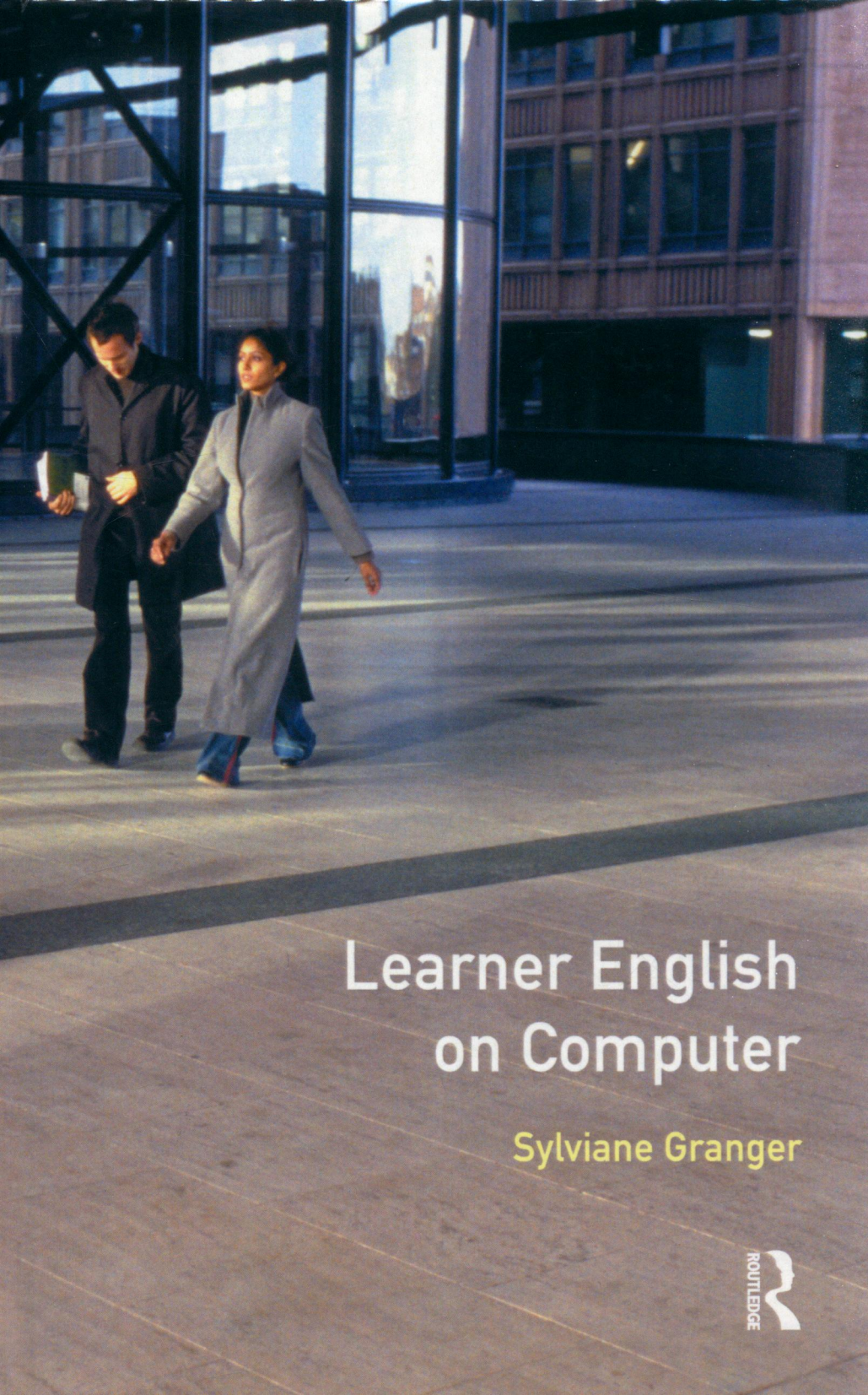

Learner English on Computer

Sylviane Granger

ROUTLEDGE

an **informa** business

ISBN 978-1-138-16275-4

Routledge
Taylor & Francis Group
www.routledge.com

Learner English on Computer

Granger

Routledge

Learner English on Computer

Edited by SYLVIANE GRANGER

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 1998 by Addison Wesley Longman Limited

2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
711 Third Avenue, New York, NY 10017, USA

Routledge is an imprint of the Taylor & Francis Group, an informa business

First issued in hardback 2016

ISBN 13: 978-0-582-29883-5 (pbk)
ISBN 13: 978-1-138-16275-4 (hbk)

British Library Cataloguing-in-Publication Data

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

Library of Congress Cataloging-in-Publication Data

Learner English on computer / edited by Sylviane Granger.
p. cm. — (Studies in language and linguistics)
Includes bibliographical references and index.
ISBN 0-582-29883-0 (pbk.)
1. English language—Study and teaching—Foreign speakers—Data processing. 2. English language—Errors of usage—Data processing.
3. English language—Computer-assisted instruction.
4. Computational linguistics. I. Granger, Sylviane, 1951– .
II. Series: Studies in language and linguistics (London, England)
PE1128.A2L364 1998
428'.00285—dc21 97-36716
CIP

Set by 35 in 9/11pt Palatino

Learner English on Computer

Studies in language and linguistics

General editors: GEOFFREY LEECH *Lancaster University*
and JENNY THOMAS *Bangor University*

Already published:

A Dictionary of Stylistics
KATIE WALES

The Communicative Competence of Young Children
SUSAN H. FOSTER

Linguistic Purism
GEORGE THOMAS

Women, Men and Language
Second edition
JENNIFER COATES

Lexical Ambiguity in Poetry
SOON PENG SU

Understanding Metaphor in Literature
GERARD STEEN

English Spelling and the Computer
ROGER MITTON

Conversational Routines in English: Convention and Creativity
KARIN AIJMER

Learner English on Computer
Edited by SYLVIANE GRANGER

In the final analysis if linguistics is not about language as it is actually spoken and written by human beings, then it is about nothing at all.

Michael Stubbs

Contributors

Jan Aarts, Department of Language and Speech, University of Nijmegen, Netherlands

Bengt Altenberg, Department of English, Lund University, Sweden

Doug Biber, Department of English, Northern Arizona University, USA

Sylvie De Cock, Centre for English Corpus Linguistics, Université Catholique de Louvain, Belgium

Adam Gadsby, ELT Division, Addison Wesley Longman, England

Patrick Gillard, ELT Division, Addison Wesley Longman, England

Przemysław Kaszubski, School of English, Adam Mickiewicz University, Poland

Geoffrey Leech, Department of Linguistics and Modern English Language, Lancaster University, England

Gunter Lorenz, Didaktik des Englischen, University of Augsburg, Germany

Tony McEnery, Department of Linguistics and Modern English Language, Lancaster University, England

Fanny Meunier, Centre for English Corpus Linguistics, Université Catholique de Louvain, Belgium

John Milton, Language Centre, Hong Kong University of Science and Technology, Hong Kong

Stephanie Petch-Tyson, Centre for English Corpus Linguistics, Université Catholique de Louvain, Belgium

Paul Rayson, Department of Computing, Lancaster University, England

Randi Reppen, Department of English, Northern Arizona University, USA

Håkan Ringbom, Department of English, Åbo Akademi University, Finland

Marie Tapper, Department of English, Lund University, Sweden

Christopher Tribble, Institute of English Language Education, Lancaster University/Centre for Applied Language Studies, University of Reading, England

Tuija Virtanen, Department of English, Åbo Akademi University, Finland

Editor's acknowledgements

I am indebted to the Université Catholique de Louvain for granting me a year's sabbatical in 1995–96, a year which was crucial for the development of my views on computer learner corpus research. I have particularly good memories of the time I spent carrying out research at the Universities of Lancaster and Nijmegen and would like to thank G. Leech and J. Aarts for the warm welcome I received in both places. I am especially grateful to the Fonds National de la Recherche Scientifique, the Commissariat Général aux Relations Internationales and the British Council for their financial support during this time. I would also like to thank the contributors to this volume for their diligence in keeping to deadlines and their patience in complying with my editorial demands. Special thanks go to Sylvie De Cock and Estelle Dagneaux for their meticulous examination of the typescript, and to Stephanie Petch-Tyson, who provided numerous suggestions for improvement. Finally, I would like to thank my family – my husband Guy and my two sons, David and Tony, for putting up with my year-long exile with only a minimum of bad grace!

Publisher's acknowledgements

We are indebted to Cambridge University Press for permission to use our Table 13.1 'A typical example of vocabulary tabulation' from *Study Writing* by L. Hamp-Lyons and B. Heasley (1987: 71); and Addison Wesley Longman for our Table 13.2 'A vocabulary table with graded stylistic information' from J. Arnold and J. Harmer *Advanced Writing Skills* (1978: 57).

List of abbreviations

BNC	British National Corpus
CA	Contrastive Analysis
CALL	Computer-Assisted Language Learning
CIA	Contrastive Interlanguage Analysis
CLC	Computer Learner Corpus
DDL	Data-Driven Learning
EA	Error Analysis
EFL	English as a Foreign Language
ELT	English Language Teaching
ESL	English as a Second Language
HKUST	Honk Kong University of Science and Technology
ICE	International Corpus of English
ICLE	International Corpus of Learner English
IL	Interlanguage
KIIC	Key Item In Context
KWIC	Key Word In Context
LD	Lexical Density
LLC	Longman Learners' Corpus
LOB Corpus	Lancaster-Oslo/Bergen Corpus
LOCNESS	Louvain Corpus of Native English Essays
MIQ	Multiple-Items Queries
MSL	Mean Sentence Length
MTTR	Mean Type/Token Ratio
MTUL	Mean T-Unit Length
NL	Native Language
NS	Native Speaker
NNS	Non-Native Speaker
POS	Part Of Speech
SGML	Standard Generalized Markup Language
SLA	Second Language Acquisition
UG	Universal Grammar

Preface

Geoffrey Leech

Learner corpora: what they are and what can be done with them

This is the first book devoted to the idea of collecting a corpus, or computer textual database, of the language produced by foreign language learners: a collection known as a learner corpus. To begin with a hypothetical but realistic example, let us suppose that higher education teacher X, in a non-English speaking country, teaches English to her students every week, and every so often sets them essays to write, or other written tasks in English. Now, instead of returning those essays to students with comments and a sigh of relief, she stores the essays (of course with the students' permission) in her computer, and is gradually building up, week by week, a larger and more representative collection of her students' work. Helped by computer tools such as a concordance package, she can extract data and frequency information from this 'corpus', and can analyse her students' progress as a group in some depth. More significant (since teacher X is also interested in building up a research profile) are the research questions which open up once the corpus is in existence; for example:

- What linguistic features in the target language do the learners in question use significantly more often ('overuse') or less often ('underuse') than native speakers do?
- How far is the target language behaviour of the learners influenced by their native language (NL transfer)?
- In which areas do they tend to use 'avoidance strategies', failing to exploit the full range of the target language's expressive possibilities?
- In which areas do they appear to achieve native-like or non-native-like performance?
- What (in order of frequency) are the chief areas of non-native-like linguistic performance which learners in country A suffer from and need particular help with?

To some extent, teacher X's interest in such questions may be directed towards improving her own teaching practices: for example, she will be able to save time where the students experience no difficulty, and concentrate remedial work on areas where more help is patently needed. In other words, she will be able to tailor teaching to need. To some extent, however, her interest will also be directed towards a more collaborative mode of research with teachers and researchers in other institutions and in other countries, who are collecting the data of their students, just as she is collecting the data of hers. Such a collaboration is needed if we are to answer more generic questions such as:

- What are the particular areas of overuse, underuse and error which native speakers of language A are prone to in learning target language T, as contrasted with native speakers of languages B, C, D . . . ?
- What, in general, is the proportion of non-native target language behaviour (overuse, underuse, error) peculiar to native speakers of language A, as opposed to such behaviour which is shared by all learners of the language, whatever their mother tongue?

It appears odd that SLA research has not yet provided a clear answer to these questions, especially to the second one, which concerns the influence of the native language on learning, and which has obvious implications for how languages should be successfully learned. The study of learner corpora for the first time provides for a research programme which will lead to its being answered.

There are many refinements and elaborations of such a research programme which can be envisaged – such as the collection of corpora of the same students at different stages of learning (a longitudinal learner corpus, in fact), or the collection of a (preferably longitudinal) corpus of the data derived from individual learners rather than from a homogeneous group. These refinements lie largely in the future. But, to answer questions such as those above, for the time being, we may look forward to the success of an international learner corpus programme which entails collecting comparable data from comparable learner groups, each of which consists of speakers of a different native language: e.g. a corpus of English produced by NSs (native speakers) of French, a similar corpus produced by NSs of Chinese, a similar corpus produced by NSs of Polish, and so on. To complete the international corpus design we also need a comparable corpus, insofar as it can be obtained, of NSs of the target language, English, as a standard of comparison, or norm, against which to measure the characteristics of the learner corpora. This is indeed the design of the International Corpus of Learner English (ICLE), founded and coordinated by Sylviane Granger, the editor of this volume.

Much of this book is devoted to the first fruits of the ICLE project, which is already beginning to yield findings of considerable interest. We should also mention the role of other large learner corpus projects now

coming to fruition: the 10-million-word Longman Learners' Corpus, rich in variety of mother tongues and levels of learner attainment, on which Gillard and Gadsby report in Chapter 12; also the homogeneous corpus of 10 million words (all from Chinese-speaking learners) collected in the HKUST project on which Milton reports in Chapter 14.

The background: corpus research and research into language learning

Rather dramatically, we may claim that the concept of a learner corpus is an idea 'whose hour has come'. Corpus linguistics (which nowadays means 'computer corpus linguistics') is a relatively new branch of linguistics which has been gathering momentum over the past 30 years, as computers have grown enormously in storage and processing ability. Its influence has spread into many branches of language research, but has been rather slow to gain a foothold in the educational sphere, for two reasons, one practical and the other theoretical. The first reason is that computers cost money, and computer-based research requires a concentration of human resources and equipment which is not easily available to those working in areas such as English language teaching (ELT) and second language acquisition (SLA), where resources are scarce. Education is the Cinderella of the academic world, particularly in language learning, where research projects are usually funded inadequately, if they are funded at all.

The second reason is that the intellectual climate current in applied linguistics over the past 20 years has not lent itself to the kinds of empirical methods that corpus linguistics fosters. If, to dramatise again, we characterise the theme of this book as 'SLA meets corpus linguistics', this is not likely to be a meeting of unalloyed joy and goodwill. Rather, it may well be an encounter marked by some suspicion and misunderstanding. Why this is so is not immediately evident. It might seem that a large and carefully compiled database of learner's language is going to be a useful resource for anyone wanting to find out how people learn languages, and how they can be helped to learn them better. After all, the notion of interlanguage (IL) research rests on the principle that understanding language learning means understanding the intermediate approximative language systems which learners, as learners, progressively acquire. And how, it may be asked, could we better study such interlanguage knowledge than by studying the language which learners produce? Surely this is the only really hard evidence we have of what progress learners are making or failing to make?

What may seem an eminently sensible course of action to a layperson does not necessarily commend itself to the academic world. Two mutually opposed intellectual currents have taken the focus of attention away