics. Our contributors were asked to (i) discuss recent work and trends, (ii) outline opportunities and challenges of combining multiple approaches, and (iii) propose directions for future research at the intersection of experimental, computational and corpus-based approaches to language learning. Each submission was peer-reviewed by several anonymous reviewers and by the editors.

In Chapter 2, Padraic Monaghan and Caroline Rowland describe the challenges of combining experimental, computational and corpus approaches to research in child language acquisition. Their paper clearly articulates the benefits of multidisciplinary approaches by providing three examples for a successful combination of methods for studying grammatical category acquisition, morphological development, and the acquisition of sentence structure. On this basis, they conclude with a discussion of future directions. In Chapter 3, Nick Ellis approaches the topic from the perspective of usage-based linguistics. Ellis clearly illustrates the essential contributions made by experimental, computational, and corpus-based research to the establishment of usage-based theories of language (see also Ellis, Römer, & O'Donnell, 2016). In Chapter 4, Detmar Meurers and Markus Dickinson provide a comprehensive review of how computational linguistics and NLP techniques can contribute to our understanding of second language learning. They focus on two contributions: First, computational linguistics can enrich the options for obtaining substantial amounts of data for language learning research, including

data obtained via Intelligent Computer-Assisted Language Learning (ICALL) interfaces (see also Ziegler et al., this volume). Second, NLP techniques can support the identification and interpretation of data of relevance to second language research via automatic linguistic annotation of large-scale corpora – which they argue requires more cross-disciplinary discussion to operationalize relevant learner language distinctions and develop annotation schemes that are adequate to support second language research.

The next three chapters focus on essential methodological considerations arising from corpus-based language learning research. In Chapter 5, Anke Lüdeling, Hagen Hirschmann, and Anna Shadrova illustrate how learner corpus data can be used to investigate acquisition patterns by concentrating on second language morphological productivity as a test case. They raise methodological points regarding linguistic modeling, the formation of target hypotheses, and error annotation. In Chapter 6, Dana Gablasova, Vaclav Brezina and Tony McEnery focus on collocations in language learning research. The interest in formulaic language has been growing in both first and second language research, and there is now a considerable number of experimental and corpus-based studies in this area (e.g., Christiansen & Arnon, in press). Gablasova et al. critically review measures of association that are frequently used to identify collocations (t-score, MI score, Log Dice) and discuss how a better understanding of these measures greatly facilitates the interpretation of trends in language production data. In Chapter 7, the same authors focus on the role of corpus-based frequency information for advancing our understanding of how languages are learned. They illustrate the issues involved in the interpretation and comparison of corpus frequencies by contrasting several first and second language corpora.

Chapters 8 and 9 provide concrete examples of the benefits of working at the intersection of experimental, computational and corpus-based approaches to language learning. In Chapter 8, Dora Alexopoulou, Marije Michel, Akira Murakami, and Detmar Meurers test hypotheses derived from instructed SLA research and task-based language teaching (TBLT) by applying techniques from computational linguistics to a very large learner corpus. They analyze the texts in the EF-Cambridge Open Language Database (EFCAMDAT, https://corpus.mml.cam.ac.uk/efcamdat), a learner corpus that contains over 70,000,000 words collected through an online language learning platform. Their paper demonstrates how large corpora and NLP techniques can contribute to contemporary language learning research by complementing experimental evidence. In Chapter 9, Nicole Ziegler, Detmar Meurers, Patrick Rebuschat, Simón Ruiz, José L. Moreno-Vega, Maria Chinkina, Wenjing Li, and Sarah Grey combine theoretical and methodological insights from SLA, NLP and ICALL research to investigate the effectiveness of input enhancement in promoting second language development. Their study is experimental, but data is collected via a web-based ICALL system (WERTi) that provides computerized pedagogical treatment of learner-selected texts and automatically tracks and collects learners' action and engagement with the input. This results in a particularly rich data set, beyond what is typically available via traditional experimental approaches.

In Chapter 10, Katrin Wisniewski provides a conceptual review of how learner corpora can contribute to language testing research, emphasizing the importance of empirical scale validity. Wisniewski focuses on the Common European Framework of Reference (CEFR), the most common European reference tool to describe levels of foreign language proficiency, and explicitly works out the opportunities and challenges of working across disciplinary and methodological boundaries. The volume concludes in Chapter 11 with an important call for the

construction of a shared platform to study second language acquisition. Brian MacWhinney argues that further advancement of SLA theory and practice requires a combination of experimental data, a better understanding of how individual differences impact learning, and corpus data that permits the investigation of acquisition patterns. The proposed platform would facilitate this by enabling the collection of substantial amounts of learner data online and by establishing a common protocol on how to share the data – in line with the Child Language Data Exchange System (CHILDES), the central repository for child language data that contributed greatly to our understanding of how children learn language (see Monaghan & Rowland, this volume). The success of such an approach rests on researchers across the world sharing data and agreeing on common protocols for adding and retrieving data.

Acknowledgements

The volume, and the symposium on which it was based, would not have been possible without the essential support and contributions of many colleagues. We are grateful to our symposium presenters and delegates for making it such a successful event, and we thank our authors for submitting excellent manuscripts for this volume. We are indebted to the anonymous peer reviewers, who thoroughly assessed the texts and provided very valuable feedback, also on how to make the contributions accessible and relevant across disciplines. At Language Learning, we are especially grateful to Nick Ellis (General Editor) and Pavel Trofimovich (Journal Editor) for their sustained support throughout this project, and to Izzat Ibrahim for his friendly assistance in the production of the volume. At Lancaster and Tübingen, we are very grateful to Lisa Becker and Abi Hawtin for their help in copy-editing the volume and to Katarina Pardula for her support in organizing the symposium. Finally, we would like to gratefully acknowledge the financial

support of the ESRC Centre for Corpus Approaches to Social Science and Language Learning's Roundtable Grant Program, without which neither the symposium nor the special issue would have been possible.

References

Andringa, S. and Rebuschat, P. (Eds.) (2015). New directions in the study of implicit and explicit learning. Special issue of Studies in Second Language Acquisition, 37(2).

Alexopoulou, T., Michel, M., Murakami, A., Meurers, D. (this volume). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques.

Christiansen, M. H. & Arnon, I. (in press). More than words: The role of multiword sequences in language learning and use. Special issue of Topics in Cognitive Science.

Christiansen, M.H. & Chater, N. (2016). Creating language: Integrating evolution, acquisition, and processing. Cambridge, MA: MIT Press.

Ellis, N. C. (this volume). Cognition, corpora, and computing: Triangulating research in usage-based language learning.

Ellis, N. C., Römer, U. & O'Donnell, M. B. (2016). Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of Construction Grammar. Malden, MA: Wiley-Blackwell.

Hoff, E. (Ed.) (2011). Research methods in child language: A practical guide. Malden, MA: Wiley-Blackwell.

Mackey, A. & Gass, S. M. (2012). Research methods in second language acquisition: A practical guide. Malden, MA: Wiley–Blackwell

MacWhinney, B. (this volume). A shared platform for studying second language acquisition.

Meurers, D. (Ed.) (2009). On the automatic analysis of learner language. Special issue of CALICO Journal 26 (3).

Meurers, D. (2012). Natural language processing and language learning. Encyclopedia of Applied Linguistics (pp. 4193–4205), edited by Carol A. Chapelle. Blackwell.

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin & F. Meunier (Eds.). The Cambridge handbook of learner corpus research (pp. 537–566). Cambridge: Cambridge University Press.

Meurers, D. & Dickinson, M. (this volume). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics.

Monaghan, P. & Rowland, C. (this volume). Combining language corpora with experimental and computational approaches for language acquisition research.

Monaghan, P. & Rebuschat, P. (2017). Aligning implicit learning and statistical learning: Two approaches, one phenomenon. Special issue of Topics in Cognitive Science.

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word-referent mappings. Cognition, 123(1), 133-143. DOI: 10.1016/j.cognition.2011.12.010

Rebuschat, P. & Williams, J. N. (Eds.) (2012). Statistical learning and language acquisition. Berlin: Mouton de Gruyter.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. Language Learning, 63(3), 595-626.

Rebuschat, P. (Ed.) (2015). Implicit and explicit learning of languages. Amsterdam: John Benjamins.

Rebuschat, P., Rohrmeier, M., Hawkins, J. H., Cross, I. (Eds.) (2012). Language and music as cognitive systems. Oxford: Oxford University Press.

Rowland, C. (2013). Understanding child language acquisition. London: Routledge.

Ushioda, E., & Dörnyei, Z. (2012). Motivation. In S. Gass & A. Mackey (Eds.). The Routledge handbook of second language acquisition (pp. 396-409). New York: Routledge.

Williams, J. N. (2012). Working memory. In S. Gass & A. Mackey (Eds.). The Routledge handbook of second language acquisition (pp. 427-441). New York: Routledge.