Corpus pragmatics: Laying the foundations

Christoph Rühlemann and Karin Aijmer

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, Taavitsainen & Jucker forthcoming).

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 in corpus pragmatic research. To begin with, our use of the term pragmatics needs to be defined (Section 2). This will be followed in Section 3 by a discussion of the basic characteristics of corpus linguistics. In Section 4 we outline how corpus pragmatics can be seen as an intersection of corpus linguistics and pragmatics. In the last section, Section 5, we aim to introduce the individual contributions to this handbook in brief detail.

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 signs user, and the situation of use. Pragmatics is concerned primarily, 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 & 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 on, not only what is said but also what is, 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 1991: 245; cf. 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?' (cf. 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 at the work place, 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 behavioral 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 Taavitsainen & Jucker forthcoming). Because of the focus on individual texts, pragmatic research is in essence qualitative: the focus is not on number of occurrence but the functional behavior 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 1:

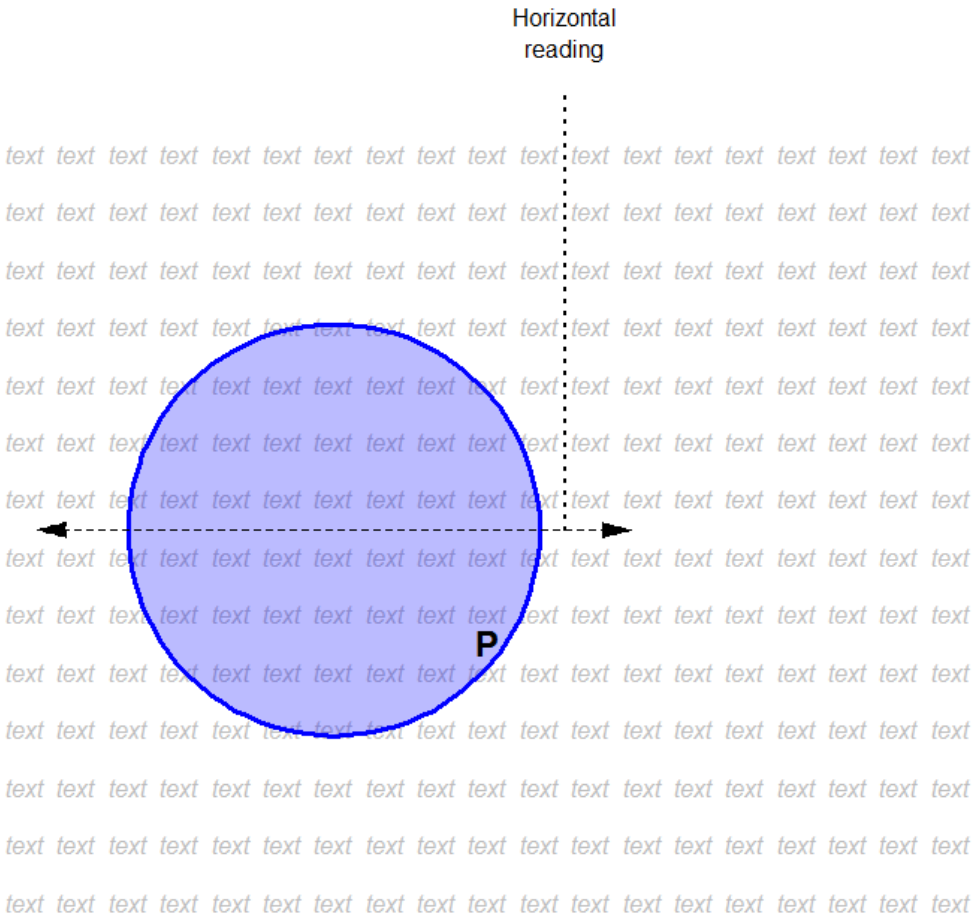

Figure 1 Horizontal-reading methodology in pragmatics (P)

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 to the thirteenth century (McCarthy & O'Keeffe 2010: 3) corpus linguistics in its modern incarnation is owed to the increasing availability of computers since the second half of the 20th century. The first electronic corpus compiled in the 1960s was the Brown Corpus, a 1-million-word corpus representing a range of written genres (Francis & 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 & Corbett 2009). Also, Google published its own large-scale corpus, the *Google N-gram Viewer,* an online resource containing hundreds of millions of books in a number of languages (cf. Michel et al. 2010). 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 (cf. McCarthy & O'Keeffe 2011).

The impact of corpora has been such that observers speak of a 'corpus revolution' (Crystal 2003: 448; Tognini Bonelli 2010: 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 2010: 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 & 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, 2001). 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).

Particularly 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., Knight & Carter 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 makes 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 & Biber perform a comparative register analysis to investigate implicit stance

expressions in a large corpus, the *Longman Spoken and Written Corpus* (LSWC) (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 corpora from 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-C is given in example (1):[i]

(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>
<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 multi-functionality: 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 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 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 Guiliana 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 analyzed 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 favored methodology is not so much horizontal as a 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	information wherever a new item in the	**corpus**	began. The package would also recognize
2	HGR 1504	suggests that perhaps the size of a	**corpus**	is more significant than its composition al
3	A03 129	subsequently 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 their
6	B77 2169	calls the manuscripts ‘the largest	**corpus**	of texts’ of them and ‘a remarkable resource’.
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 influential
9	CG6 151	language, children have access to a	**corpus**	or sample of language in the utterances they
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 2.

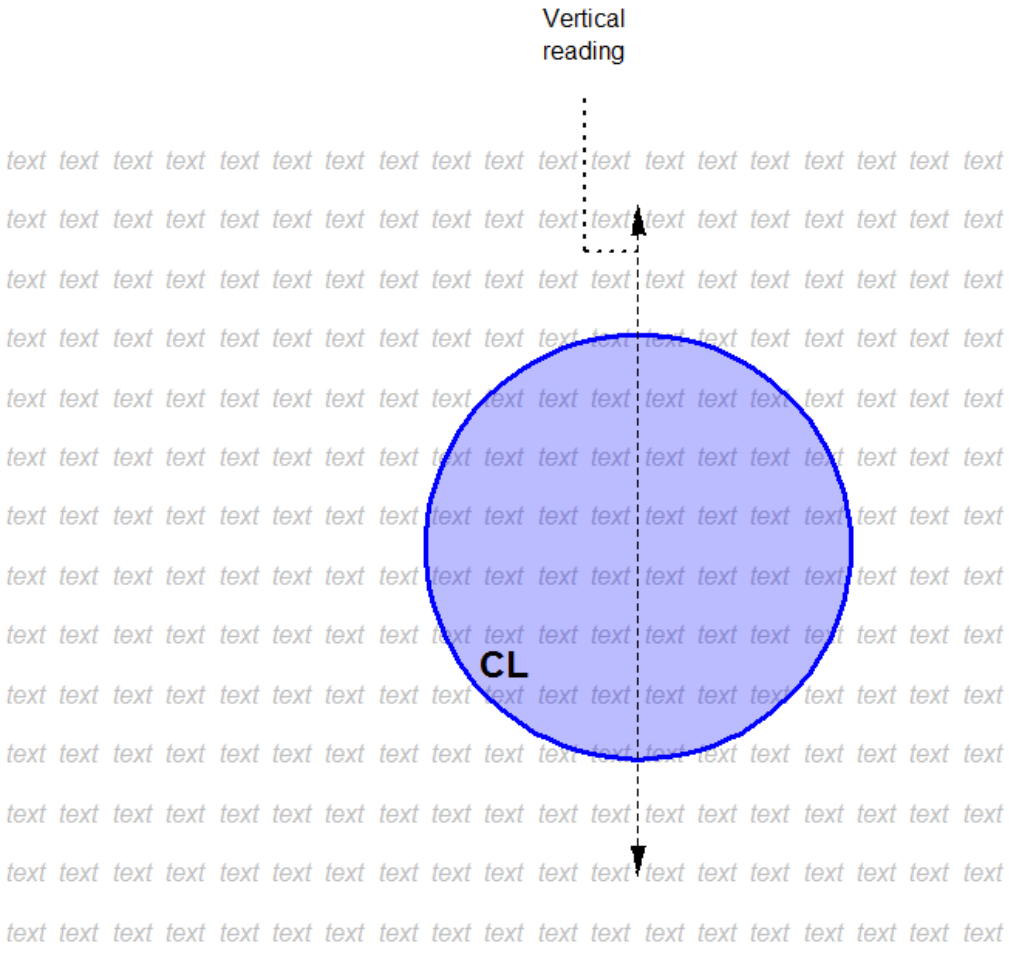

Figure 2. Vertical-reading methodology in corpus linguistics (CL)

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 *London-Oslo-Bergen* (LOB) *Corpus*. The first top five collocates of ‘corpus’ in the BNC ordered by their log-likelihood value are given in Table 1:

Table 1. Top five collocates of ‘corpus’ in the BNC

No.	Word	Total No. 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

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 (cf. also Gries 2010). To compare frequencies derived from one and the same corpus or different corpora 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 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 & Biber (this work) used a statistical program creating KWIC (Key Word In Context) lines of instances of the stance adjectives and nouns they were interested in;

Rühlemann & 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 and b.)

3. 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 earlier pragmatic work often relied on). 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 'well' used as an adverbial form of 'good') and examine the functions the target items fulfill in the concordance lines (horizontal reading). This type of analysis proceeds from pre-defined 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 & Taavitsainen). 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 do 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 analyzing 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 BNC-C (cf. Rühlemann & 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>
```

Example (2) differs from example (1) by altogether six lines containing the starting and closing tags for so-called seg (segment) tags whose attribute-values capture reporting mode types. The seg-elements mark three decisive events in the speaker's discourse, viz. changes in 'voice'. As can be seen from inspection of the larger context (see the discussion in Rühlemann 2013: 120-124), the speaker is reconstructing a conversation between a girl and her father; in so doing the speaker animates the two characters' voices using (free) direct speech (indicated by the annotation as 'Reporting_modes="MDF"' (free direct) and 'Reporting_modes="MDD"' (direct)). In the short excerpt in (2), the girl is reported as addressing her father ('Dad'), her father as replying 'what?', and the girl as inquiring about the time her mother comes back. If the text is presented as in (1) above, with all POS tags in

place but without the reporting mode tags, the switches in voice go unnoticed and, most crucially, they cannot be investigated automatically by corpus software. If the text is annotated as in (2), the switches in voice can be examined corpus-linguistically. This is no small advantage, for the switches represent discourse events which are crucial both for the speaker and the listener: in each report unit the narrator casts herself in a different role taking a different footing vis-à-vis the text (cf. Goffman 1981) while the listener needs to re-contextualize each new unit in accordance with the role shifts ('Dad' as belonging to the girl's discourse, 'what?' as attributable to the father, and so on).

Added annotation facilitates the exhaustive study of otherwise inaccessible form-function mappings. As shown, for example, in Garcia's contribution in the present volume non-conventionalized speech acts can be studied due to added annotation. The methodology Garcia used was careful line-by-line reading, assigning to "each utterance that was identified as a speech act (…) a code as to what type of speech act it represented". These annotated subsets were then processed further by use of corpus tools assigning further linguistic and contextual information to each utterance. That is, the analysis started off by means of horizontal reading (identifying speech acts) while the provision of further corpus tools to the subsets added a vertical dimension to the analysis: speech acts could be categorized not only in terms of their type but also in terms of their associated contextual factors.

Pragmatic annotation has been made available in a small yet growing number of corpora. The annotations target, for example, speech acts (e.g., Stiles 1992, Garcia 2007, Kallen & Kirk 2012), discourse markers (Kallen & Kirk 2012), quotation (Kallen & Kirk 2012, Rühlemann & O'Donnell 2012), participation role (Rühlemann & O'Donnel 2012), and politeness (Danescu-Niculescu-Mizil 2013). The reason why pragmatic annotation is not yet widely used is simple: the form-function mismatch of most pragmatic phenomena means that automatic assignment of tags will often lack precision and manual implementation of tags (which is time and resource-intensive) is unavoidable. However, as indicated in Martin

Weisser's contribution to this volume, work attempting to capture speech acts at least *semi-automatically* has made good progress: Weisser's Dialogue Annotation and Research Tool (DART) identifies speech acts such as conventionalized, dialogue-managing, and information-seeking.

So corpus-pragmatic research is more than just pragmatic research and it is more than just corpus-linguistic analysis in that it *integrates* the horizontal (qualitative) methodology typical of pragmatics with the vertical (quantitative) methodology predominant in corpus linguistics. The integrated-reading methodology underlying corpus-pragmatic research is diagrammatically shown in Figure 3.

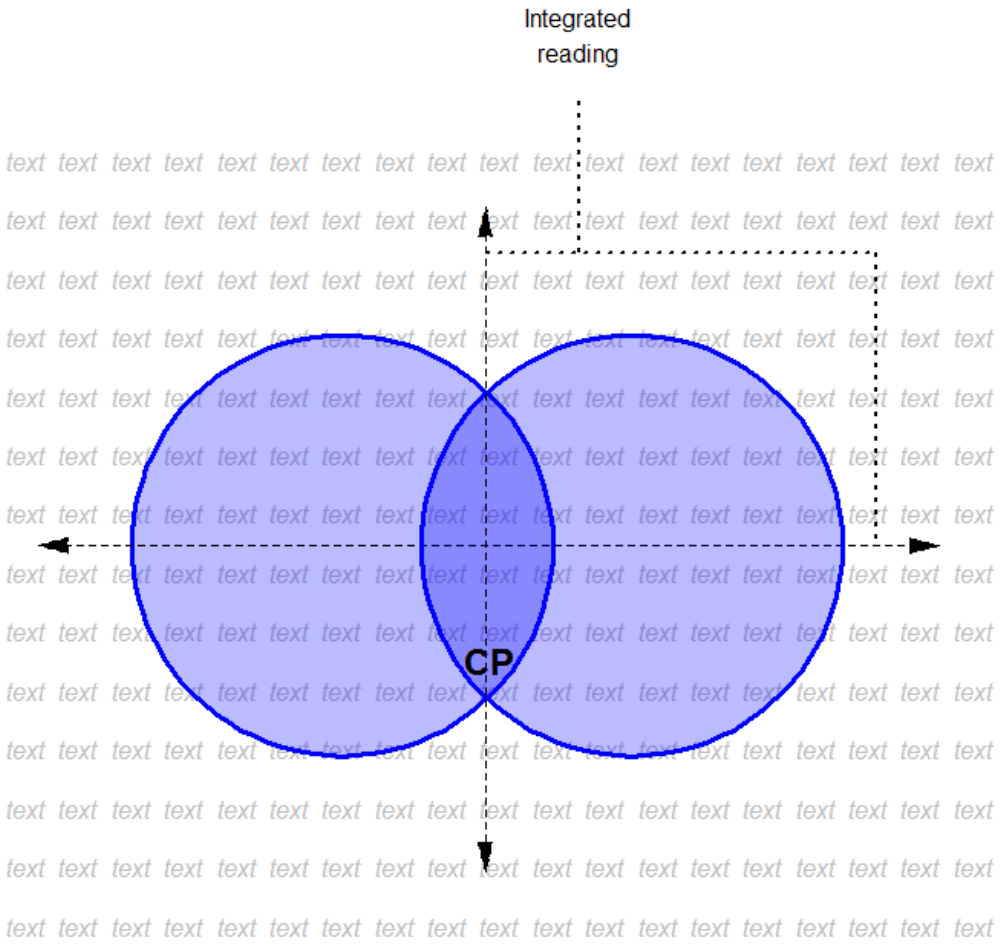

Figure 3. Integrated-reading methodology in corpus pragmatics (CP)

Given that corpus pragmatics integrates the foundational methodologies typifying the fields of corpus linguistics and pragmatics, it can be defined as the intersection between the two fields. This intersection is smaller than the two fields of which it is a composite. This figure illustrates the fact that corpus pragmatic research is neither interested in nor particularly able at many concerns that corpus linguistics and pragmatics, respectively, may be concerned with. For example, research into collocation is and will probably remain the realm of corpus linguistics, while many pragmatic functions may defy exhaustive corpus pragmatic research. However, there is considerable room for expansion of the intersection particularly if –and when– more, and more elaborate, pragmatically-annotated corpora will become available. Whether pragmatic annotation will spread across the corpus-pragmatic community will to a large extent depend on practical concerns: whether the laborious manual annotation processes can be replaced by semi-automatic and, finally, fully-automatic annotation processes that are not only more resource-economic but also more efficient allowing the precise and exhaustive targeting of hitherto intractable pragmatic phenomena in ever larger pragmatic corpora.

5. The present volume

The present volume is intended both to overview and expand this field. As shown by the contributions the range of phenomena which are regarded as pragmatic is very wide. While, then, given the overview sections, the chapters qualify as reference works, they also serve as research articles, enlarging the body of corpus pragmatic research. In the following, we briefly characterize the contributions individually.

We pursue this two-fold aim by focusing on core areas of pragmatic research and by reporting original case studies carried out in these areas. The areas covered include the following: speech acts (both in a synchronic and a diachronic perspective) (Section I),

pragmatic principles (politeness, processibility, relevance) (Section II), pragmatic markers (discourse markers, stance markers, and interjections) (Section III), evaluation (semantic prosody and use of tails) (Section IV), reference (deixis and vagueness) (Section V), and turntaking (Section VI).

5.1 Corpora and speech acts

Analysing speech acts using corpus-linguistic methods raises a number of problems both for the synchronic and diachronic analysis of speech acts. With the rise of corpus-based methods we get problems with the identification and analysis of speech acts. . A researcher might identify the speech act of apology by searching for the word *sorry.* This methodology has been used to identify speech acts of a conventional nature. However many indirect speech acts are difficult to analyse in this way because there are no pre-identified words carrying speech act meaning. Earlier corpus-based investigations have mainly used a lexical approach based on certain pre-determined speech act words. The research question addressed in the three contributions all address the following research questions: How can we use corpus tools and techniques to study speech acts if they cannot be identified by lexical means. To what extent can we use computational tools to identify speech acts automatically?

The difficulties in identifying indirect speech acts in naturally occurring discourse have also led researchers to use a combination of computerized searches and manual line-by-line analysis. In **Paula Garcia McAllister**'s case study, 'the identification-in-context' methodology is used to explore Searle's category of directives. The aim of her study is to investigate if there is a relationship between situation types and speech acts used in academic conversations. The methodology used implies a 'bottom-up' identification of the speech act by the analyst who both read through the transcript and listened to the recordings simultaneously

The corpus used was the spoken language component of the TOEFL Spoken and Written Academic Language Corpus. The corpus samples selected represent service encounters, office hours and study groups. Speaker roles were students, professors and service providers. Corpus tools were used to extract linguistic data on each utterance for example the use of modals or amplifier and the use of the utterance in a statement or a question. The results showed that there was considerable variation in the types of speech act speakers employed in different situations. The analysis also uncovered some speech acts which had been neglected in earlier speech act research. Very little has for example been written about warnings or giving instructions or directions.

As is obvious from **Thomas Kohnen's** contribution diachronic speech act analysis has now become a major field of research within the new discipline of historical pragmatics. However the historical study of speech acts encounters many problems. With the rise of corpus-linguistic methods we get problems having to do with the identification of speech acts. Another problem is specific to the historical study of speech acts: Can we for example be certain that the same pragmatic principles constrain meaning and interpretation over time and across different societies? Because of these problems diachronic corpus-based studies of speech acts tend to be eclectic. Researchers may just collect illustrations of a single speech act at a special period of the English language (illustrative eclecticism). Investigations can also start with a few selected manifestations which can be identified for example performatives or imperatives and search for them at different periods of the language.

An interesting question for the historical study of speech acts is whether indirect manifestation of speech acts such as requests were already available at earlier periods of the language. The general picture emerging from previous research is that indirect constructions only developed during the late Middle English and Early Modern English periods. On the other hand it has been shown in Kohnen's research the number of directive was almost seven

times higher in the Old English part of the *Helsinki Corpus* compared with the present-day *LOB Corpus*.

The overview section shows that most corpus-based diachronic research on speech acts has focused on individual speech acts or classes of speech acts but neglected a more comprehensive inventory of speech acts or their development through time. On this note Kohnen turns to a description of historical speech acts in the context of their 'fellow speech acts' and a systematic inventory of frequency and distribution of speech acts in corpora from the 15th, the late 17th and the late 19th century. The focus is on performatives with directive function in three historical sub-corpora. It is shown that the frequency of the performatives declines over time and that this shift takes place across genres. On the other hand this general movement of decrease is not shared by all the classes of performatives. The decline of performatives seems to be linked to socio-cultural factors such as face and literacy. The decline of directive performatives can for instance be due to the realisation that directives constitute face-threat. It has also been suggested that the spread of literacy can contribute to the decline of performatives.

Martin Weisser's overview shows that pragmatic annotation is more complex than other types of annotation due to the fact that it needs to take into account levels above the word and may have to refer to contextual information. The aim of his contribution is therefore to try and determine whether it is possible to improve on the identification and subsequent pragmatic annotation using data from a number of annotated corpora including both corpora of task-driven dialogue and the *Switchboard Corpus* (American telephone conversations).

A comparison is made between DAMSL-based schemes and Weisser's computational tool DART. DAMSL (Dialogue Act Markup in Several Layers) was arrived at through the participation of experts from various fields of linguistics and computer science within the Discourse Resource Initiative. It is shown that the DAMSL-based annotation schemes provide a useful starting-point but are unnecessarily complex and neglect syntactic aspects.

The annotations do not for example treat discourse markers as units in their own right but as part of longer units, which means that their roles as contextual clues are neglected. Moreover, the labelling of communicative functions shows that DAMSL was primarily designed to label discourse units according to their purpose on a manual basis. The DART model incorporates information from the other models and makes it more accessible from a linguistic point of view. The scheme also makes it possible to annotate large sets of data semi-automatically.

Weisser's case study demonstrates how the pragmatic annotation in DART can be used to investigate interactive strategies. The aim is investigate the strategies used by an 'agent' who has to deal with callers enquiring about train timetable information. A high number of fragments suggest that the agent avoids providing fully formulated statements where shorter, less redundant forms of expression suffice. The efficiency of the agent is also indicated by the relatively high number of yes-no questions and discourse marker initiating new topics or discourse sequences.

5.2 Corpora and pragmatic principles

The pragmatic interpretation of an utterance is distinct from it semantic interpretation. In particular it involves principles. Pragmatic principles are less binding than grammatical rules but explain the conditions under which a certain interpretation is preferred. We need a large number of principles to account for pragmatic interpretations. The processibiltiy principle accounts both for stylistic preferences and for the syntactic ordering of elements in the discourse. Much has been written about Grice's Cooperation Principle and its sub-maxims. Grice's maxim of relation (be relevant) has received a specific interpretation in relevance theory. Relevance theory is a cognitive theory of utterance interpretation. However corpus data can be analysed as manifestations of speakers' inferential processes. In this section we also consider how to deal with politeness. Politeness principles can for instance account the

interpretation of direct and indirect speech acts. They also play an important role as linguistic resource for elements which are interpreted in terms of politeness principles.

Gunther Kaltenböck examines whether the pragmatic principle of processibility as it manifests itself in the conventions of information packaging can have an effect on the development of syntactic structure. It is suggested that the emergence of a pleonastic matrix clause in examples such as *I think that a student failed the exam* is in line with the principle of processibility since it facilitates the processing of the *that*-clause. For an investigation of such structures ('presentational matrix clauses') Kaltenböck uses the *Diachronic Corpus of Present Day Spoken English* (DCPSE) and the *Corpus of Historical American English* (COHA). It is argued that such matrix clause structures offer a choice of putting the main point of the message either in the first or the second clause. Given the processibility constraint the first clause is particularly prone to convey secondary information which in turn makes it a possible candidate for grammaticalization.

Presentational matrix clauses also resemble pragmatic markers which typically occur in initial position. Potential sources for this development are matrix clauses followed by object-clauses (*I think*), matrix clauses of extraposed subjects (*it seems*), matrix clauses of copular constructions (*the thing is*), and inferential matrix clauses (*it's just that*). The concomitant effects of the development are bleaching of their original semantic content, formal invariability, loss of the *that* –complementizer, reduction and fusion to a unitary element and in some cases positional flexibility.

Gisle Andersen explores the application of a relevance-theoretic perspective on the analysis of authentic corpus data. The purpose is to show that observations from corpora can shed light on how constraints on relevance are practiced by real people in authentic discourse contexts. Like Grice's theory of conversation Relevance Theory focuses on the role of inferencing for pragmatic interpretation. However, Relevance Theory departs from Gricean pragmatics in postulating a single principle spelled out as the cognitive principle of relevance.

The principle states that human cognition is geared towards the maximization of relevance in the sense of achieving as many cognitive effects as possible for as little cognitive effort as possible. Such effects can for example be to strengthen existing assumptions or to abandon or renegotiate existing assumptions.

Central to relevance theoretical approach is the distinction between concepts and procedures. The notion of procedural encoding has been explored notably with regard to discourse markers. Such markers can be shown to have the task to guide the hearer towards the intended utterance interpretation by their capacity for imposing constraints on the interpretation. It is not necessarily the case that hearers are in need of these cues but such processing items reduce processing costs and thus contribute to relevance. Relevance theory has mostly focused on the 'classical' discourse markers such as *well*, *you know* and *I mean.* However discourse markers are not a static category but new items or 'pragmatic neologisms' readily emerge. Specifically there is a need for cross-linguistic studies. In his case study Andersen studies a number of interjections which have been borrowed from English into Norwegian. Such neologisms are for example *as if* (a marker of emphatic rejection which is found both in Norwegian and in English) and the interjection *duh.* The English examples of *as if* and *duh* were primarily represented in COCA (*Corpus of Contemporary American English*).

Politeness has been a hotly debated issue in recent years as shown in **Guiliana Diani's** contribution. An early conceptualization of politeness is represented by Leech's Politeness Principle. Leech's notion of politeness as conflict avoidance has been further expanded in Brown & Levinson 's (1987) formulation of politeness as avoiding or reducing face-threat. The framework of politeness theory has also been used together with corpus linguistics. Corpora have for example been linguistic resources for exploring politeness features such as address terms in older English.

The aim of Diani's case study is to compare the mitigation devices used to soften criticism in Italian and English book reviews. Where are mitigation devices more frequent? What kinds of mitigation devices are used? Criticisms were identified on the basis of their lexico-grammatical features and further categorized into 'direct' and 'mitigated'. The mitigation strategies identified in the corpora mainly involved the use of sequences of speech acts like praise-criticism, criticism-suggestion, praise-suggestion and hedging. In the English book direct criticism was more frequent than mitigated criticism. In the mitigated category criticism accompanied by praise was most frequent followed by hedging. The most frequent hedges were modal verbs, the epistemic verb *seem* and downtoning adverbs. A comparison with the Italian data suggested that Italian reviewers tended to avoid criticism and that they preferred mitigated criticism to direct criticism. As in the English data praise-criticism was most frequent. A difference is that the use of suggestions together with criticism was more frequent in the Italian review articles. Moreover the Italian articles contained about half the number of hedges found in English.

5.3 Corpora and pragmatic markers

We are now beginning to see many more studies of pragmatic markers which share few characteristics with the prototypical or 'classical' markers such as *you know, well, now*. Less prototypical pragmatic markers can for instance include adjectives and nouns or verbs expressing stance and have the pragmatic function to express epistemic commitment or affect in discourse. Corpus studies of pragmatic markers and stance reveal that they vary across speech and writing and across spoken and written text types.

In **Karin Aijmer's** contribution it is argued that pragmatic markers are motivated by pragmatic principles in particular by cognitive principles such as the processibility principle suggested by Leech (2003) and by social principles associated with politeness. Pragmatic

markers are difficult to define. They are characterized by reflexivity and indexicality but this is true about many other elements in language. It appears that pragmatic markers as a category are best characterized by their prototypical members. However the more one studies pragmatic markers the more exceptions one finds to the characterizations proposed. A special problem has to do with the multifunctionality of pragmatic markers and how this should be reconciled with describing pragmatic markers as combinations of form and function in context. The importance of speech context is illustrated by a corpus study of *I think* in different text types in the British component of the International Corpus of English. In conversation *I think* was used without further planning to express a spontaneous opinion or reaction. It could also be used with a polite hedging function. On the other hand, *I think* in broadcast discussion is mainly a boundary signal marking the starting-point for a new stage in the discussion.

The case study of *I think* was also used as a testing ground for the hypothesis that pragmatic markers have meaning potentials rather than fixed meanings which are realised in the same way in all situations. *I think* has core meanings (subjective opinion and epistemic meaning) from which new meanings can be inferred in interaction with contextual features.

Bethany Gray and **Douglas Biber** explore the linguistic means by which speakers and writers express stance. Analyses of stance taking a comparative register approach have shown that stance is less frequent in academic writing than in other varieties. However these studies only take into account overt stance markers. The purpose of Gray's and Biber's case study is therefore to explore the possibility that stance is expressed by implicit means in academic writing. Three sub-corpora from the *Longman Spoken and Written English* (LSWE) *Corpus* were used for the analysis: Academic prose, Newspapers and Conversation. The goal was to investigate stance adjective and nouns since they represent the stance evices that are most characteristic of academic writing. Extra-posed clauses with adjectives were shown to be especially frequent in academic writing in comparison with news and conversation. Moreover

of-phrases after stance nouns were markedly frequent in academic writing. The authors further explore the possibility that stance adjectives and nouns can be used as stance markers in contexts in which they are not contained in a complement clause (or an *of*-phrase). A number of 'new' lexico-grammatical patterns with stance adjectives and stance nouns are identified which can be regarded as the more-or-less overt expressions of a speaker/writer's attitudes, evaluations or levels of commitment to a proposition.

The section on 'Corpora and pragmatic markers' is concluded by **Neal Norrick**'s chapter on interjections. Norrick surveys the current state of research on interjections and, drawing on a number of corpora such as the *Longman Spoken and Written English Corpus* (LSWEC), the *Santa Barbara Corpus of Spoken American English* (SBCSAE), the *Saarbrücken Corpus of Spoken English* (SCoSE), presents a set of specific areas to illustrate progress and problems in the corpus investigation of interjections; these areas include exclamatory constructions, phrasal interjections, and combinations of interjections. He also presents evidence to support observations suggesting that interjections preferably occur turn-initially and, in storytelling, are intimately associated with constructed dialogue both by storytellers and listeners.

5. 4 Corpora and evaluation

In a theory of pragmatics conceptualized as speaker meaning, evaluation, as a speaker's attitude or stance towards the entities they talk about, should figure prominently. The next two chapters survey and present corpus research into this key pragmatic area.

One type of evaluation that corpus linguistics can genuinely claim to have discovered is the topic of **Alan Partington**'s chapter, viz. 'evaluative prosody', where the term 'prosody', borrowed from phonology, captures the fact that speakers co-select lexical items depending on their evaluation of the affairs mentioned. For example, a speaker who wishes to

characterize an entity as 'full of' something positive, such as *hope*, is likely to use the collocation *brimming with*, whereas the idea of 'full of' is more likely to be expressed by *fraught with* (e.g., *fraught with danger/risk/hazards*) if the speaker's evaluation of the entity is negative. Based on Hoey's (2005) theory of lexical priming, Partington argues that rather than having intrinsic context-free meaning, evaluative prosodies have primings as to how to use them, in what contexts, positions, collocations, etc. Also, Partington stresses the discourse-organizing function of evaluative prosody, where the co-selection of items with the same evaluative polarity helps not only to establish evaluative consistency but also contributes to the cohesion of the discourse (for the text-structuring effect of evaluative prosody, see also Bublitz 2003).

Evaluation, like many pragmatic phenomena, is undoubtedly "a domain of low certainty and high complexity" (Caffi & Janney 1994: 326). The uncertainty and complexity is such that it is often easier, for the (corpus) pragmaticist, to tackle not evaluation itself, which may be intractable, but its epi-phenomena. In informal conversation, one such epi-phenomen of evaluation is the use of tails, as in *People said they'd never, never catch on, teabags*, where an element ('teabags') is placed after the clause which is co-referntial with the pronoun within the clause ('they'). In his chapter, **Ivor Timmis** adds a sociopragmatic dimension to previous corpus-pragmatic descriptions by comparing tails in three different corpora: the Irish component of the *ICE Corpus*, the *British National Corpus* (BNC), and the *Bolton Corpus*; this latter corpus is a historical corpus comprised of written records of snippets of conversations in Bolton, a northern English industrial town, in the period 1937-1940.

5.5 Corpora and reference

Reference has puzzled philosophers and linguists for a long time: what is it about language that we can use it to communicate about things in the world around us? While lay wisdom has it that this is achieved by the ability of words to stand for the things they denote, pragmatic theory emphasizes the role of the speaker: referring "is not something an expression does; it is something that some one can use an expression to do" (Strawson 1950: 326). Referring is thus not a property of words but an act by the speaker (cf. Yule 1996). To resolve reference successfully the hearer's active cooperation is indispensable. This is not only the case with utterances or sentences that include deictic expressions, whose interpretation is especially context-sensitive, but it also holds for seemingly context-independent expressions. For example, as Schwarz-Friesel & Consten (2011) argue, in order to fully understand the sentence 'She dug a hole in the ground' "some slot-filling or referent-creating operation involving WITH AN INSTRUMENT , TYPICALLY A SHOVEL has to be performed" (ibid. : 351). On this view, reference resolution is an interactional achievement: speakers need not fully spell out the intended reference verbally because they can rely on the hearers to instantiate concepts through inferential processing and thus fill in missing verbal reference.

In the first of two chapters dealing with reference, **Christoph Rühlemann** and **Matthew B. O'Donnell** present corpus research into deixis. Core deictic forms such as the personal pronouns *I* and *you* figure high up among the top most frequent forms in many spoken corpora (cf. O'Keeffe et al. 2011: 44). Surprisingly, despite its ubiquity, deixis is "one of the most empirically understudied core areas of pragmatics" (Levinson 2004: 97). In recent years though, beginnings have been made in corpus research to remedy this neglect (e.g., Rühlemann 2007, Clancy 2011). Rühlemann's & O'Donnell's analysis in the present volume targets the referential patterns that demonstrative *this* enters into in conversational narrative. The analysis focuses on the so-called 'introductory *this*', where *this* is used to introduce a discourse-new referent into incipient storytelling. Given that referents marked by introductory *this* can be observed to play a leading role in the unfolding narrative the authors conclude that

introductory *this* acts as a 'theme marker', and is thus best understood as a form of discourse deixis.

In viewing reference as an interactional phenomenon in which "a basic 'intention-to-identify' and a 'recognition-of-intention' collaboration [is] at work" (Yule 1996: 17) the role of inference on the part of the hearer is paramount. One area in which the hearer's inferencing becomes particularly relevant is vagueness, understood as referential underspecification. A prototypical example is use of a general extender such as 'and that', as in 'He'll have a drink at a party an' that' (said in response to the question whether the speaker's husband drinks much) (Aijmer 2013: 128). Here, the extender cues the hearer to activate the category 'social event' and to infer that the husband is a social drinker (cf. Dines 1980). Interestingly, when underspecifiying in this way speakers need not be seen as flouting Grice's Maxim of Quantity, which demands that communication should be neither over- nor under-informative. Rather, speakers know their interlocutor will know the category; given this shared knowledge (and the mutual knowledge of the knowledge) there is no need to specify: specification would be redundant. In their chapter on vagueness, **Winnie Cheng** and **Anne O'Keeffe** survey the large body of work on vague language and present a variational case study on the approximator *about + n*, where n stands for 'number', as in 'about four or five'. The corpora used cover Hong Kong English and Irish English. Cheng and O'Keeffe discover that both the frequencies and functions of the approximator *about + n* are essentially the same in the varieties studied. Given the two varieties' extreme distality, this sameness underscores the universality of vagueness and even of some of the forms used to indicate it.

5.6 Corpora and turntaking

Conversation is generally acknowledged to be the prototypical type of language use. Researchers in the tradition of Conversation Analysis have demonstrated that conversation is

a highly structured interaction. The center piece of conversational organization is turntaking, that is, the distribution of a scarce and sought-after commodity: the 'floor', "which can be defined as the right to speak" (Yule 1996: 72). Turntaking operates according to a 'local management system' (Levinson 1983: 297) which is made up of "a set of conventions for getting turns, keeping them, or giving them away" (Yule 1996: 72). Participants use a broad range of cues to let each other know when it may become appropriate for them to take or yield the floor. The following three chapters are devoted to corpus pragmatic research into the mechanisms by which the local management systems is set into motion.

The chapter by **Gunnel Tottie** looks into the role in turn management of filled pauses, referred to as UHM. Filled pauses have been extensively researched in corpus analyses. They have been found to fulfill three functions: as signals for taking, holding, or yielding the turn. The aim of Tottie's case study based on the *Santa Barbara Corpus of Spoken American English* (SBCSAE) was to study more closely the correlation between turntaking function and position in the turn. The methodology was painstaking: instances of pauses were coded manually according to whether they appeared initially, medially, or finally in a turn. The results of the analysis only partly concur with the results of earlier studies: Tottie found support for the notions that turn-initial UHM has a turntaking function (occurring most frequently in responses to questions) and that turn-final UHM has a turn-yielding function. As regards turn-medial UHM, however, by far the largest category in her data, the turn-holding function usually ascribed to pauses occurring in this position, does not seem the most obvious. Contrary to previous research which has viewed turn-final UHM as "a speaker's last effort" (Stenström 1990:249) when faced with competition for the floor from other participants, Tottie argues that turn-medial pauses usually are best seen as symptoms of planning. Overall, her findings suggest that participants may be guided more by a concern for the continuity of the conversation than by competitive turntaking ambitions.

In their chapter, **Pam Peters** and **Deanna Wong** explore the role of high-frequency backchannels such as *mm* and *yeah*. Previous research emphasized their function as 'continuers', that is, as signals by non-current speakers acknowledging "that an extended unit of talk is underway by another [speaker] and that it is not yet, or may not yet be (…) complete" (Schegloff 1982: 81). Using acoustic analysis of telephone conversations from the Australian ICE corpus, Peters and Wong discover subtle differences in the durations of high-frequency backchannels, and the intervals before them, depending on whether the backchannel occurs as a standalone, or first in the string, or last before a change of turn. They demonstrate that due to these subtle differences the backchannel *yeah* may either signal turn continuation or turn change. The authors conclude that backchanneling plays a larger and more complex role in turn management than has so far been recognized.

While the 'floor' as a valued resource is often seen as inviting competition, turntaking can also be a collaborative effort: participants can even share turns (Schiffrin 1987). To conclude the section on 'Corpora and turntaking', and indeed the volume as a whole, **Brian Clancy** and **Mike McCarthy** investigate co-constructed turntaking, that is, their investigation focuses on complex turns which are co-constructed by two (or more) speakers in that the second-speaker turn expands or completes the first speaker's turn. Based on data from the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE) and the *Limerick Corpus of Irish English* (LCIE), the authors investigate co-construction in the use of the *if*/*when-then* pattern and sentential *which*-clauses. They conclude that these syntactic resources are 'conventionally sanctioned' opportunities for turn co-construction. Building on Sacks (1992, Vol. II: 651), who saw co-construction as "an extremely frequent and routinely doable thing", the authors also point out that the collaborative construction of turntaking is a driving forces of maintaining conversational flow.

References:

Adolphs, S. (2008). *Corpus and context. Investigating pragmatic functions in spoken discourse.* Amsterdam/Philadelphia: John Benjamins

Aijmer, K. 2013. *Understanding pragmatic markers. A variational pragmatic approach.* Edinburgh: Edinburgh University Press

Biber, D., S. Conrad and R. Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bublitz, W. and N. R. Norrick. 2011. 'Introduction: the burgeoning field of pragmatics', in Bublitz, W. and N. R. Norrick (eds.) *Handbook of pragmatics. Vol. 1 Foundations of pragmatics*. Berlin: Mouton de Gruyter, pp. 1-20

Bublitz, W. 2003. 'Emotive prosody: how attitudinal frames help construct context'. In Mengel, Ewald, Hans-Jörg Schmid and Michael Steppat (eds). *Anglistentag 2002 Bayreuth, Proceedings*. Trier: Wissenschaftlicher Verlag, pp. 381-391.

Caffi, C. and R. W. Janney. 1994. 'Towards a pragmatics of emotive communication', *Journal of Pragmatics* 22: 325–73.

Clancy, B. 2010. 'Hurry up baby son all the boys is finished with their breakfast': *A sociopragmatic analysis of Irish settled and traveller family discourse.* Unpublished doctoral thesis, Mary Immaculate College, University of Limerick, Ireland

Cook, G. 1990. 'Transcribing infinity: Problems of context presentation', *Journal of Pragmatics* 14: 1–24.

Crystal, D. 1969. *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Danescu-Niculescu-Mizil, D., M. Sudhof, D. Jurafsky, J. Leskovec, C. Potts ACL. 2013. 'A computational approach to politeness with application to social factors', available at: http://www.mpi-sws.org/~cristian/Politeness.html (last accessed October 2013)

Dines, E. R. 1980. 'Variation in discourse 'and stuff like that'', *Language in Society* 9: 13-33
Felder, E., M. Müller and F. Vogel (eds.) 2011. *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen*, Berlin/Boston: Walter de Gruyter

Garcia, P. (2007). 'Pragmatics in academic contexts: A spoken corpus study.' In: M.C. Campoy & M.J. Luzón (eds.) *Spoken corpora in applied linguistics*. Bern: Peter Lang, 97-128

Hoffmann, S., Evert, S., Smith, N., Lee, D. and Berglund Prytz, Y. 2008. *Corpus linguistics with BNCweb – A practical guide.* Frankfurt/Main: Peter Lang

Gries, S. Th. 2009a. *Quantitative corpus linguistics with R. A practical introduction*. New York/London: Routledge

Gries, S. Th. 2009b. *Statistics for linguistics with R. A practical introduction*. Berlin: de Guyter Mouton.

Gries, S. Th. 2010. 'Methodological skills in corpus linguistics: A polemic and some pointers towards quantitative methods'. In T. Harris & M. Moreno Jaén (eds.), *Corpus Linguistics in Language Teaching*. Frankfurt am Main: Peter Lang, pp. 121-146.

Hoey, M. 2005. *Lexical priming: A new theory of words and language*. London/ New York: Routledge.

Hundt, M., N. Nesselhauf and C. Biewer (eds.) 2007. *Corpus linguistics and the web*. Amsterdam/New York, NY: Rodopi

Jucker, A. H., D. Schreier, and M. Hundt. (eds.). (2009). *Corpora: Pragmatics and discourse.* Amsterdam: Rodopi

Kallen, J. L. and J. Kirk. 2012. *SPICE-Ireland: A user's guide*. Belfast: Cló Ollscoil na Banríona

Knight, D. and Adolphs, S. (2008) Multi-modal corpus pragmatics: the case of active listenership. *In* Romero-Trillo, *J. (ed.) Pragmatics and corpus linguistics. A mutualistic entente.* Berlin and New York: Mouton de Gruyter, pp.175-190

O'Keeffe, A., B. Clancy, and S. Adolphs. 2011. *Introducing pragmatics in use*. London/New York: Routledge

Leech, G. 1983. *Principles of pragmatics.* London: Longman

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2010. *Quantitative Analysis of Culture Using Millions of Digitized Books*. *Science* (Published online ahead of print: 12/16/2010)

Morris, C. 1938. 'Foundations of the theory of signs', in O. Neurath, R. Carnap and C. Morris (eds.) *International encyclopedia of unified science*, Chicago: University of Chicago Press, pp. 77-138

Romero-Trillo, J. (ed.) (2008). *Pragmatics and corpus linguistics. A mutualistic entente.* Berlin and New York: Mouton de Gruyter

Rühlemann. C. 2007. *Conversation in context: A corpus-driven approach*. London: Continuum

Rühlemann, C. and M. B. O'Donnell. 2012. 'Introducing a corpus of conversational narratives. Construction and annotation of the Narrative Corpus', *Corpus Linguistics and Linguistic Theory* 8 (2): 313-350

Scott, M. and C. Tribble. 2006. *Textual patterns. Key words and corpus analysis in language education*. Amsterdam/New York: John Benjamins

Schegloff, E. 1982. 'Discourse as an interactional achievement: Some uses of 'uh huh' and otherthings that come between sentences,' in D. Tannen (ed.) Georgetown University round table on languages and linguistics analyzing discourse: text and talk. Washington DC: Georgetown University Press, pp. 71–93

Schwarz-Friesel, M. and M. Consten. 2011. 'Reference and anaphora', in Bublitz, W. and N. R. Norrick (eds.) *Handbook of Pragmatics. Vol. 1 Foundations of Pragmatics*. Berlin: Mouton de Gruyter, pp. 347-372.

Sacks, H. 1992. *Lectures on Conversation*. Vols. I & II. Cambridge: Blackwell

Schiffrin, D. 1987. *Discourse markers.* Cambridge: Cambridge University Press.

Sinclair, J. McH. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press

Stiles, W. B. (1992). *Describing talk. A taxonomy of verbal response modes*. Newbury Park/CA: Sage Publications

Strawson, P. F. 1950. 'On referring', *Mind* 59: 320-344.

Taavitsainen, I. and A. H. Jucker (eds). (forthcoming). *Diachronic Corpus Pragmatics*. (Pragmatics & Beyond New Series). Amsterdam: John Benjamins.

Tognini Bonelli, E. 2010. 'The evolution of corpus linguistics', in A. O'Keeffe and M. McCarthy (eds.) *The Routledge handbook of corpus linguistics*, London/New York: Routledge , pp. 14-27

Yule, G. 1996. *Pragmatics.* Oxford: Oxford University Press

[i] The example is presented in a simplified version where only the c5 tag is given. In the original files in the BNC-C, 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', looks then like this:

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