Introduction
Corpus pragmatics: laying the foundations

Christoph Rühlemann and Karin Aijmer

0.1 Introduction

Corpus pragmatics is a relative newcomer on the pragmatic and the corpus-linguistic scene. For a long time pragmatics and corpus linguistics were regarded as 'parallel but often mutually exclusive' (Romero-Trillo 2008: 2). However, in recent years corpus linguists and pragmaticists have actively begun exploring their common ground. This is attested, for example, by the 2004 special issue of the *Journal of Pragmatics* dedicated to corpus linguistics, the 2007 *IPrA* conference on 'Pragmatics, corpora and computational linguistics', the 2008 *ICAME* conference on 'Corpora: Pragmatics and Discourse', and a number of recent monographs and edited collections (e.g., Adolphs 2008, Romero-Trillo 2008, Felder et al. 2011, Jucker and Taavitsainen 2014).

In this introduction we will discuss how pragmatics and corpus linguistics can profit from each other. The focus will be on the methodologies that are key to the two fields and how they can be integrated into corpus-pragmatic research. To begin with, our use of the term pragmatics needs to be defined (Section 0.2). This will be followed in Section 0.3 by a discussion of the basic characteristics of corpus linguistics. In Section 0.4 we outline how corpus pragmatics can be seen as an intersection of corpus linguistics and pragmatics. In the last section, Section 0.5, we aim to introduce the individual contributions to this handbook in brief detail.

0.2 Pragmatics defined

The origin of modern pragmatics is often credited to the work of Morris (1938), who distinguished three 'dimensions of semiosis', viz. syntax (the relation of signs to one another), semantics (the relation of signs to the objects they denote), and pragmatics (the relation of signs to their users). While semantics asks, 'What does X mean?', targeting X (the signs under scrutiny) in abstraction from the circumstances of their use, pragmatics foregrounds these circumstances, triangulating the signs, the sign user, and the situation of use. Pragmatics is primarily concerned not with sets of rules for well-formed

sentences or with inherent meanings of signs, but with how language is used in communication. Communication invariably involves at least two parties – a speaker and a listener or a writer and a reader. As a consequence, pragmatics revolves around 'language use and language users *in interaction*' (Bublitz and Norrick 2011: 4; added emphasis). Who interacts with whom is crucial in that different people share, or do not share, different background knowledge and, depending on what knowledge is activated, the same words may be interpreted differently by different respondents. So, communication is much more than the coding (by the speaker) and decoding (by the listener) of signs: it involves complex processes of inferences and interpretation, based not only on what is said but also on what is *not* – and need not be – said because it is situationally, socially or culturally 'given'. Pragmatics, in this sense, is 'the art of the analysis of the unsaid' (Mey 1993: 245; see also Yule 1996). The foundational question in pragmatics is therefore: 'What does the speaker (or writer) mean by X and how is it understood by the listener (or reader) in the given situation?' (see Leech 1983). Pragmatics can thus be defined as 'the study of the use of context to make inferences about meaning' (Fasold 1990: 119); for an elaborate discussion of the notion of pragmatics and how it can be distinguished from semantics, see Levinson (1983: Chapter 1).

Of major importance for making inferences from what is communicated is the context in which the communication occurs. Communication unfolds differently depending on the activity in which it is used: writing a tweet on the phone, exchanging greetings in the workplace, transacting with a bank clerk, discussing quotidian life's trivia with your spouse after dinner, posting a response to a query in an online forum, and so forth. The language user chooses a linguistic form variably according to the social situation, which is broadly conceived and includes such factors as speaker identity, relations to the hearer, activity type and speaker stance (Ochs 1996: 410). How and what interactants communicate is inevitably constrained by that context: tweets are severely restricted in terms of length, at work power relations co-determine communicating styles, marital talk typically involves the spouses' children. Also, the understanding of an utterance is based on cues of different kinds. The interpretation depends on verbal features together with non-verbal modalities such as prosody, kinesics, gesture and facial expressions. Indeed, listeners make inferences from a '"bundle" of interacting behavioural events or non-events from different communicational subsystems (or "modalities") simultaneously transmitted and received as a single (usually auditory-visual) impression' (Crystal 1969: 97).

Moreover, the theory of utterance interpretation must take a dialogic approach. What is said is always in response to what has been said before and it creates conditions for what comes afterwards. What I say or write to you (in whatever form or situation) provides a context for your response,

and your response provides yet more context to how I respond to your response, and so forth.

The intricate contextual embeddedness of communication poses immense challenges for pragmatic analysis (see Cook 1990). What are the relevant contextual features, i.e., the features which are activated in the communication situation? How do the contextual parameters differ depending on the communication situation? The challenges are particularly serious for diachronic pragmatic analyses. As expressed by Kohnen (this volume), 'Can we recover enough information about the communicative practice of past ages in order to faithfully reconstruct and interpret the pragmatic meaning of the written documents that have come down to us?' (see also Jucker and Taavitsainen 2014). Because of the focus on individual texts, pragmatic research is in essence qualitative: the focus is not on the number of occurrences but on the functional behaviour observable in the texts of the phenomena under examination. Given the dependence on context, pragmatic research has methodologically relied on the analysis of small numbers of texts where careful 'horizontal' reading is manageable, that is, where large and often whole texts are received and interpreted in the same temporal order in which they were produced and received – a methodology which, as will be shown below, contrasts sharply with the 'vertical' methodology prevalent in corpus linguistics. The horizontal-reading methodology is illustrated in Figure 0.1.

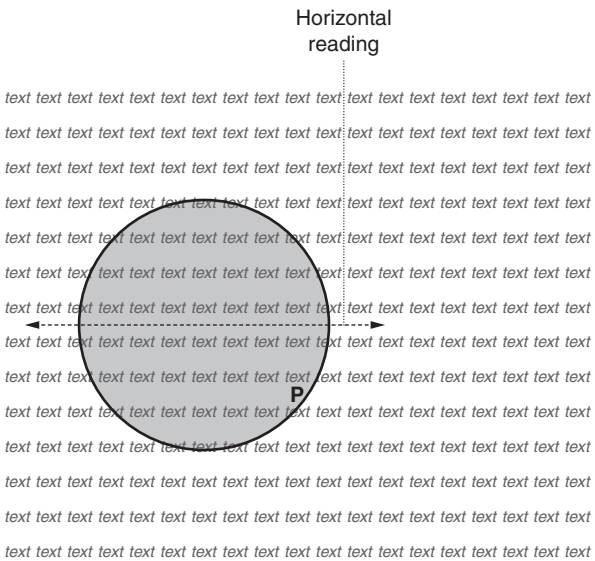

Figure 0.1 Horizontal-reading methodology in pragmatics (P).

0.3 Corpus linguistics

While pragmatics is a relatively young discipline, the history of corpus linguistics is even younger. Although the use of concordances, as 'the most basic way of processing corpus information' (Hunston 2002: 38), can be traced back as far as the thirteenth century (O'Keeffe and McCarthy 2010: 3), corpus linguistics in its modern incarnation is owed to the increasing availability of computers since the second half of the twentieth century. The first electronic corpus compiled in the 1960s was the Brown Corpus, a one-million-word corpus representing a range of written genres (Francis and Kučera 1964). Subsequently, due to the enormous advances made in computer technology which allowed ever greater storage and faster processing of ever larger quantities of data, corpora quickly made inroads into linguistic research. In recent years corpora have even 'begun to be freely available online to the casual browser, language learner and relatively novice student' (Anderson and Corbett 2009). Also, Google published its own large-scale corpus, the *Google Ngram Viewer*, an online resource containing hundreds of millions of books in a number of languages (see Michel et al. 2011). Corpora have come to be applied in a wide range of linguistic disciplines including lexicography, grammar, discourse analysis, sociolinguistics, language teaching, literary studies, translation studies, forensics and pragmatics (see O'Keeffe and McCarthy 2010).

The impact of corpora has been such that observers speak of a 'corpus revolution': see Crystal (2003: 448) and Tognini Bonelli (2012: 17). The revolutionary potential is due to the fact that now language samples can be collected and searched in such large quantities that 'patterns emerge that could not be seen before' (Tognini Bonelli 2012: 18). The impact of the revolution has been felt most dramatically in the study of what Sinclair (1991) termed the 'idiom principle', demonstrating that lexis and grammar interact in intricate ways and calling into question the long-held categorical distinction between grammar and lexis. Corpus analyses have shaken the foundations of linguistics such that 'by the late 20th century lexis came to occupy the centre of language study previously dominated by syntax and grammar' (Scott and Tribble 2006: 4).

Spoken corpora take centre-stage when it comes to studying language use. Their collection is time-consuming and the transcription of the data poses special problems. For instance, the *London–Lund Corpus of Spoken English* (LLC) from the 1960s and 1970s has been used to study discourse markers and conversational routines (Aijmer 1996, 2002). However, the corpus is small (half a million words). We now find very large collections of spoken data such as the spoken components of the *British National Corpus* (BNC) and the *Corpus of Contemporary American English* (COCA).

In particular, research based on multimodal corpora is currently growing exponentially, promising to facilitate important insights into the interplay

between linguistic and non-linguistic semiotic systems (e.g., Carter and Adolphs 2008). There are now special multilingual tools available for audio and visual data facilitating the study of feedback in the form of gesture, body posture and gaze as well as their integration with discourse.

In recent years we have seen a broadening of pragmatics to new languages and regional varieties, to new text types and spoken or written registers. This broadening is made possible by the development of new corpora which can be used to study pragmatic phenomena in different text types and situations. A number of 'sociolinguistic' corpora have emerged which provide information about the speakers (age, gender and class). They make it possible to study the sociolinguistic distribution of pragmatic markers and speech acts (Macaulay 2005). The focus is on the factors which make one variant use more acceptable than another. Timmis (this work), for example, compares utterance-final 'tails' in three different corpora, offering information about social and regional variation and changes over time. We can also use corpora to compare language use across registers. In the present work, Gray and Biber perform a comparative register analysis of implicit stance expressions in a large corpus, the *Longman Spoken and Written English* (LSWE) corpus (Biber et al. 1999: 24–35).

Corpora are further distinguished by whether they are raw (that is, text-only) or annotated, with corpus annotation defined as 'the practice of adding *interpretative, linguistic* information to an electronic corpus of spoken and/or written language data' (Garside et al. 1997: 2; emphasis in original). The most widely used corpus annotation is part-of-speech (POS) tagging, whereby every word in the corpus is automatically assigned to its grammatical class. A number of corpora, such as the British Component (ICE-GB) of the ICE family, are 'parsed', that is, the texts contained in them are automatically segmented 'into constituents, such as clauses and phrases' (Hunston 2002: 19). Further, a small number of corpora have been fitted with phonetic, semantic, discourse and pragmatic annotation (more on the latter follows below). An example of POS tagging in the BNC is given in extract (1):[1]

(1)
```
<w c5="NN1"       >Dad </w>
<w c5="DTQ"       >what</w>
<c c5="PUN"       >? </c>
<w c5="AVQ"       >How </w>
<w c5="AJ0-AV0"   >long</w>
<w c5="VBZ"       >'s </w>
```

[1] The example is presented in a simplified version where only the c5 tag is given. In the original files in the BNC, each word element receives not only one but three attributes: c5 (the full CLAWS 5 tag set), hw (headword), and pos (a reduced set of word classes). The first word in example (1), 'how', then looks like this:

```
<w c5="AVQ" hw="how" pos="ADV">How </w>
```

```
<w c5="DPS"       >our </w>
<w c5="NN1"       >Mum </w>
<w c5="VVG"       >going </w>
<w c5="TO0"       >to </w>
<w c5="VBI"       >be </w>
<w c5="CJS"       >before </w>
<w c5="PNP"       >she </w>
<w c5="VVZ"       >comes </w>
<w c5="AVP-PRP"   >in</w>
<c c5="PUN"       >?</c>
```

Each word is identified in terms of its grammatical word class: 'Dad' as a singular noun (NN1), 'what' as a question determiner (DTQ), ''s' as the third-person present tense form of the verb BE (VBZ), 'our' as a possessive determiner and so on. For some words the automatic assignment was inconclusive, such as for 'long', where an ambiguous tag was used (AJ0-AV0). The most obvious advantage of POS annotation is the enhanced precision with which words can be retrieved. The form ''s', for example, can be the short form of 'is', 'has', and even 'does' or the genitive –'s. Further advantages of POS annotation, as noted by Leech (1997: 5), are its re-usability, which saves other researchers precious time and effort, and its multifunctionality: POS annotation can be 'a kind of "base camp" annotation towards more difficult levels of annotation' (Leech 1997: 5) such as those of syntax, semantics or, as we will see below, pragmatics. Such 'more difficult levels of annotation' may be necessary if the analysis is intended to capture not only what is manifest in the surface structure but also what is going on beneath this surface in terms of discourse and pragmatic structure. For example, in (1), which is part of an extended utterance by a single speaker, it will be hard for the reader to make sense of how the first few words cohere. Their coherence can only be appraised by inspection of more context. Below, in Section 0.4, we will view the excerpt in its larger context and see how pragmatic annotation can help enlighten discourse and pragmatic structure.

As regards size, corpora range from small specialized corpora containing fewer than a million words to mega-corpora of more than a billion words (for example, the *Cambridge International Corpus*) to the web-as-corpus, which counts trillions of words (e.g., Hundt et al. 2007). In the present volume both large and small corpora are used. Besides mega-corpora such as the BNC we find small specialized corpora such as Giuliana Diani's corpus of academic book review articles (this work).

Not only large corpora but even small specialized corpora contain far more words than could possibly be read and analysed by any one researcher in the same way as the select texts which pragmaticists are used to working with. Corpus-linguistic methodology is adapted to this size of corpora: the favoured methodology is not so much horizontal as vertical reading. The

vertical-reading methodology can best be illustrated using the KWIC (key word in context) format, also referred to as concordance line display, where the word under scrutiny (the node word) is located in, and retrieved from, all the texts in the corpus in which it occurs and aligned in the centre of the concordance lines. (Only when the co-text as provided in the concordance lines is insufficient will the researcher inspect larger contexts.) Consider the first ten hits of a KWIC search in the BNC for the word 'corpus' (the second column on the left indicates the texts in which the node word was found):

1	J0V 2050	mation wherever a new item in the	**corpus**	began. The package would also recogni
2	HGR 1504	uggests that perhaps the size of a	**corpus**	is more significant than its composition
3	A03 129	bsequently annulled the habeas	**corpus**	on grounds of procedural irregularites
4	HU9 1148	the city where the Feast of	**Corpus**	Christi originated. However, because of
5	A68 1367	In those days the Fellows of	**Corpus**	were rather proud of the briskness of the
6	B77 2169	alls the manuscripts 'the largest	**corpus**	of texts' of them and 'a remarkable reso
7	CMH 935	work on the effects of cutting the	**corpus**	callosum in humans (Gazzaniga 1985)
8	CFF 333	at the latter's old college of	**Corpus**	Christi at Oxford. Here his most influen
9	CG6 151	uage, children have access to a	**corpus**	or sample of language in the utterances
10	EES 1839	dictionary derived from the LOB	**corpus**	can make a significant contribution to the

The concordances can be scanned by the researcher 'for the repeated patterns present in the context of the node' (Tognini Bonelli 2012: 19). In the case of 'corpus', it leaps out that 'corpus' co-occurs with 'christi', together forming the compound 'Corpus Christi', which, in the two instances in the KWIC display given above, refers to the religious holiday and, respectively, an Oxford college (upon closer examination it turns out that, indeed, 'christi' is *the* top most frequent collocate of 'corpus' in the BNC!). Perhaps less surprisingly, in five instances 'corpus' refers to a collection of texts in the corpus-linguistic sense. To establish these patterns researchers go through the texts focusing on the node word and the minimal co-texts surrounding the node word. That is, the analysis essentially cuts across the texts following occurrences of the node word in a vertical direction, as shown in Figure 0.2.

The outcome of a corpus-linguistic vertical analysis is typically a frequency list of some sort, ordering the items searched for in terms of the number of instances found in the corpus. For example, as noted, the word 'corpus' most frequently collocates with 'christi', the second most common collocate is 'habeas' (forming the law term 'habeas corpus'), the third most frequent is 'lob', a reference to the *Lancaster–Oslo–Bergen* (LOB) *Corpus*. The first top five collocates of 'corpus' in the BNC ordered by their log-likelihood value are given in Table 0.1.

Indeed, Gries argues that 'strictly speaking at least, the only thing corpora can provide is information on frequencies' (Gries 2009a: 11). On this view corpus linguistics is essentially a quantitative discipline (see also Gries 2010). To compare frequencies derived from one and the same corpus or different corpora

Table 0.1 *Top five collocates of 'corpus' in the BNC.*

Number	Word	Total number in whole BNC	Expected collocate frequency	Observed collocate frequency	Log-likelihood value
1	christi	82	0.003	60	1126.032
2	habeas	49	0.002	49	997.3541
3	lob	190	0.007	57	928.1762
4	british	35,431	1.355	50	264.1342
5	callosum	17	0.001	13	245.9446

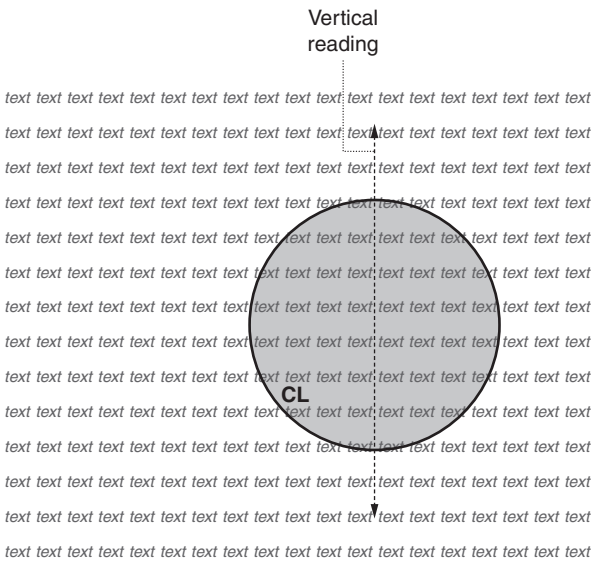

Figure 0.2 Vertical-reading methodology in corpus linguistics (CL).

and to establish whether the frequencies are due to chance or a reflection not only of the distribution in the corpus (which is, whatever its size, just a minute sample) but in the language or language variety as a whole (what statisticians call the 'population'), the use of statistical operations is necessary. For example, to compare frequencies between (sub-)corpora of unequal sizes, corpus linguists calculate normalized frequencies (e.g., frequencies per 100 utterances, per 1,000 words, and so on; see, for example, Biber et al. 1998: 33–34). Or, to gauge whether a word co-occurs with a node word just because the word itself is very frequent and the odds are greater that it will appear next to the node word, or whether it occurs more often in the company of the node than would be expected on the basis of the word's overall occurrence in the corpus, a number of measures

can be used (for an accessible discussion of association measures see Hoffmann et al. 2008: Chapter 8). One such measure is log-likelihood, the measure given in Table 0.1. To illustrate, it is no surprise that in the BNC the word 'british' has a much higher overall occurrence (35,431 occurrences) than 'christi' (82 occurrences). However, 'christi' co-occurs with 'corpus' 60 times whereas 'british' co-occurs with 'corpus' 50 times. Hence, the strength expressed in the log-likelihood value that binds 'christi' to 'corpus' is much greater than the bond between 'british' and 'corpus'. Other techniques involve even more sophisticated statistical analysis. For example, Gray and Biber (this work) used a statistical program creating KWIC lines of instances of the stance adjectives and nouns they were interested in; Rühlemann and O'Donnell (this volume) test distributions of *this* and *these* across different positions in narratives for sameness using Kolmogorov–Smirnov tests. (For worthwhile introductions to statistics for (corpus) linguists, see Gries 2009a, 2009b.)

0.4 Corpus pragmatics

Corpus pragmatics, as a combination of pragmatics and corpus linguistics, combines the key methodologies of both fields. Given the context-dependence of pragmatic phenomena, merely vertical analyses of corpus data are rare in corpus pragmatics. Similarly, analyses in which corpus data merely serve to illustrate a pre-existing theory are far from prototypical too (although they are possible and maybe a step ahead compared to the often completely invented sentences relied upon in earlier pragmatic work). Most typically, corpus-pragmatic research integrates vertical and horizontal analysis in some way.

To begin with, corpus-pragmatic analyses can take lexical words or constructions which previous pragmatic analyses have shown to have recurring pragmatic functions as their starting points; examples would be pragmatic markers such as *well* and *you know*. Using the KWIC function, occurrences of the forms can easily be captured and displayed in concordances both in raw-text and POS-tagged corpora (vertical reading). In a second step, the researcher can examine the use of the forms in context, weed out unwanted uses (such as *well* used as an adverbial form of *good*) and examine the functions the target items fulfil in the concordance lines (horizontal reading). This type of analysis proceeds from predefined forms to the range of functions performed by the forms (form-to-function). A closely related approach takes the inverse direction, starting from a function and investigating the forms used to perform it (function-to-form). However, the function cannot be retrieved itself, only surface forms 'orbiting' it can. For example, speakers may not only perform speech acts but also talk about them, using so-called meta-communicative expressions such as *threaten*, *request*, which 'name a particular speech act, for instance, or they may flag specific ways of speaking or communicating' (Jucker and Taavitsainen 2014: 12). These expressions can be searched for and the range of forms used to

talk about threats or requests can be examined in the specific contexts. So, again we have vertical reading preceding horizontal reading.

For most pragmatic phenomena there is no one-to-one relationship between form and function. Corpus-based studies of speech acts have therefore usually focused on fixed or conventionalized speech acts (Aijmer 1996, Deutschmann 2003, Adolphs 2008). One can for example use the corpus to search for information about 'speech act words' such as *sorry* or *thanks* (their frequency, distribution and collocations). However, speech act words by no means always accompany the relevant speech acts. Thus, while searches for occurrences of *sorry* in a corpus may achieve very high 'precision' (meaning they effectively retrieve all instances of apologies co-occurring with the word *sorry*), they may perform badly in terms of 'recall' (meaning all the apologies in which no *sorry* was used are overlooked); for a discussion of precision and recall, see Hoffmann et al. (2008: 77–79). For diachronic speech act analysis the problems of identifying speech acts are even larger, since such studies are based on written material and speech acts may change over time (see Kohnen, this work).

One way to achieve both high precision and optimal recall when analysing the many pragmatic phenomena characterized by form–function mismatch is by adding annotation targeted at the phenomena one wishes to study. The work with added pragmatic annotation is illustrated in (2), an excerpt from the *Narrative Corpus* (NC), a corpus of conversational narratives extracted from the conversational subcorpus of the BNC (see Rühlemann and O'Donnell 2012). The extract contains the same words as example (1) above.

(2)
```
<seg Reporting_modes="MDD">
<w c5="NN1"          >Dad </w>
</seg>
<seg Reporting_modes="MDF">
<w c5="DTQ"          >what</w>
<c c5="PUN"          >?</c>
</seg>
<seg Reporting_modes="MDF">
<w c5="AVQ"          >How </w>
<w c5="AJ0-AV0"      >long</w>
<w c5="VBZ"          >'s </w>
<w c5="DPS"          >our </w>
<w c5="NN1"          >Mum </w>
<w c5="VVG"          >going </w>
<w c5="TO0"          >to </w>
<w c5="VBI"          >be </w>
<w c5="CJS"          >before </w>
<w c5="PNP"          >she </w>
<w c5="VVZ"          >comes </w>
<w c5="AVP-PRP"      >in</w>
<c c5="PUN"          >?</c>
</seg>
```