Learner Corpus Research Meets Second Language Acquisition

Advances in Learner Corpus Research (LCR) and Second Language Acquisition (SLA) have brought these two fast-moving fields significantly closer in recent years. This volume brings together contributions from internationally recognised experts in both LCR and SLA to provide an innovative, cross-collaborative examination of how each area can provide rich insights for the other. Chapters present recent advances in LCR and illustrate in a clear and accessible style how these can be exploited for the study of a broad range of key topics in SLA, such as complexity, tense and aspect, cross-linguistic influence vs. universal processes, phraseology, and variability. It concludes with two commentary chapters written by eminent scholars, one from the perspective of SLA, the other from the perspective of LCR, allowing researchers and students alike to reflect upon the mutually beneficial harmony between the two fields and to link up LCR and SLA research and theory.

Bert Le Bruyn is an Assistant Professor at Utrecht University. He is an expert on semantic language variation with a solid background in corpus linguistics. He received a VENI grant in 2014 for his work focusing on the semantics and L2 acquisition of referentiality, and another grant in 2016 for his corpus-based project with Henriëtte de Swart, Time in Translation.

Magali Paquot is an FNRS research associate at the Centre for English Corpus Linguistics, UCLouvain. She specialises in the use of learner corpora to study key topics in SLA and is particularly interested in methodological issues. She is co-editor in chief of the International Journal of Learner Corpus Research, one of the founding members of the Learner Corpus Research Association, and a member of the IRIS Digital Repository Advisory Group.
THE CAMBRIDGE APPLIED LINGUISTICS SERIES
The authority on cutting-edge Applied Linguistics research
Series Editors 2007–present: Carol A. Chapelle and Susan Hunston
For a complete list of titles please visit: www.cambridge.org
Recent Titles in This Series:
Learner Corpus Research Meets Second Language Acquisition
Bert Le Bruyn and Magali Paquot
Second Language Speech Fluency From Research to Practice
Parvaneh Tavakoli and Clare Wright
Ontologies of English Conceptualising the Language for Learning, Teaching, and Assessment
Edited by Christopher J. Hall and Rachel Wicaksono
Task-Based Language Teaching Theory and Practice
Rod Ellis, Peter Skehan, Shafqefong Li, Natsuko Shintani and Craig Lambert
Feedback in Second Language Writing Contexts and Issues
Edited by Ken Hyland and Fiona Hyland
Language and Television Series A Linguistic Approach to TV Dialogue
Monika Bednarek
Intelligibility, Oral Communication, and the Teaching of Pronunciation
John M. Lewis
Multilingual Education Between Language Learning and Translanguaging
Edited by Jasone Cenoz and Durk Gorter
Learning Vocabulary in Another Language 2nd Edition
J. S. P. Nation
Narrative Research in Applied Linguistics
Edited by Gary Barkhuizen
Teacher Research in Language Teaching A Critical Analysis
Simon Borg
Figurative Language, Genre and Register
Alice Deignan, Jeannette Littlemore and Elena Semin
Exploring ELF Academic English Shaped by Non-native Speakers
Anna Mauranen
Genres across the Disciplines Student Writing in Higher Education
Hilary Nesi and Sheena Gardner
Disciplinary Identities Individuality and Community in Academic Discourse
Ken Hyland
Replication Research in Applied Linguistics
Edited by Graeme Porte,
The Language of Business Meetings
Michael Handford
Reading in a Second Language Moving from Theory to Practice
William Grabe
Modelling and Assessing Vocabulary Knowledge
Edited by Helmut Daller, James Milton and Jeanine Trefmers-Daller
Practice in a Second Language Perspectives from Applied Linguistics and Cognitive Psychology
Edited by Robert M. DeKeyser,
Task-Based Language Education From Theory to Practice
Edited by Kris van den Branden,
Second Language Needs Analysis
Edited by Michael H. Long,
Insights into Second Language Reading A Cross-Linguistic Approach
Keiko Koda
Research Genres Exploration and Applications
John M. Swales
Critical Pedagogies and Language Learning
Edited by Bonny Norton and Kelleen Toohey
Exploring the Dynamics of Second Language Writing
Edited by Barbara Kroll,
Understanding Expertise in Teaching Case Studies of Second Language Teachers
Amy B. M. Tsui
Learner Corpus Research Meets Second Language Acquisition

Edited by

Bert Le Bruyn
UIL-OTS, Utrecht University

Magali Paquot
FNRS – Centre for English Corpus Linguistics, UCLouvain
Contents

List of Figures ix
List of Tables x
List of Contributors xii
Series Editors’ Preface xiii

Learner Corpus Research and Second Language Acquisition: an attempt at bridging the gap 1
Bert Le Bruyn and Magali Paquot

Article Use in Russian and Spanish Learner Writing at CEFR B1 and B2 Levels: Effects of Proficiency, Native Language, and Specificity 10
Tania Ionin and María Belén Díez-Bedmar

L1 Influence vs. Universal Mechanisms: An SLA-Driven Corpus Study on Temporal Expression 39
Valentin Werner, Robert Fuchs, and Sandra Götz

The Interplay between Universal Processes and Cross-Linguistic Influence in the Light of Learner Corpus Data: Examining Shared Features of Non-native Englishes 67
Lea Meriläinen

Exploring Multi-Word Combinations as Measures of Linguistic Accuracy in Second Language Writing 96
Charlene Polio and Hyung-Jo Yoon

Using Syntactic Co-occurrences to Trace Phraseological Complexity Development in Learner Writing: Verb + Object Structures in LONGDALE 122
Magali Paquot, Hubert Naets, and Stefan Th. Gries

Understanding the Long-Term Evolution of L2 Lexical Diversity: The Contribution of a Longitudinal Learner Corpus 148
Nicole Tracy-Ventura, Amanda Huensch, and Rosamond Mitchell
Contents

L2 Developmental Measures from a Dynamic Perspective  
Marjolijn Verspoor, Wander Lowie, and Martijn Wieling  

Exploring Individual Variation in Learner Corpus Research: Methodological Suggestions  
Stefanie Wulff and Stefan Th. Gries  

Building an Oral and Written Learner Corpus of a School Programme: Methodological Issues  
Philippa Bell, Laura Collins, and Emma Marsden  

Commentary: Have Learner Corpus Research and Second Language Acquisition Finally Met?  
Sylviane Granger  

Commentary: An SLA Perspective on Learner Corpus Research  
Florence Myles  

Index
Figures

1 Importance of variables in explaining differences between native and non-native usage 53
2 Results of regression analyses 115
3 The main effect of Topic 136
4 The main effect of OQPT 136
5 Average MI scores for dobj dependencies per year 137
6 Percentage of A2 to C2 texts per year 140
7 Scatterplot of Narrative regression results 164
8 Scatterplot of Interview regression results 165
9 Jorge’s (13-year-old Spanish learner of L2 English) development of negative constructions 176
10 Average holistic rating (five-point scale) of first two (Early) and last two (Late) measurements of the writing samples 183
11 MLTU for individual learners over time in weeks, scaled per individual by z-transformation 184
12 Development of Guiraud for individual learners over time in weeks, scaled per individual with z-transformations 185
13 General effects over time in weeks for the MLTU measure, the Guiraud measure, and their difference 186
14 The effect of LENGTHDIFF 200
15 The effect of PossedNUMBER × PossorAnim 201
16 The effect of SegAltDiff × PossorNumber 203
17 The effect of L1 × PossorComplexity 204
18 The effect of L1 × PossedNumber 205
19 Percentages of non-native-like uses against genitive frequencies 206
20 DEVIATION-values against genitive frequencies 206
21 DEVIATION-values of 10 speakers 207
22 Example of transcription in CHILDES 230
23 The International Corpus of Learner English 250
### Tables

1. Number of texts, number of words, mean and standard deviation in the corpus (by level and L1 background) 18
2. Tagging variables with examples of tags 20
3. Medians and IQRs of correct and incorrect article uses, by level and L1 background 24
4. Normalized data: medians and IQRs (article uses per 100 words, text and level) 25
5. Specific and non-specific indefinites: total number of uses, by level and L1 background 28
6. Specific and non-specific indefinites: medians and IQRs of correct and incorrect uses per 100 words and text, by level and L1 background 28
7. Properties of *the* overuse in place of *a*: number of instances involving additional information about the referent, via modification and/or further mention 30
8. Breakdown of native and non-native corpora 46
9. Instances of PP and SP across corpora 47
10. Independent variables 48
11. Classification of **VARIETYSPECIFICITY** based on the CART analysis 54
12. Corpora examined for embedded inversion 76
13. Indirect questions and embedded inversions 77
14. Embedded inversion in WH-questions and Yes/No-questions 78
15. Corpora examined for preposition omission 82
16. Frequencies of preposition omission 82
17. Summary of measures 105
18. Examples of absent bigrams and trigrams coded as errors or non-errors 108
19. Correlations between individual measures 110
20. Pattern matrix of the factor loadings 111
21. Hierarchical regression analysis to predict error counts 112
22. Hierarchical regression analysis to predict holistic accuracy scores 112
23. Hierarchical regression analysis for traditional error counts’ prediction of holistic accuracy scores 113
List of Tables  xi

24  Number of texts in LONGDALE sample used for this study  127
25  Prompts used in the LONGDALE sample  128
26  Number of EFL learners in LONGDALE sample  129
27  Corpus preprocessing workflow  132
28  Summary results of the final regression model  135
29  Language proficiency development: Individual trajectories in LONGDALE  140
30  Age and years studying L2 of the LANGSNAP 3.0 participants  154
31  Project timeline  155
32  LANGSNAP corpus word counts by task  156
33  Descriptive statistics of types and tokens over time  160
34  Descriptive statistics for $D$ scores over time  160
35  Descriptive statistics for MATTR scores over time  163
36  Regression results  163
37  Composition of the data set  196
38  Observed frequencies of genitives for animate possessors with plural possesseds  202
39  Participants divided by grade level  220
40  Quantity of text produced as a function of language of task instructions  222
41  Number of French L1 words used as a function of language of task instructions  222
42  Topic preferences for the argumentative written task  223
43  Number of words by grade and task type  225
44  Descriptive statistics for lexical diversity  226
45  Descriptive statistics for syntactic complexity  226
46  Post-hoc analyses for D-values  227
47  Post-hoc analyses for mean number of morphemes per utterance  227
Contributors

Philippa Bell (Université du Québec à Montréal, Canada)
Laura Collins (Concordia University, USA)
María Belén Díez-Bedmar (University of Jaén, Spain)
Robert Fuchs (Universität Hamburg, Germany)
Sandra Götz (Justus-Liebig-Universität Giessen, Germany)
Sylviane Granger (UCLouvain, Belgium)
Stefan Th. Gries (UC Santa Barbara, USA & Justus-Liebig-Universität Giessen, Germany)
Amanda Huensch (University of Pittsburgh, USA)
Tania Ionin (University of Illinois at Urbana-Champaign, USA)
Bert Le Bruyn (Utrecht University, The Netherlands)
Wander Lowie (University of Groningen, The Netherlands)
Emma Marsden (University of York, UK)
Lea Meriläinen (University of Eastern Finland, Finland)
Rosamond Mitchell (University of Southampton, UK)
Florence Myles (University of Essex, UK)
Hubert Naets (UCLouvain, Belgium)
Magali Paquot (UCLouvain, Belgium)
Charlene Polio (Michigan State University, USA)
Nicole Tracy-Ventura (West Virginia University, USA)
Marjolein Verspoor (University of Groningen, The Netherlands)
Valentin Werner (Universität Bamberg, Germany)
Martijn Wieling (University of Groningen, USA)
Stefanie Wulff (University of Florida, USA)
Hyung-Jo Yoon (California State University, Northridge, USA)
Series Editors’ Preface

This volume brings into dialogue two areas of central interest to Applied Linguistics: Second Language Acquisition (SLA) and Learner Corpus Research (LCR). As the editors of this volume note in their introduction, the shared concerns of these research communities should suggest productive collaboration between the two, but in the past they have not addressed each other consistently. This can be explained at least partly by differences in their research processes: designing a corpus of learner texts is not the same as compiling a database for SLA research; the questions answered by consulting a learner corpus are not necessarily those asked in SLA.

This volume advances communication between these two areas with studies by experts who address issues of common interest. The chapters explore the challenges that corpus research poses to theoretical assumptions in SLA as well as the methodological challenges SLA offers to LCR. The volume addresses three overarching questions of substantial significance.

The first question asks the extent to which learning a second or additional language follows paths determined by universal mechanisms or by the character of the learner’s L1. Several chapters in the book offer evidence from a variety of learner corpora to find the optimum fit to data between the two positions.

The second question considers how changes in learner proficiency can be measured. Here, necessary developments in corpus research come under scrutiny, in particular the need to automate measures of accuracy and complexity to take account of large amounts of data and the need to build corpora that are subdivided by time. A key issue under this theme is the interaction between individual differences and overall learner trends, and how these might be modelled.

Finally, the volume addresses the question of how the design of corpora and the statistics used to process them may be modified in order to support SLA research. Recent developments in this area are illustrated and evaluated.

The work in this volume offers an excellent example of interdisciplinary research, in which theory and methods from complementary perspectives are used to challenge and support each other. It will be of interest to students and researchers working in both SLA and LCR.

xiii