Corpus Pragmatics

Corpus linguistics is a long-established method which uses authentic language data, stored in extensive computer corpora, as the basis for linguistic research. Moving away from the traditional intuitive approach to linguistics, which used made-up examples, corpus linguistics has made a significant contribution to all areas of the field. Until very recently, corpus linguistics focused almost exclusively on syntax and the lexicon; however, corpus-based approaches to the other subfields of linguistics are now rapidly emerging, and this is the first handbook on corpus pragmatics as a field. Bringing together a team of leading scholars from around the world, this handbook looks at how the use of corpus data has informed research into different key aspects of pragmatics, including pragmatic principles, pragmatic markers, evaluation, reference, speech acts and conversational organization.

KARIN AIJMER is Professor Emerita in English Linguistics in the Department of Languages and Literature at the University of Gothenburg, Sweden.

CHRISTOPH RÜHLEMMANN is a Researcher at Philipps University Marburg, Germany.
Corpus Pragmatics

A Handbook

Edited by

Karin Aijmer
University of Gothenburg

and

Christoph Rühlemann
Philipps University Marburg
Contents

List of figures
List of tables
Notes on contributors

Introduction
Corpus pragmatics: laying the foundations
CHRISTOPH RÜHLEMANN AND KARIN AIJMER

Part I: Corpora and speech acts
1 Speech acts: a synchronic perspective
   PAULA GARCIA MCALLISTER
2 Speech acts: a diachronic perspective
   THOMAS KOHNEN
3 Speech act annotation
   MARTIN WEISSER

Part II: Corpora and pragmatic principles
4 Processibility
   GUNThER KALTENBÖCK
5 Relevance
   GISLE ANDERSEN
6 Politeness
   GIULIANA DIANI

Part III: Corpora and pragmatic markers
7 Pragmatic markers
   KARIN AIJMER
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Stance markers</td>
<td>219</td>
</tr>
<tr>
<td></td>
<td>Bethany Gray and Douglas Biber</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Interjections</td>
<td>249</td>
</tr>
<tr>
<td></td>
<td>Neal R. Norrick</td>
<td></td>
</tr>
<tr>
<td>Part IV: Corpora and evaluation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Evaluative prosody</td>
<td>279</td>
</tr>
<tr>
<td></td>
<td>Alan Partington</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Tails</td>
<td>304</td>
</tr>
<tr>
<td></td>
<td>Ivor Timmis</td>
<td></td>
</tr>
<tr>
<td>Part V: Corpora and reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Deixis</td>
<td>331</td>
</tr>
<tr>
<td></td>
<td>Christoph Rühlemann and Matthew Brook O'Donnell</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Vagueness</td>
<td>360</td>
</tr>
<tr>
<td></td>
<td>Winnie Cheng and Anne O'Keeffe</td>
<td></td>
</tr>
<tr>
<td>Part VI: Corpora and turn-taking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Turn management and the fillers <em>uh</em> and <em>um</em></td>
<td>381</td>
</tr>
<tr>
<td></td>
<td>Gunnel Tottie</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Turn management and backchannels</td>
<td>408</td>
</tr>
<tr>
<td></td>
<td>Pam Peters and Deanna Wong</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Co-constructed turn-taking</td>
<td>430</td>
</tr>
<tr>
<td></td>
<td>Brian Clancy and Michael McCarthy</td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>454</td>
</tr>
</tbody>
</table>
List of figures

Figure 0.1  Horizontal-reading methodology in pragmatics (P).  page 3
Figure 0.2  Vertical-reading methodology in corpus linguistics (CL).  8
Figure 0.3  Integrated-reading methodology in corpus pragmatics (CP).  12
Figure 3.1  XML representation of dialogue structure.  87
Figure 3.2  Tags + speech acts with a frequency of over 40 occurrences.  106
Figure 4.1  Clause-initial I think in COHA (per million words).  126
Figure 4.2  It may be that in COHA (per million words).  127
Figure 4.3  Clause-initial maybe in COHA (per million words).  127
Figure 4.4  Relative change of it may be that, clause-initial Maybe/Perhaps within the set in COHA (Wilson confidence intervals for $p < 0.05$).  128
Figure 4.5  Clause-initial (it) turns out (that) and non-initial (it) turns out in COHA (per million words).  129
Figure 4.6  Relative change of adverbial uses of (it) turns out compared to all uses of turns out in COHA (Wilson confidence intervals for $p < 0.05$; Wallis 2009).  130
Figure 4.7  Matrix clause uses of The thing is in COHA (per million words).  132
Figure 4.8  Relative change of (The) thing is (that) + clause compared to all uses of thing is in COHA (Wilson confidence intervals for $p < 0.05$; Wallis 2009).  132
Figure 4.9  It’s just that matrix clauses in COHA (per million words).  137
Figure 4.10  Sentential adverb Only in COHA (per million words).  137
Figure 4.11  Increasing use of It’s just that as a proportion of the set {it’s just that, sentence-initial Only} in COHA (Wilson confidence intervals for $p < 0.05$; Wallis 2009).  137
Figure 8.1  Stance adjective + complement clauses across register.  230
viii List of figures

Figure 8.2 Stance meanings expressed by adjective + complement clauses: *that*-clauses. 231
Figure 8.3 Stance meanings expressed by adjective + complement clauses: *to*-clauses. 233
Figure 8.4 Stance noun structures across registers. 234
Figure 8.5 Stance meanings expressed by adjective + complement clauses: *that*-clauses. 235
Figure 9.1 Histograms of positions of *yeah*, *oh* and *well* in utterances in the non-narrative conversational components in the Narrative Corpus. 259
Figure 9.2 Percentages of occurrence of *well*, *oh* and *yeah* in first position in dialogue and utterances in the Narrative Corpus. 264
Figure 10.1 Evaluative meaning expressed in terms of prototypicality. 290
Figure 12.1 Componential structure of texts in the Narrative Corpus. 343
Figure 12.2 Comparison of estimated probability densities for positions of *this* and *these* in textual components (CPR, pre-narrative; CNN, narrative; CPO, post-narrative). 345
Figure 12.3 Frequencies of reference patterns of *this* in narrative-initial utterances (CNI). 349
Figure 12.4 Breakdown of frequencies of reference patterns for *this* in CNI by semantic group. 351
Figure 13.1 HKCSE (conversation) and LCIE: comparison of thirteen VL type 1 sub-types. 369
Figure 13.2 Breakdown of semantic categories used with ‘*about + n*’, per 100 occurrences in HKC, NEC, and LCIE. 371
Figure 14.1 Frequency of UHM per 1,000 words in the fourteen texts of SBC1, from Tottie (2014). 388
Figure 15.1 Measuring the duration of a backchannel (*yeah*), and the intervals before and after it. 413
Figure 15.2 Lengths of speaker turns and numbers of backchannels received. 416
Figure 15.3 Median durations of *yeah* and *mm* and median lengths of the preceding interval, when occurring first and finally in a string. 419
Figure 15.4 Median durations of *yeah* and *mm* and median lengths of the preceding interval, as standalones and first in a string. 420
Figure 15.5 Median durations of *yeah* and *mm* and the intervals before and after, as standalones and as string-final utterances. 422
List of figures

Figure 15.6 Median intervals before and after standalone yeah and mm. 424
Figure 15A.1 Scatter plots showing the relationship between the backchanneler-managed interval before yeah and mm and the speaker-controlled interval after them. 429
List of tables

Table 0.1 Top five collocates of ‘corpus’ in the BNC.  
page 8

Table 1.1 Speech act definitions with examples from corpus extracts.  
33

Table 1.2 Number of words and utterances per situation type in the TOEFL Corpus.  
36

Table 1.3 Raw and normalized frequency counts of speech acts in each situation type in the TOEFL Corpus.  
38

Table 2.1 Frequency (per 10,000 words) and incidence of performatives in the three historical corpora.  
69

Table 2.2 Distribution of the performatives across genres (incidence and proportions in per cent) in the three historical corpora.  
69

Table 2.3 Frequency (per 10,000 words) and incidence of six classes of performatives in the three historical corpora.  
70

Table 2A.1 Distribution of the different classes of performatives across the genres in the three historical corpora.  
80

Table 2A.2 Distribution of speech act verbs across classes of performatives in the three historical corpora.  
81

Table 3A.1 Comparison of speech act labels in DART, DAMSL, and SWDB-DAMSL.  
112

Table 6.1 Type and distribution of mitigation strategies for criticism in English book review articles, from the EN_HIBRA corpus.  
179

Table 6.2 Most frequent hedges in the English corpus (EN_HIBRA).  
180

Table 6.3 Type and distribution of mitigation strategies for criticism in Italian book review articles, from the IT_HIBRA corpus.  
183

Table 7.1 I think in the spoken and written components of the ICE-GB corpus.  
206

Table 7.2 I think in different text types in the ICE-GB corpus.  
206

Table 7.3 I think in different positions in the utterance: sample of 100 examples.  
208
| Table 7.4 | *I think* in different positions in the broadcast discussion: sample of 100 examples. | 211 |
| Table 8.1 | Extent to which stance is attributed to the speaker/writer. | 222 |
| Table 8.2 | Description of the *Longman Spoken and Written English* sub-corpora used in the analysis. | 227 |
| Table 8.3 | Stance noun and adjective framework. | 228 |
| Table 8A.1 | All stance adjectives and nouns used in the analysis. | 247 |
| Table 9.1 | Most frequent initial and free-standing interjections in LSWE-AC. | 258 |
| Table 9.2 | Secondary interjections in LSWE-AC. | 260 |
| Table 11.1 | Number of words in the Bolton, ICE-Ireland and BNC corpora. | 306 |
| Table 11.2 | Frequency of tails in the LSWE, London–Lund, CANCODE and York corpora. | 311 |
| Table 11.3 | Frequency of tails in the BNC, ICE-Ireland and Bolton corpora. | 315 |
| Table 11.4 | Distribution of different tail variants in the BNC, ICE-Ireland and Bolton corpora. | 316 |
| Table 11.5 | Variation with noun phrase tails in the BNC, ICE-Ireland and Bolton corpora. | 319 |
| Table 11.6 | Variation with pronoun tails in the BNC, ICE-Ireland and Bolton corpora. | 320 |
| Table 11.7 | Variation with operator tails in the BNC, ICE-Ireland and Bolton corpora. | 321 |
| Table 11.8 | Variation with inverted operator tails in the Bolton Corpus. | 322 |
| Table 11.9 | Emphatic tag clauses in ICE-Ireland. | 323 |
| Table 11.10 | Emphatic tag clauses in ICE-Ireland and inverted operator tails in the Bolton Corpus. | 323 |
| Table 12.1 | Distribution of *this* and *these* across textual components (RF, raw frequency; NF, normalized frequency, per 1,000 words). | 346 |
| Table 12.2 | Pearson residuals for the distribution of *this* across textual components. | 347 |
| Table 12.3 | Key items in narrative-initial utterances (CNI) (Freq. A) compared to narrative-medial (CNM) and narrative-final utterances (CNF) (Freq. B). | 348 |
| Table 12.4 | Semantic groups of nouns collocating with *this* in R1–4 in narrative-initial utterances (CNI). | 350 |
| Table 13.1 | Breakdown of the Hong Kong and Limerick corpus data used in the study. | 367 |
xii  List of tables

Table 13.2  HKC versus NES in HKCSE (conversation): comparison of thirteen sub-types of VL type 1. 368
Table 13.3  LCIE and HKCSE (conversation): comparison of thirteen sub-types of VL type 1. 369
Table 14.1  The placement of UHM in Kjellmer’s Cobuild corpus, based on Kjellmer (2003: 184). 386
Table 14.2  The composition of the first quarter of the Santa Barbara Corpus of Spoken American English, with number of instances of UHM per text and frequency of UHM per 1,000 words. 387
Table 14.3  The distribution of UHM in SBC1, compared with Kjellmer (2003) and Rühlemann (2007). 393
Table 14A.1  The distribution of UHM according to turn-position in SBC1. 406
Table 14A.2  Transcription conventions in the Santa Barbara Corpus, from Du Bois et al. (1993). 407
Table 15.1  Frequencies of mm and yeah in the ICE-AUS corpus, as standalones and in strings of 3 to 6 backchannels. 414
Notes on contributors


GISLE ANDERSEN is Professor of English Linguistics at NHH Norwegian School of Economics in Bergen, Norway. His research focuses mainly on various aspects of spoken interaction. His work deals especially with the use of corpora for studies in pragmatics and sociolinguistics, covering topics such as discourse markers and interjections, vague language and listenership. He has also worked on written communication, lexicography and terminology, and the influx of English language into Norwegian. He is the author of Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents (2001) and he co-authored Trends in Teenage Talk: Corpus Compilation, Analysis and Findings (2002). He edited the book Exploring Newspaper Language (2012) and co-edited Pragmatic Markers and Propositional Attitude (2000) and Pragmatics of Society (2011). Andersen has been involved in various corpus compilation projects, including COLT (The Bergen Corpus of London Teenage Language) and the Norwegian Newspaper Corpus.

DOUGLAS BIBER is Regents’ Professor of English (Applied Linguistics) at Northern Arizona University. His research efforts have focused on corpus linguistics, English grammar, and register variation (in English and cross-linguistic; synchronic and diachronic). He has written over 190 research articles, 7 edited books, and 14 authored books and monographs; these include a textbook on Register, Genre, and Style (Cambridge 2009), the

**Winnie Cheng** is Associate Dean, Faculty of Humanities, and Professor of English and Director of the Research Centre for Professional Communication in English (RCPCE), Department of English, The Hong Kong Polytechnic University (egwcheng@polyu.edu.hk). Her research interests include pragmatics, corpus linguistics, critical discourse analysis, conversation analysis, lexical semantics, discourse intonation, and intercultural communication.

**Brian Clancy** lectures in academic writing and research methods at Mary Immaculate College, University of Limerick, Ireland. His research work focuses on the blend of a corpus-linguistic methodology with the discourse-analytic approaches of pragmatics and sociolinguistics. His primary methodological interests relate to the use of corpora in the study of language varieties and the construction and analysis of small corpora. His published work addresses these areas and also explores corpus-based discourse analysis in intimate settings, such as between family and close friends, and the language variety Irish English. He is also involved in research projects on academic discourse, both spoken and written, and has published in this area. He is co-author, with Anne O’Keeffe and Svenja Adolphs, of *Introducing Pragmatics in Use* (2011).

**Giuliana Diani** is a tenured researcher in English Language and Translation at the University of Modena and Reggio Emilia. She holds an MA in Language Studies from the University of Lancaster (UK) and a PhD in English Linguistics from the University of Pisa (Italy). She has worked on various aspects of discourse analysis and EAP, with special reference to argumentation, metadiscourse and evaluative language. Her recent work centres on language variation across academic genres, disciplines and cultures through the analysis of small specialized corpora. Her recent publications include *Variation and Change in Spoken and Written Discourse: Perspectives from Corpus Linguistics* (co-edited with Julia Bamford and Silvia Cavalieri, 2013), *Reviewing Academic Research in the Disciplines: Insights into the Book Review Article in English* (2012), and *Academic Evaluation: Review Genres in University Settings* (co-edited with Ken Hyland, 2009).

**Bethany Gray** is an Assistant Professor of English at Iowa State University, where she teaches in the Applied Linguistics and Technology program. Her research relies on the use of corpus linguistics methodologies to investigate register variation in English, with an emphasis on academic writing and...
disciplinary variation, as well as on methodological considerations in corpus-based research. Her research has appeared in journals such as the International Journal of Corpus Linguistics, Corpora, the Journal of English for Academic Purposes, English for Specific Purposes, TESOL Quarterly, English Language and Linguistics, and the Journal of English Linguistics.

Gunther Kaltenböck is Senior Lecturer at the English Department of Vienna University. He holds an MA from the University of London and a PhD and habilitation (postdoctoral thesis) from the University of Vienna. He has published numerous book chapters and articles in international journals, mainly on cognitive–functional grammar and corpus linguistics, a book on *It-Extraposition and Non-Extraposition in English* (2004), and he co-edited *Tracing English Through Time: Explorations in Language Variation* (2007), *Proceedings of Anglistentag 2001* (2002) and *New Approaches to Hedging* (2011). His research interests include cognitive–functional grammar, grammatical variation, phonetics and pragmatics.

Thomas Kohnen is Professor of English Historical Linguistics at the University of Cologne. His major fields of study include historical pragmatics and historical text linguistics, corpus linguistics (both synchronic and diachronic), historical syntax, speech act theory, as well as orality and literacy and the language of religion. He is co-editor of the book series *English Corpus Linguistics*. He is also in charge of the *Corpus of English Religious Prose* (COERP), which is being compiled at the English Department of the University of Cologne. His recent publications include studies on politeness, religious discourse and English Bible translations.

Paula Garcia McAllister received a Master’s degree in Applied Linguistics from Northeastern Illinois University in 1991, and a PhD in Applied Linguistics from Northern Arizona University in 2004. Her dissertation, titled *Meaning in Academic Contexts: A Corpus-Based Study of Pragmatic Utterances*, involved line-by-line analysis of speaker meaning, applying a taxonomy of pragmatic functions based on Searle’s (1969) speech acts. She has taught undergraduate and graduate level courses on general linguistics, second language assessment, language teaching methodology, and research methods at several American universities. She has published articles in peer-reviewed journals such as the *Bilingual Research Journal* and *Language Awareness*, and has authored and co-authored several book chapters. Her current position involves facilitating research at a healthcare organization in Northern Arizona. She currently teaches courses in an online Master’s programme at Shenandoah University.
MICHAEL MCCARTHY is Emeritus Professor of Applied Linguistics, University of Nottingham, UK, Adjunct Professor of Applied Linguistics, University of Limerick, Ireland, and Visiting Professor in Applied Linguistics at Newcastle University, UK. He is author/co-author/editor of 50 books, including Touchstone, Viewpoint, Cambridge Grammar of English, English Grammar Today, The Routledge Handbook of Corpus Linguistics, Spoken Language and Applied Linguistics, Issues in Applied Linguistics, and English Vocabulary in Use, as well as more than 100 academic papers. He has lectured on language and language teaching in 44 countries and has been actively involved in the profession for 48 years.


ANNE O’KEEFFE is a senior lecturer in Applied Linguistics at Mary Immaculate College, University of Limerick, Ireland. She has published widely in the area of applied linguistics, and in the sub-field of corpus linguistics, including Investigating Pragmatics in Use (2011, with Brian Clancy and Svenja Adolphs), English Grammar Today (Cambridge 2011, with Ronald Carter, Michael McCarthy and Geraldine Mark), From Corpus to Classroom: Language Use and Language Teaching (Cambridge 2007, with Ronald Carter and Michael McCarthy), Investigating Media Discourse (2006) and The Vocabulary Matrix (2009, with Michael McCarthy and Steve Walsh). She is also co-editor of the Routledge Handbook of Corpus Linguistics (2011, with Michael McCarthy).

MATTHEW BROOK O’DONNELL is a Research Assistant Professor at the Annenberg School for Communication at the University of Pennsylvania and based in the Communication Neuroscience Lab. His research
background includes corpus linguistics, natural language processing and data mining with a focus on extracting linguistic patterns and networks from large textual databases (or corpora). His current research combines linguistic analyses of (social-)media language and persuasive discourse with behavioural and neuroscience approaches. He was previously a Research Fellow in Corpus Linguistics at the English Language Institute at the University of Michigan.

ALAN PARTINGTON is Associate Professor of English Linguistics at the University of Bologna. His research interests include corpus linguistics, corpus-assisted discourse studies (especially of political language), corpus research methodology, corpus-assisted stylistics, diachronic language studies, evaluation and evaluative prosody, irony and metaphor. He is the author of Patterns and Meanings, The Linguistics of Political Argument: The Spin-Doctor and the Wolf-Pack at the White House, Persuasion in Politics (with C. Taylor), The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-Talk, Patterns and Meanings in Discourse (with A. Duguid and C. Taylor), editor of Modern-Diachronic Corpus-Assisted Discourse Studies and co-editor of Corpora and Discourse.

PAM PETERS is a Fellow of the Australian Academy of the Humanities, and an Emeritus Professor of Macquarie University. She previously held a Personal Chair in Linguistics, and was Director of the University’s Dictionary Research Centre. She was a member of the Editorial Committee of Macquarie Dictionary from 1986 to 2006, and authored the Cambridge Guide to English Usage (Cambridge 2004). She continues to conduct research in lexicography, terminography, and Australian and international English.

CHRISTOPH RÜHLEMANN is currently a researcher at Marburg University. He is the author of Conversation in Context: A Corpus-Driven Approach (2007) and Narrative in English Conversation (Cambridge 2013). He has published on different topics relating to conversational English in edited collections as well as journals such as Applied Linguistics, the ICAME Journal, the International Journal of Corpus Linguistics, and the Journal of English Linguistics.

IVOR TIMMIS is Reader in English Language Teaching at Leeds Metropolitan University (UK), where he supervises PhD students and teaches on the MA in ELT. His research interests lie mainly in spoken corpora and the relationship between corpus linguistics and language teaching. In recent years he has been compiling an unusual corpus of spoken English – the Bolton/Worktown Corpus, 1937–1940 – which is a record of informal conversations among working-class people in the northern industrial town of Bolton (UK). He recently received a British Academy/Leverhulme grant to continue with this work with a view to making the corpus publicly available.
GUNNEL TOTTIE is Professor Emerita of English Linguistics at the University of Zurich. She is a corpus linguist with a variationist approach, whose publications include *Negation in English Speech and Writing* (1991) and *An Introduction to American English* (2002), an English–Swedish dictionary, and articles on relativization and tag questions. Her current work deals with pragmatic markers in contemporary British and American English and relativization in Early Modern English, especially the use of zero subject constructions.

MARTIN WEISSER is currently a Visiting Professor in the Center for Linguistics and Applied Linguistics at Guangdong University of Foreign Studies, China. He obtained his *venia legendi* in English Linguistics from the University of Bayreuth, Germany, in 2011, and has taught English Linguistics at different universities in the UK, Germany, and Hong Kong. He specializes in the computational analysis and identification of speech acts, but his general research interests include all aspects of corpus-linguistic analysis, related to both spoken and written language, including learner data. He is the author of *Essential Programming for Linguistics* (2009), and has written various freeware programs for linguistic analysis, including the Dialogue Annotation and Research Tool (DART).

DEANNA WONG’s doctoral dissertation focused on *Dynamic Listening: The Use of Backchannels in Telephone Conversations* (2012), including papers published in the *International Journal of Corpus Linguistics and Corpora*. Her postdoctoral research interests are in strategic use of pauses in conversation, and feedback in online communication. She is currently a Lecturer in Linguistics at Macquarie University, Sydney.