DE GRUYTER
MOUTON

Geoffrey Sampson,
Anna Babarczy

GRAMMAR WITHOUT GRAMMATICALITY

GROWTH AND LIMITS OF GRAMMATICAL PRECISION

TRENDS IN LINGUISTICS

DE
G

Geoffrey Sampson
Anna Babarczy

Grammar Without Grammaticality

Growth and Limits of Grammatical Precision

DE GRUYTER
MOUTON

ISBN 978-3-11-028977-0
e-ISBN 978-3-11-029001-1
ISSN 1861-4302

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Typesetting: RoyalStandard, Hong Kong
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Geoffrey Sampson and Anna Babarczy
Grammar Without Grammaticality

Trends in Linguistics Studies and Monographs

Volume 254

Preface

This book sets out to explain and give evidence for a distinctive view of the nature of human language, a view which contrasts sharply with assumptions that are taken for granted by many or most linguists publishing today. The book has had a long gestation, but the bulk of it emerges from collaborative work carried out at Sussex University in England, where one of us (Sampson) was Professor of Natural Language Engineering, and the other (Babarczy) was assisting him as Research Fellow.

Some chapters of the book are primarily conceptual in nature, aiming to make as clear and explicit as possible what we think language is like, where we disagree with others, and how we answer their objections to our point of view. Other chapters go into details of the empirical, quantitative researches that have led us to those conceptual conclusions. (Some chapters mix these modes of discussion.)

We realize, of course, that many readers are more comfortable with conceptual discussion than with the nitty-gritty of numbers and statistics. Some might have preferred us to leave out the second kind of material, feeling that they could trust us to get that part right and give them just the conclusions which emerge. We believe that it is important not to omit the quantitative material. There is a style of linguistic publication which is quite influential nowadays, that might be described as a pastiche of scientific writing. General statements are expressed in words, but here and there one sees a sprinkling of algebraic notations or other formalisms which create an impression of hard scientific analysis underlying the generalities – only, if you press, that analysis never seems to be forthcoming. (The kind of intellectual charlatanry skewered by Alan Sokal and Jean Bricmont (1997) is familiar in modern linguistics, although it happened that none of the individual academics Sokal and Bricmont discussed was a linguist.) We want it to be clear to readers that our work is not like that. If some passages of the book seem hard going, we hope that readers will be patient with them for the sake of receiving a sense of the detailed research which, in our eyes at least, has entitled us to assert the conclusions we have drawn.

If some parts of the book may seem dry, in another respect the book may offer *more* human interest than the average work on linguistics. Since language is an aspect of human behaviour, and an important one, any fresh insights on the nature and structure of language will be of some interest to human beings – but often those insights are rather abstract. Linguistics is frequently seen as a subject which does not have the weighty practical or moral implications

for society that commonly attach to fundamental theorizing in other human sciences, such as economics or politics. But we believe that the contrasting views about language structure discussed in this book – ours, and the current majority view with which we disagree – do indeed entail contrasting consequences for how societies should be organized and lives should be led. The central aim of our book is to establish a truth about language, but professional academics have a duty to society to spend some time discussing the practical implications of their work. We take this issue up in our final chapter.

Some of our chapters are based on papers previously published in academic journals or conference proceedings, with redrafting and additional material to fit them into a volume that reads coherently as a whole. Only in this way can separate small-scale investigations be appreciated as contributing to a single novel and consistent model of the nature of language. Other chapters are newly published here.

Where material was published previously, this was sometimes done under one of our names, sometimes the other, and sometimes we published as co-authors. That creates a problem about pronouns for this book. It would be tedious for readers if we chopped and changed between "I" and "we", depending on details of the original published version of different passages here. Furthermore, even when Sampson published under his sole name, often his writing drew on work done by Babarczy.

Luckily, in English the authorial "we" is vague between singular and plural: for the sake of readability, we shall be "we" throughout this book, even in contexts where the reference is obviously singular. For instance, at the beginning of chapter 1 we mention "our copy" of an old book by John Meiklejohn: of course the book belongs to only one of us, and we have separate bookshelves, now in different countries – but we spare the reader the need to wade through irrelevant references to "one of us", "the other co-author", and so forth.

The exception is that at a few points it matters which of us is which. For instance, in chapter 2 we describe an experiment in which we played separate roles and where it could have been relevant that GRS but not AB is an English native speaker. In such contexts we identify ourselves by initials, as we have just done. Broadly, as will commonly be the case when a senior academic collaborates with a full-time researcher, the bulk of the hard quantitative work tended to be done by AB, and the general conclusions, together with the ways to make them engage with the ongoing context of debate that constitutes the discipline of linguistics, tended to be settled by GRS. But each of us contributed at each level; the book is a genuine collaboration.

The programme of research from which this book emerges has centred round various English-language "corpora" – samples, held in electronic form,

of different genres of the language. Since many chapters will refer to one or another of these resources, we list here those which are relevant to our book, with access details. (Sampson and McCarthy 2004 contains papers discussing these and many other corpora, and illustrating various ways in which they are used for linguistic research.)

The "raw" English corpora mainly relevant to our work – where "raw" means that the electronic files contain the wording of the language samples but little more[1] – are the *Brown Corpus*, the *Lancaster–Oslo/Bergen (LOB) Corpus*, and the *British National Corpus*. Brown and LOB are the resources which first enabled corpus linguistics to get off the ground: they are matched collections of one million words each of published writing in, respectively, American and British English of the 1960s. The British National Corpus is a more recent collection comprising one hundred million words of British English, of which ninety million is written language but ten million is spoken, and of the latter about 4·2 million words are "demographically sampled" speech – broadly, the spontaneous conversational talk of a cross-section of the UK population. For Brown and LOB Corpora see e.g. <khnt.hit.uib.no/icame/manuals/>; for the British National Corpus, <www.natcorp.ox.ac.uk>.

Initially as a member of the corpus research group at Lancaster University led by Geoffrey Leech, and later as leader of his own research group of which AB became a key member, GRS has used these and similar materials in order to generate a maximally refined and comprehensive structural taxonomy for written and spoken English, attempting to do for the grammar of the English language something akin to what Linnaeus in the eighteenth century did for plant life. (On this analogy, cf. Sampson 2000a.) In practice this entailed using subsamples of the raw corpora as testbeds to uncover and resolve shortcomings in a scheme of structural annotation for the language, making the scheme ever more adequate to record and classify the grammatical detail that occurs in samples of the language in practice – a process which can never be complete. The resulting annotated subcorpora, or "treebanks", then became research resources in their own right:

- the SUSANNE Corpus comprises 64 files drawn from four genre categories of the Brown Corpus, for a total of about 130,000 words of written American English
- the CHRISTINE Corpus comprises about 80,500 words, plus many ums and ers and the like, of spoken British English (taken from the demographically-sampled speech section of the British National Corpus)

---

1 In fact, for these resources, recent versions of the files do also include wordtags, i.e. part-of-speech classifications of the individual words.

- the LUCY Corpus comprises 165,000 words of written British English representing a spectrum of genres ranging from polished published writing to the writing of young children.

(The name SUSANNE was chosen to stand for "surface and underlying structural analyses of natural English", and also to celebrate links with the life of St Susanna. The names CHRISTINE and LUCY likewise refer to saints whose reputed careers were in different ways appropriate to the language genres in question.)

These research resources are freely available for others to download and work with (and many researchers internationally have done and are doing so). For details of access, and documentation of these resources, see <www.grsampson.net/Resources.html>.

The structural annotation scheme which emerged from this programme of research, called the SUSANNE Scheme after the earliest of the treebanks listed, was published in the form of a 500-page book (Sampson 1995). Various refinements to that published scheme which derived from work since 1995, particularly to handle the special features of spontaneous speech, are covered in the electronic documentation files for the corpora.

Geoffrey Sampson
University of South Africa

Anna Babarczy
Budapest University of Technology and Economics

Acknowledgements

We thank Alan Morris for his work on the electronic resources used in the research discussed in this book, and Anna Rahman for permission to use the material on which chapter 6 is based, of which she was a co-author. For discussion of various of our ideas, we are grateful to Anders Ahlqvist, John Carroll, Gerald Gazdar, Adam Kilgarriff, and the late Larry Trask, as well as numerous participants at conferences and anonymous referees for journals. We ask anyone whose name we have overlooked to forgive us.

Much of the research presented here was sponsored by the Economic and Social Research Council (UK), whose support we gratefully acknowledge.

The origin of the various chapters is as follows:

Chapter 1 is new material.

Chapter 2 is based on a paper in the *Journal of Natural Language Engineering*, vol. 14, pp. 471–494, 2008, reprinted with permission of Cambridge University Press.

Chapter 3 is based on a paper in the *Proceedings of the Workshop on Linguistically Interpreted Corpora, LINC–2000*, Luxemburg, 6 Aug 2000 (ed. by Anne Abeillé et al.), pp. 28–34.

Chapter 4 is based on a "target paper" which appeared in *Corpus Linguistics and Linguistic Theory*, vol. 3, pp. 1–32 and 111–129, 2007.

Chapter 5 is new material.

Chapter 6 is based on a paper in J. M. Kirk (ed.), *Corpora Galore*, 2000, pp. 295–311, reprinted by permission of Rodopi of Amsterdam.

Chapter 7 is based on a chapter in *Empirical Linguistics*, © Geoffrey Sampson 2001, reprinted by permission of Continuum Publishing, an imprint of Bloomsbury Publishing plc.

Chapter 8 is based on a paper in Sylviane Granger and Stephanie Petch-Tyson (eds), *Extending the Scope of Corpus-Based Research*, 2003, pp. 177–193, reprinted by permission of Rodopi of Amsterdam.

Chapter 9 is new material.

Chapter 10 is based on a paper in Mickael Suominen et al. (eds), *A Man of Measure: Festschrift in honour of Fred Karlsson on his 60th birthday*, 2006, pp. 362–374, reprinted with permission of the Linguistic Association of Finland.

Chapter 11 is based on a paper in *English Language and Linguistics*, vol. 6, pp. 17–30, 2002, reprinted with permission of Cambridge University Press.

Chapter 12 is based on a paper in the *Journal of Natural Language Engineering*, vol. 9, pp. 365–380, 2003, reprinted with permission of Cambridge University Press.

Chapter 13 is based on a paper in the *International Journal of Corpus Linguistics*, vol. 10, pp. 15–36, 2005.

Chapter 14 is new material.

Chapter 15 is based on a keynote address to the International Association for Dialogue Analysis workshop (IADA 2006), Mainz, September 2006, published in Marion Grein and Edda Weigand (eds), *Dialogue and Culture*, pp. 3–25, 2007, and reprinted with kind permission of John Benjamins Publishing Company.

List of figures

List of tables

Table of contents