The origins of learner corpus research go back to the late 1980s, when large electronic collections of written or spoken data started to be collected from foreign/second language learners, with a view to advancing our understanding of the mechanisms of second language acquisition and developing tailor-made pedagogical tools. Engaging with the interdisciplinary nature of this fast-growing field, The Cambridge Handbook of Learner Corpus Research explores the diverse and extensive applications of learner corpora, with 27 chapters written by internationally renowned experts. This comprehensive work is a vital resource for students, teachers and researchers, offering fresh perspectives and a unique overview of the field. With representative studies in each chapter, which provide an essential guide on how to conduct learner corpus research in a wide range of areas, this work is a cutting-edge account of learner corpus collection, annotation, methodology, theory, analysis and applications.

SYLVIANE GRANGER is Professor of English Language and Linguistics at the University of Louvain, and Director of the Centre for English Corpus Linguistics.

GAËTANELLE GILQUIN is a Research Associate with the Belgian National Fund for Scientific Research at the University of Louvain.

FANNY MEUNIER is Professor of English Language, Linguistics and Didactics at the University of Louvain.
CAMBRIDGE HANDBOOKS IN LANGUAGE AND LINGUISTICS

Genuinely broad in scope, each handbook in this series provides a complete state-of-the-field overview of a major sub-discipline within language study and research. Grouped into broad thematic areas, the chapters in each volume encompass the most important issues and topics within each subject, offering a coherent picture of the latest theories and findings. Together, the volumes will build into an integrated overview of the discipline in its entirety.

Published titles

The Cambridge Handbook of Phonology, edited by Paul de Lacy
The Cambridge Handbook of Endangered Languages, edited by Peter K. Austin and Julia Sallabank
The Cambridge Handbook of Sociolinguistics, edited by Rajend Mesthrie
The Cambridge Handbook of Pragmatics, edited by Keith Allan and Kasia M. Jaszczolt
The Cambridge Handbook of Language Policy, edited by Bernard Spolsky
The Cambridge Handbook of Second Language Acquisition, edited by Julia Herschensohn and Martha Young-Scholten
The Cambridge Handbook of Biolinguistics, edited by Cedric Boeckx and Kleanthes K. Grohmann
The Cambridge Handbook of Generative Syntax, edited by Marcel den Dikken
The Cambridge Handbook of Communication Disorders, edited by Louise Cummings
The Cambridge Handbook of Stylistics, edited by Stockwell and Whiteley
The Cambridge Handbook of Linguistic Anthropology, edited by Enfield, Kockelman and Sidnell
The Cambridge Handbook of English Corpus Linguistics, edited by Douglas Biber and Randi Reppen
The Cambridge Handbook of Bilingual Processing, edited by John W. Schwieter
The Cambridge Handbook of Learner Corpus Research, edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier

Further titles planned for the series

The Cambridge Handbook of Morphology, edited by Hippisley and Stump
The Cambridge Handbook of Historical Syntax, edited by Ledgeway and Roberts
The Cambridge Handbook of Formal Semantics, edited by Maria Aloni and Paul Dekker
The Cambridge Handbook of English Historical Linguistics, edited by Merja Kytö and Paivi Pahtsa
The Cambridge Handbook of Linguistic Multicompetence, edited by Li Wei and Vivian Cook
The Cambridge Handbook of Areal Linguistics, edited by Raymond Hickey
The Cambridge Handbook of Learner Corpus Research

Edited by
Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier
University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.
It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107041196

© Cambridge University Press 2015
This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2015
Printed in the United Kingdom by TJ International Ltd, Padstow, Cornwall

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
The Cambridge handbook of learner corpus research / edited by Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier.

pages cm
Includes bibliographical references and index.
ISBN 978-1-107-04119-6 (hbk)
P128.C68C36 2016
410.1’88–dc23
2015014244

ISBN 978-1-107-04119-6 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.
Contents

List of figures .......................... vii
List of tables ............................ viii
List of contributors ...................... x
Acknowledgements ....................... xii

1 Introduction: learner corpus research – past, present and future  Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier  1

Part I Learner corpus design and methodology  7
2 From design to collection of learner corpora  Gaëtanelle Gilquin  9
3 Learner corpus methodology  Marcus Callies  35
4 Learner corpora and psycholinguistics  Philip Durrant and Anna Siyanova-Chanturia  57
5 Annotating learner corpora  Bertus van Rooy  79
6 Speech annotation of learner corpora  Nicolas Ballier and Philippe Martin  107
7 Error annotation systems  Anke Lüdeling and Hagen Hirschmann  135
8 Statistics for learner corpus research  Stefan Th. Gries  159

Part II Analysis of learner language  183
9 Learner corpora and lexis  Tom Cobb and Marlise Horst  185
10 Learner corpora and phraseology  Signe Okseffjell Ebeling and Hilde Hasselgård  207
11 Learner corpora and grammar  Tom Rankin  231
12 Learner corpora and discourse  JoAnne Neff-van Aertselaer  255
13 Learner corpora and pragmatics  Nina Vyatkina and D. Joseph Cunningham  281
Contents

Part III Learner corpus research and second language acquisition 307
14 Second language acquisition theory and learner corpus research Florence Myles 309
15 Transfer and learner corpus research John Osborne 333
16 Learner corpora and formulaic language in second language acquisition research Nick C. Ellis, Rita Simpson-Vlach, Ute Römer, Matthew Brook O’Donnell and Stefanie Wulff 357
17 Developmental patterns in learner corpora Fanny Meunier 379
18 Variability in learner corpora Annelie Ädel 401
19 Learner corpora and learning context Joybrato Mukherjee and Sandra Götz 423

Part IV Learner corpus research and language teaching 443
20 The learner corpus as a pedagogic corpus Angela Chambers 445
21 Learner corpora and language for academic and specific purposes Lynne Flowerdew 465
22 The contribution of learner corpora to reference and instructional materials design Sylviane Granger 485
23 Learner corpora and language testing Fiona Barker, Angeliki Salamoura and Nick Saville 511

Part V Learner corpus research and natural language processing 535
24 Learner corpora and natural language processing Detmar Meurers 537
25 Automatic grammar- and spell-checking for language learners Claudia Leacock, Martin Chodorow and Joel Tetreault 567
26 Learner corpora and automated scoring Derrick Higgins, Chaitanya Ramineni and Klaus Zechner 587
27 Learner corpora and native language identification Scott Jarvis and Magali Paquot 605

Bibliography 629
Author index 725
Subject index 735
Corpus index 743
Software index 747
Figures

5.1 Annotation proposed by Hirschmann et al. (2007) page 96
6.1 Screen capture of LeaP in Praat (Gut 2009) 116
6.2 Learner distinctions of short /ɪ/ and long /i:/ (after Méli 2013) 126
6.3 A sample of ANGLISH using Prosogram as a device for computer-aided decision 131
8.1 Visualisation of the data in Hyland and Milton (1997: 189) 162
8.2 Visualisation of the final model on Hasselgård and Johansson’s (2011: 46) data 167
8.3 The effect of the interaction of PossessorNumber: LengthDifference on the choice of s-genitives in Gries and Wulff (2013) 171
8.4 Dendrogram of Hasselgård and Johansson’s (2011) Table 3 178
16.1 The Zipfian type-token frequency distribution of verb lemmas in the VOL VAC in the native English participants of the ESF project (based on Ellis and Ferreira-Junior 2009a) 376
17.1a Graphical representation of a longitudinal study design 382
17.1b Graphical representation of a pseudo- or quasi-longitudinal study design 382
22.1 ‘Get it right’ box for used to in MEDAL (2007) 491
22.2 Error warning for non-count nouns in the CGE (2006: 339) 492
24.1 An example for the multi-tier representation of the CzeSL corpus (Rosen et al. 2014: 72) 555
27.1 Principles of support vector machines 608
27.2 Principles of discriminant analysis 609
## Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Overall accuracy of POS tags in the 5,168-token sample from the TLE</td>
<td>93</td>
</tr>
<tr>
<td>5.2</td>
<td>Contribution of learner errors to tagging errors of the three taggers</td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>A typology of spoken learner corpora (adapted from Ballier and Martin 2013)</td>
<td>111</td>
</tr>
<tr>
<td>6.2</td>
<td>Representative spoken learner corpora and their features</td>
<td>112</td>
</tr>
<tr>
<td>6.3</td>
<td>A simplified presentation of the phonetic domains for the annotation of spoken learner data</td>
<td>115</td>
</tr>
<tr>
<td>6.4</td>
<td>The phonetics/phonology divide in transcriptions of learner speech at the segmental level</td>
<td>118</td>
</tr>
<tr>
<td>6.5</td>
<td>Some speech software features and their strengths for the analysis of spoken learner data</td>
<td>123</td>
</tr>
<tr>
<td>8.1</td>
<td>Raw/normalised frequencies per million words (pmw) of <em>quite</em> (from Hasselgård and Johansson 2011: 46)</td>
<td>163</td>
</tr>
<tr>
<td>8.2</td>
<td>Laufer and Waldman’s (2011: 660) extended frequency data on V–N collocations</td>
<td>163</td>
</tr>
<tr>
<td>8.3</td>
<td>Hasselgård and Johansson’s (2011) Table 3</td>
<td>178</td>
</tr>
<tr>
<td>9.1</td>
<td>Two calculations of lexical sophistication (k = 1,000-word frequency band)</td>
<td>204</td>
</tr>
<tr>
<td>14.1</td>
<td>Acquisition of syntax and morphology implicationally scaled for all learners (Bonilla 2015: 64)</td>
<td>324</td>
</tr>
<tr>
<td>14.2</td>
<td>Corpus and experimental data collected</td>
<td>327</td>
</tr>
<tr>
<td>15.1</td>
<td>Data for studying transfer between languages</td>
<td>339</td>
</tr>
<tr>
<td>15.2</td>
<td>Corpora used in Neff van Aertselaer (2008)</td>
<td>349</td>
</tr>
<tr>
<td>17.1</td>
<td>Select list of learner corpora collected according to a longitudinal or pseudo-longitudinal design</td>
<td>384</td>
</tr>
<tr>
<td>20.1</td>
<td>Questions for judgemental analysis of CALL appropriateness (Chapelle 2001: 59)</td>
<td>451</td>
</tr>
<tr>
<td>23.1</td>
<td>Metadata in CLC and ICLE</td>
<td>519</td>
</tr>
<tr>
<td>Table</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>24.1</td>
<td>Inter-annotator agreement for selected CzeSL error tags (Rosen et al. 2014: 76)</td>
<td>557</td>
</tr>
<tr>
<td>25.1</td>
<td>Systematic spelling errors (adapted from Rimrott and Heift 2008)</td>
<td>572</td>
</tr>
<tr>
<td>25.2</td>
<td>Spelling-error categorisation (adapted from Bestgen and Granger 2011)</td>
<td>577</td>
</tr>
<tr>
<td>26.1</td>
<td>Distribution of test takers’ first languages for TPO training and TPO evaluation data sets</td>
<td>598</td>
</tr>
<tr>
<td>26.2</td>
<td>Performance evaluation of multiple regression scoring model on the TPO evaluation data</td>
<td>600</td>
</tr>
<tr>
<td>27.1</td>
<td>A summary of the features used in Tetreault et al. (2012)</td>
<td>620</td>
</tr>
</tbody>
</table>
Contributors

Annelie Ädel Dalarna University, Sweden
Nicolas Ballier Université Paris Diderot, Sorbonne Paris Cité, France
Fiona Barker Cambridge English Language Assessment, UK
Marcus Callies Universität Bremen, Germany
Angela Chambers University of Limerick, Ireland
Martin Chodorow City University of New York, USA
Tom Cobb Université du Québec à Montréal, Canada
D. Joseph Cunningham Georgetown University, USA
Philip Durrant University of Exeter, UK
Signe Oksefjell Ebeling University of Oslo, Norway
Nick C. Ellis University of Michigan, USA
Lynne Flowerdew Hong Kong University of Science and Technology, SAR China
Gaëtanelle Gilquin Université catholique de Louvain – FNRS, Belgium
Sandra Götz Justus Liebig University Giessen, Germany
Sylviane Granger Université catholique de Louvain, Belgium
Stefan Th. Gries University of California Santa Barbara, USA
Hilde Hasselgård University of Oslo, Norway
Derrick Higgins Civis Analytics, USA
Hagen Hirschmann Humboldt-Universität zu Berlin, Germany
Marlise Horst Concordia University, Canada
Scott Jarvis Ohio University, USA
Claudia Leacock McGraw-Hill Education/CTB, USA
Anke Lüdeling Humboldt-Universität zu Berlin, Germany
Philippe Martin Université Paris Diderot, Sorbonne Paris Cité, France
Fanny Meunier Université catholique de Louvain, Belgium
Detmar Meurers Eberhard Karls Universität Tübingen, Germany
Joybrato Mukherjee Justus Liebig University Giessen, Germany
Florence Myles University of Essex, UK
JoAnne Neff-van Aertselae Complutense University of Madrid, Spain
List of contributors

Matthew Brook O’Donnell  University of Pennsylvania, USA
John Osborne  Université Savoie Mont Blanc, France
Magali Paquot  Université catholique de Louvain, Belgium
Chaitanya Ramineni  Educational Testing Service, USA
Tom Rankin  Vienna University of Economics and Business, Austria
Ute Römer  Georgia State University, USA
Angeliki Salamoura  Cambridge English Language Assessment, UK
Nick Saville  Cambridge English Language Assessment, UK
Rita Simpson-Vlach  University of Michigan, USA
Anna Siyanova-Chanturia  Victoria University of Wellington, New Zealand
Joel Tetreault  Yahoo Labs, USA
Bertus van Rooy  North-West University, South Africa
Nina Vyatkina  University of Kansas, USA
Stefanie Wulff  University of Florida, USA
Klaus Zechner  Educational Testing Service, USA
Acknowledgements

We wish to thank Andrew Winnard for the confidence he has shown in entrusting us with the task of editing this volume and Bethany Gaunt for her help at the production stage. We are extremely grateful to all the contributors to the volume for their interest and willingness to conform to our guidelines and deadlines. We would also like to thank the many reviewers for their insightful comments and suggestions, which have greatly contributed to enhancing the quality of the volume. Special thanks are owed to Maité Dupont for her invaluable help at various stages in the preparation of the manuscript.

The editors and authors gratefully acknowledge the permission granted to reproduce the following copyright material:


Figure 22.1. From Macmillan English Dictionary for Advanced Learners, 2nd edn. 2007. Oxford: Macmillan Education. Reproduced by permission of Macmillan Education.
Acknowledgements


Every effort has been made to obtain the necessary permissions for the use of copyright material. We apologise for any inadvertent omissions and would be pleased to make the appropriate acknowledgements in future reprints or editions of this book if notified by the copyright holders.