22. Corpus-based pragmatics II: Quantitative studies

Christoph Rühlemann

Introduction

Due to the massive dependence of pragmatic phenomena on context, corpora, as a relatively decontextualized medium, have long been seen by some researchers as unfit for use in pragmatic research. Nonetheless, corpus linguistic analyzes, both qualitative and quantitative in orientation, have produced a wealth of new insights into key pragmatic phenomena. The aim of this paper is to illustrate key quantitative corpus studies into phenomena of pragmatic interest. The paper is divided into six sections. The opening section addresses the question of how context-sensitive corpora are. Section 2 presents a case study into semantic prosody, an attitudinal phenomenon at the semantics/pragmatics interface. Section 3 presents a relevance-theoretic study of the pragmatic marker *like*. Section 4 is concerned with quantitative studies on reference. Section 5 introduces corpus research into speech acts. The concluding section looks to the future, outlining recent attempts at building multimodal corpora.

1. Pragmatics and quantitative corpus linguistics: a troubled relationship?

Pragmatics is concerned with meaning in context. Because speakers can mean more than they say, pragmatics is "the art of the analysis of the unsaid" (Mey 1991: 245). The relationship between pragmatics, thus understood, and corpus linguistics is seen by some as a troubled one. The reason is simple: corpora record text, not meaning, and they record context only crudely, particularly in spoken corpora.

Spoken corpora are based on transcriptions made from audio recordings. Since tape recorders cannot filter out non-speech noises, thus selectively 'listening' to speech – as humans can – many, and often large, stretches of corpus transcripts may be inaudible or unintelligible. Further, the information provided in spoken corpora about the contexts in which the spoken texts were produced is fairly minimal; we get to know who (sex, age, class, etc.) was talking to whom when. Moreover, other types of context are recorded only abstractly: we learn in what kinds of setting the talk occurred, or what type of interaction it was. Yet another type of 'abstract' context, which is considered so crucial that in most large corpora the data are categorized according to it, is 'type of situation', or register, such as academic writing, fiction, and conversation. The types of context largely missing from

spoken corpora include the almost infinite wealth of concrete situational, nonverbal, and social context that conversationalists in their specific contexts of situation are connected to (cf. Cook 1990). In sum, we may have impoverished textual evidence, and only rudimentary or abstract contextual evidence. Therefore, corpora have long been seen by some researchers as unfit for use in pragmatic research.

Indeed, some pragmatic features inevitably escape corpus linguistic analysis. This is because in part-of-speech (POS)-tagged corpora, only those phenomena can be studied fully whose lexical form(s) and pragmatic function(s) display a straightforward one-to-one relationship. This relationship is found, for example, in the words *sorry* and *pardon*, which are regularly included in apologies (cf. Jucker 2009). The form-function match is already weaker in compliments which need not necessarily be realized using typical conventionalized patterns (cf. Jucker et al. 2008). Where there is a complete form-function mismatch, as in cases of conversational implicature, a quantitative corpus study will be useless: what listeners take to be implicated in an utterance cannot be retrieved exhaustively from a corpus but can only be inferred (*post hoc*) with varying degrees of confidence.

What a corpus *can* do even in those cases where the form-function mismatch of a phenomenon prevents exhaustive searches, is provide the analyst with illustrative examples that are not only attested and, in this sense, authentic but also embedded in their co-texts, thus giving *some* evidence of the context in which they were used. Such corpus illustrations can usefully complement, or even replace, the invented and often *completely* decontextualized examples that have formed the basis of much pragmatic enquiry.

Another approach to studying pragmatics corpus-linguistically is to use pragmatically-annotated corpora. For example, a small subcorpus of the *Michigan Corpus of Academic Spoken English* (MICASE) is tagged for some speech acts (cf. Maynard and Leicher 2007) and in the *Corpus of Verbal Response Mode* (VRM) *Annotated Utterances*, all utterances are coded twice: once for their literal meaning and once for their pragmatic meaning, using a principled taxonomy of speech acts (cf. Stiles 1992). However, the number of such pragmatically-annotated corpora is still small and they are coded for selected aspects of pragmatic interest only.

Given these limitations, it may be surprising that quantitative corpus analyses of pragmatic phenomena have grown into a large body of literature and produced a wealth of new insights. In the following sections I will present some corpus linguistic studies into pragmatic units. The approach will be selective: only a few key studies can be presented. References to related studies, however, will be provided. Unless otherwise indicated, the illustrative examples will be taken from the *British National Corpus* (BNC), a 100-million-word corpus of contemporary British English (cf. Hoffmann et al. 2008).

2. The semantics/pragmatics interface: semantic prosody

One of the clearest strengths of corpus linguistics is the analysis of what Sinclair (1991) refers to as the 'idiom principle', that is, broadly, how language patterns at the level of phraseology. A very large body of corpus linguistic literature points to 'idiom' phenomena such as collocation, collostruction, colligation and so forth as important building blocks of phraseological patterning. Another, crucial, phraseological phenomenon is 'semantic prosody', which I will focus on in this section. Semantic prosodies probably best illustrate Channell's (2000) claim that some pragmatic phenomena can only be revealed in studies of large corpora. This is because semantic prosodies are normally hidden not only from introspection but also from observation of small numbers of examples. I will first briefly outline essential characteristics of semantic prosody and then illustrate the phenomenon in a case study of *BREAK out*.

Semantic prosody is closely related to 'semantic preference', which is defined by Stubbs as "the relation, not between individual words, but between a lemma or word-form and a set of semantically related word forms" (2001: 65). Stubbs cites the example of the adjective *large*, which often collocates with nouns denoting 'quantity' and 'size' such as *number*, *scale*, *part*, *amounts*, etc. Semantic prosody, on the other hand, can, as a rule of thumb (Partington 2004: 149), be seen as a subcategory of semantic preference. The distinguishing feature of semantic prosody is that the type of semantic preference is related to polarity: the collocates of a word are either typically positive or, more often, negative. Because of the inherently evaluative nature of semantic prosody, alternative terminologies have been proposed, such as 'pragmatic meaning' (Channell 2000), 'emotive prosody' (Bublitz 2003) or 'evaluative prosody' (Morley and Partington 2009). Further, because of the evaluation they convey, semantic prosodies are said to be "on the pragmatic side of the pragmatics/semantics continuum" (Sinclair 2004: 34).[1] Bublitz (1996) and Morley and Partington also note the cohesive role of semantic prosody, which the latter describe as "the mechanism which shows how one elemental type of meaning – evaluative meaning – is frequently shared across units of discourse and, by ensuring consistency of evaluation or *evaluative harmony*, plays a vital role in keeping the discourse together, in its cohesion" (2009: 139). An illustrative example from the *Corpus of Contemporary American English* (COCA) (cf. Davies 2009) is (1):

(1) Hello, everyone. I am Dr. Sanjay Gupta. Your health, there is nothing more important. # And now that **winter weather** has **set in**, there are certain health and safety issues tat [sic] you need to be more concerned about, including how to avoid succumbing to sickness. (COCA: CNN_YourHealth 2002)

This example involves the phrasal verb *SET in*, of which Sinclair (1991) observed that its subjects "refer to unpleasant states of affairs" (Sinclair 1991: 74). Sinclair cites as main vocabulary for *SET in* words like *rot*, *decay*, *malaise*, *despair*, *ill-will*,

etc. In (1), the subject is *winter weather*, a rather neutral phrase. However, winter weather *is* negatively evaluated. This becomes clear when we consider the context: *health and safety issues*, *need to be concerned about*, and *succumbing to sickness* are clear indices of the presence of negative evaluation. The evaluative harmony, to use Partington's term, in the excerpt is thus ensured by the accumulation of negative items. See also Louw's (1993: 173) observation that "in many cases semantic prosodies 'hunt in packs' and potentiate and bolster one another", an observation Bublitz (2003: 387) refers to as "the need to establish a common emotive ground by accumulating equi-polar means."

Excerpt (1) is illustrative of yet another crucial aspect of semantic prosody: it can serve as an effort-saving device "lighten[ing] the burden on the listener and free[ing] the speaker of the tedium of labouring a point" (Morley and Partington 2009: 144). This is, in Louw's (1993: 157) words, because semantic prosody is defined as "an aura of meaning with which a form is imbued by its collocates." In other words, because *SET in* typically co-selects negative subjects, the negativity of its collocates 'colours' the meaning of *SET in* in such a way that it is perceived as negative even if it co-occurs with a subject that is not clearly negative, such as *winter weather*. Louw's notion of 'imbuing' has met with criticism. Whitsitt (2005), for example, argues that "there is no evidence for assuming that we can see the results of a diachronic process of *imbuing*" (2005: 296; emphasis in original; see also Bednarek 2008; Morley and Partington 2009). Whitsitt cites as evidence words such as *alleviate* and *heal*, which habitually co-occur with clearly unpleasant words and yet "certainly [do] not come to have an unpleasant meaning because of that company" (2005: 297). On the other hand, the fact that speakers/writers can achieve certain rhetorical effects such as irony by deliberately deviating from typical collocational patternings, as in *outbreak of sanity,* shows that the "underlying semantic prosody clearly persists, even and especially, in collocational clashes" (Morley and Partington 2009: 150). How can this evaluative persistence be explained?

An answer that has recently been given is the theory of 'priming' (Hoey 2005). As a word, such as *SET in*, is learnt "through encounters with it in speech and writing, it is loaded with the cumulative effects of those encounters such that it is part of our knowledge of the word that it co-occurs with other words" (Hoey 2003), such as, in the case of *SET in*, heavily negative items like *rot*, *decay* and so on. That is, in (1), the speaker need not spell out that winter weather is taken as bad, for example by adding a negative adjective such as *cold*, *severe*, or *harsh.* S/he can trust the addressee to know the primings of *SET in*, that is, a set of 'instructions' on how to use it, how it normally interacts with other items. Among these primings is a word's semantic prosody, "an instruction, which tells the reader 'when you find me in a text, read the surrounding discourse in a favourable/infavourable light, unless there's something else around which tells you not to'" (Partington forthcoming).

The evaluation conveyed by semantic prosodies may be obvious or 'hidden to the naked eye'. Morley and Partington (2009: 151) suggest that it is "best con-

sidered as a cline", with items such as *murder* and *venerable* on the 'overt-evaluation' end expressing fairly clearly unfavourable and favourable evaluation, items such as *peddle* and *fraught with* somewhere in the middle, and items such as *SET in* and *not + BUDGE* (see below) occupying the 'covert-evaluation' end where the evaluation can only be identified using corpus linguistic methods.

The linguistic analysis and description of semantic prosody faces several challenges. One problem is that while it may be easy to achieve a consensus that words such as *rot*, *decay* etc. describe negative states of affairs, the detection of negative or positive semantic prosody may be much less straightforward in other cases. Whether a word is used and interpreted positively or negatively depends in part on the wider context in which it is used and in part on the speaker who uses it. Stubbs (1996) cites the example of *intellectuals*, whose collocates – e.g., *activist*, *student*, *leftwing*, *liberal* – may have a negative 'ring' in some circles but be positively evaluated in others (cf. Bednarek 2008: 122). Semantic prosody therefore "cannot be 'objectively' derived from corpus data and requires a lot of inference on the part of the analyst" (Bednarek 2008: 132).[2]

A second problem is the evaluative inconsistency of semantic prosodies. Prosodies can be 'switched off' or even reversed. Morley and Partington (2009: 149) discuss the example of *not + BUDGE*, which usually carries a negative prosody. However, under certain circumstances, for example when the subject is first person, it can be used favourably. An example of positive evaluation in the use of *not + BUDGE* is given in Hunston (2007):

(2) The Prime Minister rejected resounding calls for the resignation of the government, "I will not budge", he said

Further, the distribution of semantic prosody and word form may be 'asymmetrical' (Bublitz 1996). That is, it may be misleading to claim that 'verb X' or 'noun Y' has positive or negative prosody because in many cases different forms of one and the same word have different evaluative tendencies. Consider the phrasal verb *BREAK out*: when used transitively, as in *Just as linotype operators at the Sun were* breaking out *the champagne to celebrate the arrival of £1,000 pay packets* (CHU 308), its prosody is clearly positive (cf. Louw 2000); when used intransitively, as in *Fires keep* breaking out (A18 1206), its prosody is decidedly negative (cf. the discussion below).

Another problem is that semantic prosodies can be subject to variation across registers. O'Halloran (2007: 15) demonstrates that, for example, the past tense form *erupted* "has largely positive associations in the sports report register, but largely negative ones in the hard news register." That is, the non-register specific concept of semantic prosody may, in some cases, have to be replaced by the concept of 'register prosody', which indicates that "some prosodies have probabilistic relationships to register" (O'Halloran 2007: 4; see also Bublitz 1996).

Finally, semantic prosody cuts across the traditional lexis/grammar dichotomy. The *GET*-passive has been shown to have a negative prosody in that it typically at-

tracts past participles sharing an 'adversative' core meaning (Carter and McCarthy 1999; Stubbs 2001; Rühlemann 2007b). Typical examples are *GET stuck*, *GET caught*, and *GET killed*. Further, semantic prosody may also be observed at a level below the word. Rühlemann (2006) demonstrates that the prefix *dys-*, as in *dyslexia*, *dysfunction*, *dysplasia*, etc., as well as the prefix *dis-*, when used as a productive morpheme, as in *disease*, *disappear*, *disabled*, etc. form words that have negative prosody.

In the remainder of this section a case study into the pattern '(inanimate) N + (intransitive) *BREAK out*' will bring into relief some of the above said.

A search for the verbal lemma*BREAK* immediately followed by *out* finds 1,126 occurrences in the BNC. (3) shows a random excerpt from the concordances. As is suggested by the inanimate nouns preceding the string (underlined in (3)), *negative* things such as hostilities, war, mayhem etc. break out.

(3) if hostilities ever **broke out**. We've
When war **broke out** he returned to
complete mayhem **broke out**. Hats were flung
a fight **broke out** at the Zuwaya
resembling panic **broke out** among the ladies

Is this association with negative items systematic? Table 1 presents the 15 most frequent nouns preceding *BREAK out* in the BNC; the top ten Z-scores, which measure the strength of attraction between collocates, are shaded.

Table 1. Top 15 nouns collocating with *BREAK out*

No.	Collocate	Freq.	Z-score
1	war	259	167.8
2	fire	89	82.8
3	fighting	46	113.8
4	fight	39	75.3
5	row	36	48.5
6	world	32	12.6
7	violence	20	29.4
8	blaze	17	72.4
9	riot	17	44.7
10	argument	17	16.3
11	hostility	16	43.5
12	scuffle	14	132.3
13	rioting	13	82.1
14	way	12	1.1
15	sweat	11	37.2

As can be seen from Table 1, *BREAK out* most commonly collocates with inanimate nouns that share physical violence and, to a lesser degree, fire meanings. *War*, the most extreme manifestation of violence, is not only by far the most frequent collocate but also has the highest Z-score; note also that the 259 occurrences of *war + BREAK out* account for more than a fifth of all occurrences of *BREAK out*. *War* is thus a very strong collocate of *BREAK out*. Other nouns in the *violence* group include *fighting* (rank 3), *fight* (4), *row* (5) etc.; *world*, ranked 6th, occurs in combination with *war* forming the compound *world war.* The *fire* group includes the nouns *fire* (2) and *blaze* (8). Inspection of concordance lines suggests that it is invariably destructive, not warming or cosy, fire that breaks out. Since destructive fire is an inevitable concomitant of war, we might subsume the fire meanings under the heading of the violence meanings, defining the negative prosody of *BREAK out* thus: *BREAK out* co-selects inanimate nouns that typically express violence (including destructive fire) meanings. *BREAK out* illustrates that the concepts of semantic prosody and semantic preference are not mutually exclusive in that the phrasal verb is a semantic prosody with a clear semantic preference.

Interestingly, there are six occurrences of *peace + BREAK out* in the BNC. Inspection of the concordance lines suggests that covert evaluation is the driving force behind these uses. Consider (4) (from the *Economist*, 1991):

(4) Where will the dollar head when *peace* **breaks out**? (ABH 2961)

In (4), the prospect of peace (at the time of the Gulf war) is taken as a threat to the dollar, which has risen before the war but, it is suggested, might fall when the war is over. That peace may have bad consequences for the dollar is not made explicit. Rather, the negative 'aura of meaning' surrounding *BREAK out* is tacitly projected onto peace.

In sum, prosodies are a useful resource for the diagnosis of covert speaker evaluation. Further, as Louw argues, there is the same well-calculated collocational deviance in instances of irony (Louw 1993). The prospect, then, is that semantic prosodies may help to computationally uncover irony, a pragmatic phenomenon that has so far escaped quantitative empirical study. Moreover, semantic prosodies may assist in the study of persuasion: "Propaganda, advertising and promotional copy will now be gradable against the semantic prosodies of the whole language" (Louw 1993: 173). Exploiting the diagnostic potential of semantic prosodies for these kinds of speaker meaning has only just begun and more insights may be expected from this promising avenue of research.

3. Pragmatic markers: *like* in a relevance-theoretic perspective

Pragmatic markers are words or phrases that do not add so much to the propositional content of utterances as they metalingually flag how discourse *relates* to

other discourse. Pragmatic markers thus play a crucial role in facilitating processes of pragmatic inference.

Because they are lexically relatively fixed and thus relatively easily retrievable from a corpus, corpus research into pragmatic markers has been extremely productive.[3] This section presents Andersen (2001), a corpus study on a marker that is particular in that it attends to several discourse functions and, not surprisingly, has attracted a wealth of recent research: *like*. Andersen's study, which is based on the *Bergen Corpus of London Teenage Language* (COLT), is notable because the approach to *like* is informed by relevance theory (Sperber and Wilson 1995). This theory holds that human communication is geared towards the maximisation of relevance such that from the wealth of information and stimuli in our cognitive environments we tend to select that information and those stimuli that are most relevant to us in any given communicative situation.

Andersen identifies five broad functions of *like* used as a pragmatic marker. They are illustrated in (5)-(9). In (5), *like* carries out an approximation function indicating that the numerical information is approximate rather than precise; in (6); *like* is in an exemplification function. The common denominator of uses of *like* classified as approximation and exemplification is "that they involve non-identical resemblance between the encoded and the communicated concepts" (Andersen 2001: 237). The third function Andersen identifies is 'metalinguistic use', which concerns the speaker's relation to the proposition "in terms of its formal linguistic characteristics" (Andersen 2001: 243); in example (7), *like* marks a particular word choice (*reminder*) as potentially inexact or inappropriate. Another function *like* can carry out is as 'interpretive use', a term Andersen prefers over the more commonly used term 'quotative' on the grounds that *like* may be used to preface not only presentations of speech, as in (8), but also "gestures and facial expressions that can be seen as metarepresentations of speaker attitude" (Andersen 2001: 254; see also Buchstaller 2008). Finally, Andersen stresses the role of *like* as a hesitational/linking device, that is, its function in accompanying false starts, self repairs and cut-off utterances and in providing a discourse link between syntactically distinct units of discourse (Andersen 2001: 259). In (9), for example, *like* acts as a pause filler granting the speaker planning time (note the co-occurrence with the hesitation form *er*):

(5) I mean I've been in two shops now there's fifty pound difference **like**, you know (KB2 2401)

(6) think it's the way he looks, like, if you know what I mean, you know **like** […] pull his face and **like**, look over glasses (KB1 768)

(7) PS007 >: Oh yes, I was assuming that … erm … a I do, I didn't really set it out as a formal agenda just as a
PS002 >: Just as a note
PS007 >:a reminder, **like**. (KB0 1337)

(8) Yeah that's what I, why, that's what I said to Susanna and she was **like** don't be ridiculous! (KP5 2163)

(9) PS03T >: well you should wipe the outside of the frame didn't you?
PS03S >: er, yes, **like**, you like just flick round it (KBB 6472)

The functional distribution of *like* in the COLT is shown in Table 2:

Table 2. Functional distribution of *like* (cf. Andersen 2001: 266)

Function	**%**
Approximation	21
Exemplification	19
Metalinguistic focus	19
Interpretive use (quotative)	7
Hesitational/discourse link	35

As is shown in Table 2, the hesitational/linking functions in the corpus are most salient, followed by approximation, exemplification, and metalinguistic use. The quotative function, by contrast, is relatively infrequent. This may be contrary to expectations, given the higher rates for quotative *like* in Tagliamonte and Hudson (1999). The discrepancy between Andersen and Tagliamonte and Hudson, however, may be due to the fact that quotative *like* is a recent development that has been spreading to many regional varieties of English (cf. Buchstaller 2008) and that Tagliamonte and Hudson's data were collected in 1996 whereas the COLT data stem from the early 1990s.

Do all of the different functions of discourse marker *like* have a common root? Andersen (2001) interprets discourse marker *like* as a 'looseness marker', that is, as "a signal that the relation between an utterance and its underlying thought is not a one-to-one relationship, but a relation of non-identical resemblance" (2001: 230). As such, it is a highly co-constructive item inviting recipients to collaborate in the negotiation of meaning.

4. Reference

Reference undoubtedly touches on one of the most fundamental questions concerning language: how can we speak and, in speaking, communicate something meaningful *about* things, people, and states of affairs? The answers linguists and, more importantly, philosophers have suggested are by far too varied and complex to go into in sufficient detail in this short section (see Schwarz-Friesel and Consten this volume). A few remarks must suffice.

It is often claimed that referring expressions refer to their referents, such that, for example, the expression 'the present U.S. American president' 'refers' to Barack Obama. On this view, the expression *denotes* an object in the real world; see Russell's (1905) influential paper, entitled "On denoting". Denotation, in Russell, is taken to be "a relation between an expression, considered in abstraction, and the thing that is the expression's referent or denotatum" (Lycan 2008: 19). Russell distinguishes between *meaning* and *denotation*. An expression such as 'the present U.S. American prime minister' would have to be assigned meaning because it could be perfectly understood, but, since the US political system does not provide for a prime minister, it would fail to denote anything. In Russell's view, the sentence "The present U.S. American prime minister is wise" would be considered to be *false*. A number of objections have been brought against Russell's theory of definite descriptions, chiefly by Strawson's (1950) article aptly entitled "On referring". Strawson thought of referring not as an abstract relation between an expression and an object but argued that referring "is not something an expression does; it is something that some one can use an expression to do" (1950: 326). The variable that he brings into the equation is "the context of an utterance […] and by 'context' I mean, at least, the time, the place, the situation, the identity of the speaker, the subjects which form the immediate focus of interest, and the personal histories of both the speaker and those he is addressing" (1950: 336). Whether 'the present U.S. American president' is used to refer to Mr Obama or his predecessor or his successor or any other U.S. American president crucially depends on the circumstances of its use. Also, the sentence 'The present U.S. American prime minister is wise' would not be seen as lacking truth-value; since it builds on a false presupposition, the question whether the statement is true or false simply does not arise. On this view, which links reference intimately to context and, hence, inference, reference is deeply pragmatic.

The following two subsections section present corpus-based research into two types of referring expressions: definite noun phrases and deixis.

4.1. Reference through definite noun phrases

One type of material that corpus linguists standardly work with are frequency lists, that is, lists in which the words in a corpus are ranked in order of their frequency in that corpus.

Table 3 displays what are, according to Kilgarriff (1998), the 20 most frequent items in the conversational subcorpus and the written subcorpus of the BNC.

Table 3. 20 most frequent items in the conversational subcorpus (C) and the written subcorpus (W) of the BNC

	C			W		
Rank	**Frequency**	**Item**	**Tag**	**Frequency**	**Item**	**Tag**
1	167,640	i	pnp	5,776,384	the	at0
2	135,217	you	pnp	2,789,403	of	prf
3	128,165	it	pnp	2,421,302	and	cjc
4	115,247	the	at0	1,939,617	a	at0
5	92,239	's	vbz	1,695,860	in	prp
6	90,886	and	cjc	1,468,146	to	to0
7	77,611	n't	xx0	892,937	is	vbz
8	68,846	a	at0	845,350	to	prp
9	62,382	that	dt0	839,964	was	vbd
10	58,810	yeah	itj	834,957	it	pnp
11	48,322	he	pnp	768,898	for	prp
12	47,391	to	to0	606,027	with	prp
13	43,977	they	pnp	605,749	he	pnp
14	42,241	do	vdb	603,178	be	vbi
15	41,654	oh	itj	590,305	on	prp
16	38,515	what	dtq	580,267	i	pnp
17	35,156	is	vbz	561,041	that	cjt
18	34,901	of	prf	490,673	by	prp
19	34,837	was	vbd	435,574	at	prp
20	34,477	in	prp	426,207	you	pnp

In Table 3, all personal pronouns are shaded pink, while all items related to noun phrases (NP), such as articles and prepositions, are shaded grey. An initial comparison of the shaded cells reveals clear differences in the ways that language users refer in conversation (C) and writing (W) respectively. Among the 20 most frequent items in C, there are (i) more personal pronouns (the pronoun *they* is not included in the top 20 in W) and (ii) far less NP-related items (prepositions and the definite and the indefinite article) than in W: there are ten NP-related items in W but only four in C. Table 3 further shows that the definite article *the* is by far the most common word in writing (roughly twice as common as the next one, the preposition *of*). Indeed, *the* is not only the most frequent word in writing but also the most frequent word in the whole of the BNC (cf. Kilgarriff 1998) and in most other general corpora, such as the *Cambridge International Corpus* (CIC) (cf. McCarthy 1998) or the *Bank of English* (BoE) (cf. Sinclair 1999). With its roughly 6 million

occurrences in the 90-million-word written subcorpus of the BNC, *the* accounts for more than six percent of all word tokens in that subcorpus. It is ironic that, a good 100 years on, Russell's verdict that "to the philosophical mathematician [*the*] is a word of great importance" (1919: 167) should be empirically confirmed by corpus frequency counts.

This evidence from the BNC suggests two major interpretations: (i) the most common referring technique in writing is the use of both the definite NP and, less importantly, the indefinite NP, while (ii) in conversation, reference is expressed most commonly by means of pronouns. While this latter finding will be discussed in the next subsection (5.2), the remainder of this subsection takes a closer look at the use of the definite article both in conversation and writing.

The overriding function of the definite article is to specify "that the referent of the noun phrase is assumed to be known to the speaker and the addressee" (Biber et al. 1999: 263). That is, broadly speaking, *the* functions as a marker of given information. It does so in a variety of ways. In a study carried out on the *Longman Spoken and Written English* (LSWE) *Corpus*, Biber et al. (1999) identify the following reference patterns of definite NPs marked by *the*.[4] They are illustrated by relevant examples:

(10) Anaphoric: A MAN died and a girl was badly injured when fire ripped through a
house yesterday. **The** girl, who had been clinging to a third-floor window ledge, fell just as firemen were about to grab her. (CBF: 2921)
('the girl' is understood as the girl that 'was badly injured when fire …'; reference is backward-looking)

(11) Indirect anaphoric:
A woman died yesterday after being knocked down by a shoplifter
fleeing with a £2.58 descaler. Mrs Lillian Amelia Smith, 81, sustained a fractured skull at the store at Newham, east London, on Friday. Police are treating **the** incident as murder. (AKH: 260)
('the incident' can be identified via inference as the events described in the preceding report)

(12) Cataphoric: The next main point is about **the** complexity of the system. (HHW: 3638)
('the complexity' in question is identifiable via the *of*-phrase following it; reference is forward-looking)

(13) Situational: I think there might be parcel for you at **the** door is it? (FPU: 358)
(based on 'frame knowledge' the reference of 'door' is understood as 'the front door' where parcels are normally delivered)

(14) Generic: Just for once, a Frenchie has conceded that **the** Brits do something better. (A0C 1386)
('the' denotes 'the whole class of Brits' without specific reference to particular Brits)

(15) Idiom: But, quite frankly, what's **the** point? (A5Y: 230)
('the' is an integral part of the idiom 'what's the point?'; it has no referring or denoting function; see also Searle 1969: 72)

Intriguing results emerge from the analysis of the extent of use of these patterns in different registers by Biber et al. (1999). The registers considered include three written registers (fiction, news, and academic writing) and conversation. The results are summarized in Table 4:

Table 4. Percentage use of reference patterns in four registers (C: conversation; F: fiction; N: news reportage; A: academic writing) (cf. Biber et al. 1999: 266)

	C	**F**	**N**	**A**
Situational	55	10	10	10
Anaphoric	25	30	30	25
Indirect anaphoric	5	10	15	15
Cataphoric	5	15	30	40
Generic	>2.5	>2.5	5	5
Idiom	>2.5	>2.5	>2.5	>2.5
Uncertain	10	15	5	5

The findings presented in Table 4 allow for a number of observations. Generally, there are marked cross-register differences in the use of the reference patterns. Specifically, anaphoric reference, which may intuitively be seen as the major reference type of definite NPs, accounts for less than a third in all four registers. Cataphoric reference represents, respectively, 30 percent in news reportage, 40 percent in academic writing, but only five percent in conversation. Situational reference, conversely, accounts for 55 percent in conversation while it is found in only ten percent in the written registers.

The analysis in Biber et al. (1999) also enables us to understand more deeply the relative frequencies of the definite article in the word frequency list discussed above. We saw that *the* is by far the most frequent word in the written part of the BNC but less highly ranked in the conversational part of the BNC. Applying the Biber et al. analysis to these frequency-based rankings, we can now see that definite NP reference is not only relatively less important in conversation, where it is second to reference through personal pronouns, than in writing, where it is by far the most important reference type. We can also see that the difference in the use of the definite NP between conversation and writing is in fact much more dra-

matic because reference through definite NP fulfills different functions in the two modes: while it is largely (indirect) anaphoric and cataphoric and, hence, *endophoric* in writing, it is overwhelmingly situational and, hence, *exophoric* in conversation.

Another type of exophoric reference is through the use of deixis. Its use in discourse presentation will be the focus of the next section. That section will also facilitate a more detailed understanding of the overrepresentation of personal pronouns in conversation we observed in Table 3 above.

4.2. Deixis and discourse presentation[5]

Deictic reference is a prime example of exophoric reference because establishing referents of deictic words necessarily requires extra-linguistic context. For example, the references of the person-deictic word *I* and the time-deictic word *tonight* are completely dependent on the speech situation in which the two words are uttered. Deictic words presuppose a deictic centre ('origo') relative to which they are computed. The deictic centre is in most cases associated with the current speaker. It can also be seen as the 'viewpoint' from which the speaker refers to the dimensions of person, time, and place (Lyons 1977: 638). This viewpoint constrains the use of deictic words such that person, time, and place deictics normally conform to this, one, viewpoint, thus forming a 'deictic system' (Levinson 1983: 68).

One area of language use in which speakers can use deictic words that do not conform to their own deictic system is discourse presentation, alternatively referred to as 'speech reporting' and 'constructed dialogue.' This will be explained with reference to corpus research carried out by McIntyre et al. (2004), a study based on the *Lancaster Speech, Writing and Thought Presentation Spoken Corpus*, a small, balanced corpus of contemporary spoken English drawn from the conversational subcorpus of the BNC and oral history archives from Lancaster University containing elicited interviews. The authors propose categories for the presentation of not only speech but also thought and writing. For space limitations, only the categories as well as the study's initial results for speech presentation will here be reported on.

McIntyre et al.'s (2004) model of speech presentation provides for six major categories. With reference to the examples listed below, the categories include the two direct categories 'Direct', as in (16), and 'Free Direct', as in (17). Both modes represent speech "in the form in which it is directly manifest to a listener" (Leech and Short 1981: 345) in an anterior situation. A distinction between the two is that Direct requires a preceding reporting clause, whereas Free Direct is not accompanied by a reporting clause. The categories further include Indirect, as in (18), and Free Indirect, as in (19). Like Free Direct, Free Indirect has no accompanying reporting clause. In (Free) Indirect mode the propositional content of the

original speech is specified, "but no claim is made to present the words and structures originally used to utter that proposition" (McIntyre et al. 2004: 61). Further, the categories include Representation of Speech Act, as in (20). This category presents "the illocutionary force of an utterance or text (part) with an optional noun or prepositional phrase indicating the topic" (McIntyre et al. 2004: 61) but does not claim to represent the propositional content or the original wording of that content. The final category is 'Representation of Voice', as in (21), which "captures minimal references to speech with no indication of the illocutionary force, let alone the propositional content or form of the utterance (part)" (McIntyre et al. 2004: 62).

(16) Direct (D): Yeah she said erm **you have it my love don't you worry** but I said **no**
I'll give you some no no I don't want no money for it at all she said, **you take it my love**.
(KB6 1355)

(17) Free Direct (FD): [Speaker is reporting how someone asked him/her for change for a fiver].
I said no! […] only. So … **well can you lend me a pound**? I said no!
(KD5 7945)

(18) Indirect (I): And I thought she might have come today cos she said **she'd bring up**
the money but she hasn't, not yet.
(KB6 474)

(19) Free Indirect (FI): Father said can my girls come? **No they couldn't come**
(McIntyre et al. 2004: 60)

(20) Representation of Speech Act (RSA): **The only, the only er thing I complained about with you is,**
is the butter.
(KBC 6240)

(21) Representation of Voice (RV): the children didn't even know **he was talking to them**.
(KB0 375)

Which of these modes is the most frequent in everyday speech? The answer McIntyre et al.'s analysis suggests is straightforward:

Table 5. Proportions of speech presentation categories (McIntyre et al. 2004)

Category	Frequency	Rank
D	38 %	1
FD	4 %	5
I	12 %	4
FI	2 %	6
RSA	27 %	2
RV	17 %	3

As is shown in Table 5, Direct is by far the most common speech presentation mode in McIntyre et al.'s spoken corpus, accounting for more than a third of all presentations. The second most frequent mode is RSA, accounting for more than a quarter, followed by RV, accounting for less than a fifth. Free Direct is relatively underrepresented in the corpus (ranking fifth), a fact which may be due to the non-conversational interview sections in the corpus. In spontaneous conversation, Free Direct may to be much more common (cf. Stenström et al. 2002: 110 ff.).

In order to make sense of the dominance of the direct mode in speech it is important to consider this mode in terms of reference and deixis. A fundamental difference between the direct modes and all other modes lies in the speaker perspective (Coulmas 1986: 2). In Direct and Free Direct, the perspective is that of the presentee: all deictic features are appropriate to the speaker in the anterior situation. By using words which are ostensibly marked as the words of a non-present speaker, the presenter assumes the role of the presentee: he/she enacts the presentee and his/her speech. By contrast, in the non-direct categories (Free) Indirect, RSA, and RV, the perspective is that of the presenting speaker: all deictic features are appropriate to the speaker in the posterior, discourse presenting, situation (cf. McIntyre 2004: 60).

The switch into the presentee's perspective and deictic system that characterizes (Free) Direct has important implications on the referential plane. Each time the perspective and hence the deictic system changes, so do the referents of the same deictics. Prime examples of such referential changes are the pronouns *I* and *you*. The two pronouns are core person deictics referring to people immediately present in the speech situation. (As such, they are distinguished from third-person pronouns such as *he* or *she* which typically refer anaphorically or cataphorically to entities in the text.) Reconsider example (16), reprinted here as (22). The speaker is presenting three utterances: two by a speaker who is not present in the posterior, presenting, situation (utterances 1 and 3) and one she made herself in that situation (utterance 2). For illustration purposes let us call the presenter 'speaker A' and the presentee 'speaker B'. The letters A and B in the excerpt indicate co-referential deictics (presented speech is in bold and deictics are underlined):

(22) 1 Yeah she said erm **you(A) have it my love don't you(A) worry**
2 but I(A) said **no I(A)'ll give you(B) some**
3 **no no I(B) don't want no money for it at all** she said, **you(A) take it my love**.

In the excerpt, a number of switches in deictic reference can be observed. In utterance 1, speaker A uses *you* twice, the person deictic appropriate to speaker B's deictic system to refer to an addressee, in that case, speaker A. Then, in utterance 2, speaker A uses *I* twice to refer to herself (speaker A), and *you* to refer to speaker B thus deploying the deictics appropriate to her own deictic system. Finally, in utterance 3, speaker A switches back into speaker B's deictic system using *I*, whose referent is speaker B, and *you,* whose referent she is herself (speaker A). That is, because in Direct presentation, speakers assume different perspectives and the corresponding deictic systems, reference in Direct mode jumps back and forth between referents: deictic-system oscillation leads to reference oscillation. Further, we need to be aware that such discourse presentational changes in perspective are not isolated events occurring here and there but pervasive – not only in longer conversations but often within one and the same utterance, particularly in narrative.

The study by McIntyre et al. is thus significant on two counts. First, it suggests one answer (out of several) to the question why *I* and *you* are among the most frequent words in conversation, as seen in the previous section.[6] *I* and *you* are so common in conversation because the most frequent discourse presentation mode is Direct: when presenting discourse, conversationalists mostly use the mode in which deictic references – such as *I* and *you* – used by speakers in anterior situations are 'copied and pasted' from that anterior situation into a posterior presenting situation.

Second, McIntyre et al. underscore a view of deixis as a far more flexible system than is often thought. The study provides empirical evidence to suggest that presenters are by no means confined to presenting discourse and the references therein from their point of view, with all deictic references conforming to their origo and the corresponding deictic system. Presenters are free, and make ample use of that freedom, to shuttle between various origos, deictic systems and perspectives thus creating a continuous oscillation of reference. To judge by the commonness of direct speech presentations, presenters perform this referential oscillation with great ease. What about the listeners? How can they resolve the ambiguities in reference that the constant oscillation is likely to bring with it? Clearly, different 'voices' may be marked off by intonational means, speaker change may be signalled by use of reporting clauses such as *he said*, *I says*, etc., 'utterance openers' such as *oh* and *well* may flag the start of a direct speech presentation. But in many cases no such additional 'processing instructions' are being used. Still, listeners seem to process switches in perspective and deictic system equally easily as pres-

enters do. Precisely what enables listeners to resolve reference, which presenters handle with such flexibility, is still underresearched. Therefore, investigating reference resolution in discourse presentation might add valuable insights to pragmatic theories of inference.

5. Speech acts

As noted in the introductory section, the relation between quantitative corpus analysis and speech act analysis is not a one-to-one match because corpora record surface forms while speech acts "are defined on the basis of their function, not their form" (Jucker 2009: 7). However, speech acts are often realised using 'illocutionary force indicating devices' (Searle 1969: 30) or largely routinized forms. Such features have proven useful anchors for corpus searches for speech acts. The number of corpus studies into speech acts is to date still small. Two such studies will be presented in this section, one on the speech act of complimenting and one on the indirect speech act of suggesting.[7]

5.1. Compliments

Compliments have received a fair amount of attention in pragmatic research (for an overview see Jucker 2009). In a pioneering study, Manes and Wolfson (1981), using a 'notebook method' (cf. Jucker 2009), claim specific frequencies for the compliment patterns they found in their data. The study that will be briefly reviewed here, Jucker et al. (2008), aims to assess the accuracy of Manes and Wolfson's findings with the help of the BNC. The study is also intended to highlight methodological problems involved in approaching speech acts via corpora.

The two studies report strikingly different pattern frequencies of compliments. Using search strings which correspond to the compliment patterns established by Manes and Wolfson, hand-searching subsets of matches and extrapolating their frequencies to the entire sets, Jucker et al. conclude that there are approximately 343 compliments in the 100-million-word BNC; Manes and Wolfson, by contrast, collected 686 compliment sequences. Further, Manes and Wolfson established nine compliment patterns; the frequencies Jucker et al. found for these patterns diverge considerably from the frequencies reported by Manes and Wolfson. The frequencies are shown in Table 6:

Table 6. Frequencies in Manes and Wolfson (1981) and Jucker et al. (2008)

Pattern	Manes & Wolfson	Jucker et al.
1	53.6 %	76.4 %
2	16.1 %	3.2 %
3	14.9 %	2.3 %
4 and 6	5.7 %	5.0 %
5	2.7 %	6.4 %
7	1.6 %	3.5 %
8	1.6 %	3.2 %
9	–	1.0 %

The differences are most marked with regard to patterns 1–3. While pattern 1 accounts for slightly more than half of all compliments in Manes and Wolfson's data, this pattern accounts for more than two thirds in the BNC. Even more striking, in relative terms, are the differences for patterns 2 and 3: the frequencies reported in Manes and Wolfson are more than five times as high as in Jucker et al.[8] The three patterns are illustrated in (23)-(25) (optional elements in brackets):

(23) NP+copula+(intensifier)+adjective:
'**You look so beautiful**,' he whispered.
(FSF 1954)

(24) I+(intensifier)+verb of liking+NP
You know **I really like you**!
(KE1 1859)

(25) Pronoun+ copula+(intensifier)+(indefinite article)+adjective+NP
'Another two pounds a week for that?' He looked anxiously at her. 'Oh, **that's very generous**, thank you. That would be marvelous.'
(CCM 1399)

The significance of the Jucker et al. corpus study lies not in an assumed superiority of the corpus method over the notebook (or any other non-corpus linguistic) method and thus in the claim that the Jucker et al. frequencies were more reliable than Manes and Wolfson's. By contrast, Jucker et al. demonstrate and discuss in great detail the merits and, more importantly, shortcomings of the corpus method as an alternative approach to speech acts. As the authors point out, the surface patterns used are crude (Jucker et al. 2008: 290). This is because the 'hits' either include too many irrelevant matches (a problem of 'precision') or they fail to retrieve all relevant examples (a problem of 'recall'). Indeed, precision and recall were very low, ranging between less than 1 and 20 percent. For example, poor recall may be the reason why a much lower number of compliments was found in the BNC: al-

though highly differentiated search strings were used, not all compliments could be retrieved because compliments, particularly in spoken sections, may include some minor deviation from the search string, such as pauses or repairs.

Indeed, as the authors point out, "[a]lmost every query method fails to have complete precision and recall" (Jucker et al. 2008: 276). To solve this problem, large pragmatically-annotated corpora would be needed. These, however, are still in very short supply. Given that the notebook (and any other non-corpus-based) method have their drawbacks too (see Jucker 2009 for a balanced discussion), the corpus approach, even in its present imperfect form, is nonetheless a welcome addition to the variety of methods used to study speech acts.

5.2. Suggestions

Indirect speech acts are "cases in which one illocutionary act is performed indirectly by way of performing another" (Searle 1975: 60), as in *Can you pass the salt?* A much debated question in pragmatics is how listeners resolve the ambiguity between, in this case, question and request. It is often assumed that the adequate interpretation of indirect speech acts requires a complex chain of inferences by which a recipient first decodes the literal force (in this case, that a question has been posed to him/her about his/her *ability* to pass the salt), then realizes that something is 'up' with this question (for example, that it is not in accordance with the Cooperative Principle because it fails to be relevant) and only then infers that the question is not to be taken literally but rather indirectly, as a request by the speaker to be passed the salt. Corpus studies, by contrast, emphasize the role less of inferences but more of collocational patterning. This section will highlight one such study, namely Adolphs's (2008), which is based on the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE). The focus in this study is on 'speech act expressions' introducing suggestions; the expressions investigated include, among others, *why don't you.*

Why don't you can be used both in direct and indirect speech acts: when used to introduce a genuine question, a direct speech act is performed; when used to make a suggestion, a question form is used to 'put forward a proposal for consideration' by the addressee (in Searle's taxonomy of speech acts suggestions would count as 'directives'):

(26) PS52K>: I don't believe that.
PS52C>: **Why don't you** believe it? It's a survey
(KP6 1737)

(27) PS1C1>: but I've got nobody to go with!
PS1JA>: Oh! **Why don't you** come with us?
(KDW 2752)

The majority of uses of *why don't you*, Adolphs observes, perform a suggestion. She elaborates a functional profile of *why don't you*, consisting of its collocations, the discourse factors bearing upon its use and its contextual distribution. Thus, she observes that suggestion-*why don't you,* henceforth S-WDY, is often preceded by a form of SAY introducing speech presentation, as in (28). No such association with speech presentation is reported for question-*why don't you*, henceforth Q-WDY.

(28) oh I think it's because I was *saying* **why don't you** come up like this week (KDM: 7460)

As to right-hand collocates, at N+1 (that is, in the 'slot' immediately after *why don't you*) S-WDY is regularly followed by a group of transitive verbs including *ask*, *get*, *tell*, and *use*. The set of verbs that Q-WDY collocates with is distinctly different: they include *like*, *want*, and *have to*. Another key collocate is also found at N+1: the marker *just* which, in association with S-WDY, serves to down-tone the imposition implied in suggestions (remember that suggestions are a type of 'directives'). Consider:

(29) Why **don't you just** sit down somewhere? (KBH 4400)

Further, S-WDY and Q-WDY are distinguished by the type of response they typically trigger: while Q-WDY require a "more detailed answer" (Adolphs 2008: 62), the responses to S-WDY "range from minimal acknowledgment tokens to agreement, or evaluations of the suggestion" (ibid.). Finally, Adolphs found that S-WDY was most frequent in the 'intimate' speaker relationship category in the CANCODE, that is, it occurs most frequently in interactions between people whose social closeness is maximal (partners, family, very close friends). In this context type, speakers are most 'off-guard' and the imposition implied in S-WDY as a directive is felt as less threatening.

Thus, we see that the uses of Q-WDY and S-WDY are, indeed, 'idiomatic' in the sense that a large number of distinctly different co-textual and contextual factors bear upon their use and it appears plausible to assume that these 'idiom' factors are salient enough for communicative partners to disambiguate the two speech acts.

This is not to say that all indirect speech acts are idioms and do not require inferential processes for their interpretation. What Adolphs's case study shows is merely that some 'indirect' speech acts may be less indirect than rather idiomatic, and it is as yet by no means clear whether the share of such idiomatic speech acts in all possible types of (indirect) speech act is large or small. No doubt, it is still early days for corpus linguistic research into speech acts. However, the beginnings are promising and the prospects are that corpora can make important contributions to speech act theory.

6. Looking to the future

Current corpora facilitate fascinating observations of how words are actually used. However, they fail to represent communication beyond the word. A particular challenge for current research is therefore to integrate corpus linguistic methods and theories of multimodal linguistic research (Carter and Adolphs 2008: 276). At present, corpora targeted on aspects of multimodal communication are both small in size and number. Their central purpose is to facilitate explorations of how meaning is made through 'multimodal patterns', that is, patterns of interaction between verbal and nonverbal choices (Carter and Adolphs 2008: 281).

Research faces major challenges before this goal is reached. Given that nonverbal meaning seems to make up a very large chunk of overall meaning (Birdwhistell 1970: 157–8 estimates that "probably no more than 30 to 35 percent of the social meaning of a conversation or an interaction is carried by the words") it will not be surprising that the nonverbal semiotic systems may be as highly differentiated as the verbal system. For example, Rimé and Schiaratura (1991: 248) present a taxonomy of speech-related hand gestures which includes six broad variables: speech markers, ideographs, iconic gestures, pantomimic gestures, deictic gestures, and symbolic gestures. Ekman and O'Sullivan (1991: 176) discuss evidence for the cross-cultural recognition of at least six emotions expressed via facial actions: happiness, anger, fear, sadness, surprise, and disgust. Variables of prosody include rhythm, volume, tempo, voice quality, and intonation with its manifold subvariables. Obviously, the task of developing a coding scheme to capture such a wealth of individual variables corpus-linguistically is daunting. It becomes even more daunting considering that the variables *interact* rather than act independently of one another. Therefore, an important goal is the development of "tools that provide an *integrated* approach to the representation of the data" (Carter and Adolphs 2008: 283; emphasis in original). Another complicating factor is that verbal and nonverbal choices are hard to align since "within any sequence a substantial number of utterances and gestures made by speaker and hearer overlap" (Carter and Adolphs 2008: 284). That is, unlike speaking turns which are taken 'orderly' in the sense that normally 'one speaker speaks at a time', nonverbal 'turns' are much less restricted: while a speaker is speaking (and acting nonverbally) the listener(s) may produce, in response to the speaker's unfolding utterance, nonverbal signals and actions themselves.

Given these challenges, it is small wonder that current multimodal corpus analyses are decidedly selective, focussing on narrow multimodal phenomena rather than trying to study multimodal patterning in its (at present overwhelming) complexity. One such work in progress is Baldry and Thilbault (2006), who take a systemic-functional approach to analyze gaze in a corpus of TV car advertisements. Another work in progress is Carter and Adolphs's (2008) 'Headtalk' project. This project, which is based on a small corpus of several hours of video-taped MA and

PhD supervision sessions at Nottingham University, is intended to explore the patterning of multimodal backchannels, focussing specifically on head nods as a type of nonverbal backchannel in co-occurrence with verbal backchannels.

To conclude, multimodal corpus linguistics "is very much in its infancy" (Baldry and Thibault 2006: 181). However, since this strand of research offers intriguing prospects for an enhanced description of how speakers mean more than they say, the construction, annotation, and exploitation of multimodal corpora may in future become a major site of corpus pragmatic research.

Notes

1. Corpus linguistic explorations into evaluative language are numerous. They include, among many others; Aijmer (1989) on tails, that is, postponed items succeeding the core of the clause, as in *Pathetic behaviour that is, innit?*, which primarily seem to fulfill an affective-stance function; Biber and Finegan (1988) on adverbial stance types across text clusters; and Norrick (2009) on interjections in narrative.
2. An initial attempt at quantifying "good" and "bad" prosodies without relying on the researcher's subjective, evaluative judgments is Dilts and Newman (2006), who use a method based "on experimentally measured judgments of *good*ness and *bad*ness obtained prior to, and independently of, corpus-based studies" (2006:240; emphases in original).
3. A few selected corpus papers on discourse markers include: Aijmer (1987) on the mental processes signaled by *oh* and *ah;* Stenström (1998) on *cos* as a continuation (rather than a cause) marker; Lenk (1998) on the coherence inducing function of discourse markers. For a recent collection of papers on discourse markers see Jucker and Ziv (1998).
4. Definite NPs can be marked not only by *the* (by far the most frequent definite determiner) but also by possessive determiners (e.g., *his*, *her*), and demonstrative determiners (e.g., *this*, *that*) (Biber et al. 1999:269–10). In Biber et al.'s study, definite NPs were identified using an interactive programme which looked for NPs co-occurring with the definite article only (Biber, personal communication).
5. For deixis and indexicality cf. Hanks this volume.
6. The reasons why *I* and *you* are so highly common in spontaneous speech are undoubtedly manifold: they include planning-induced repetition, as in *But I cert I I I I I ju I it it just sounds* […] (KB7: 3681), use of *I* and *you* in high-frequent discourse markers such as *I know*, *I see*, *you know* etc. and the fact that, in conversation, speaker and addressee "are in immediate contact, and the interaction typically focuses on matters of immediate concern" (Biber et al. 1999: 333) (for a more detailed discussion see Rühlemann 2007a).
7. Further corpus studies into speech acts include Aijmer (1996) on routinized speech act expressions based on the *London-Lund Corpus* (LLC) and Kohnen (2000), a pilot study into explicit performatives based on the *Lancaster Oslo/Bergen* (LOB) *Corpus*, the *London-Lund Corpus* (LLC), and the (historical) *Helsinki Corpus* (HC); for speech acts in general see Collavin this volume.
8. Note that the differences in patterns 5–7 are more important, in relative terms, than the differences in 1, 4 and 6; however, since the percentage values for patterns 5–7 are low, the differences cannot be assigned much significance.

References

Adolphs, Svenja
2008) *Corpus and context. Investigating pragmatic functions in spoken discourse.* Amsterdam: Benjamins.

Aijmer, Karin
1987 *Oh* and *Ah* in English conversation. In: Meijs, Willem (ed.), *Corpus Linguistics and Beyond.* 61–68. Amsterdam/Atlanta, GA: Rodopi.

Aijmer, Karin
1989 Themes and tails: The discourse functions of dislocated elements. *Nordic Journal of Linguistics* 12: 137–54.

Aijmer, Karin
1996 *Conversational Routines in English.* London: Longman

Andersen, Gisle
2001 *Pragmatic Markers and Sociolinguistic Variation. A Relevance-Theoretic Approach to the Language of Adolescents.* Amsterdam: Benjamins.

Baldry, Anthony and Paul Thibault
2006 Multimodal corpus linguistics. In: Geoff Thompson and Susan Hunston (eds.), *System and Corpus: Exploring Connection,* 164–83. London/Oakville: Equinox.

Bednarek, Monika
2008 Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory* 4(2): 119–139.

Biber, Douglas and Edward Finegan
1988 Adverbial stance types in English. *Discourse Processes* 11: 1–34.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan
1999 *Longman Grammar of Spoken and Written English.* Harlow: Pearson.

Birdwhistell, Ray L.
1970 *Kinesics and Context: Essays on Body Motion Communication*. Philadelphia: University of Pennsylvania Press.

Bublitz, Wolfram
1996 Semantic prosody and cohesive company: *somewhat predictable*. *Leuvense Bijdragen* 85: 1–32.

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

Buchstaller, Isabelle
2008 The localization of global linguistic variants. *English World-Wide* 29(1): 15–44.

Carter, Ronald and Svenja Adolphs
2008 Linking the verbal and the visual: new directions for corpus linguistic. *Language and Computers* 64: 275–291

Carter, Ronald and Michael J. McCarthy
1999 The English *get*-passive in spoken discourse: description and implications for an interpersonal grammar. *English Language and Linguistics* 3(1): 41–58.

Channell, Joanna
2000 Corpus-based analysis of evaluative lexis. In: Susan Hunston and Geoff Thompson (eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse*, 38–55. Oxford: Oxford University Press.

Cook, Gu
1990 Transcribing infinity: Problems of context presentation. *Journal of Pragmatics* 14: 1–24.
Coulmas, Florian (ed.)
1986 *Direct and Indirect Speech*. Berlin: Mouton de Gruyter.
Davies, Mark
2009 The 385+ million word Corpus of Contemporary American English (1990–2008+). Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14(2): 159–190.
Dilts, Philip and John Newman
2006 A note on quantifying "good" and "bad" prosodies. *Corpus Linguistics and Linguistic Theory* 2(2): 233–242.
Ekman, Paul and Maureen O'Sullivan
1991 Facial expression: Methods, means, and moves. In: Robert S. Feldman and Bernard Rimé, *Fundamentals of Nonverbal Behaviour,* 163–199. Cambridge: Cambridge University Press.
Hoey, Michael
2003 Lexical priming and the properties of text. Available at: http://www.monabaker.com/tsresources/LexicalPrimingandthePropertiesofText.htm (last visited October 2009)
Hoey, Michael
2005 *Lexical priming. A new theory of words and language.* London: Routledge.
Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee and Ylva Berglund Prytz
2008 *Corpus Linguistics with the BNCweb – a practical guide.* Frankfurt am Main: Peter Lang
Hunston, Susan
2007 Semantic prosody revisited. *International Journal of Corpus Linguistics* 12 (2): 249–268.
Jucker, Andreas H.
2009 Speech act research between armchair, field and laboratory. The case of compliments. *Journal of Pragmatics* doi:10.1016/j.pragma.2009.02.004
Jucker, Andreas H. and Yael Ziv (eds.)
1998) *Discourse Markers. Descriptions and Theory.* Amsterdam: Benjamins.
Jucker, Andreas H., Gerold Schneider, Irma Taavitsainen and Barb Breustedt
2008 Fishing for compliments. Precision and recall in corpus-linguistic compliment research. In: Andreas H. Jucker and Irma Taavitsainen (eds.), *Speech Acts in the History of English*, 316–341. Amsterdam: Benjamins.
Kilgarriff, Adam
1998) BNC database and word frequency lists. Available at: http://www.kilgarriff.co.uk/bnc-readme.html (last visited October 2009).
Kohnen, Thomas
2000 Corpora and speech acts: The study of performatives. In: Christian Mair and Marianne Hundt (eds.), *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20). Freiburg im Breisgau 1999*, 177–86. Amsterdam: Rodopi.

Lampert, Andrew, Robert Dale and Cécile Paris
2006 Classifying Speech acts using Verbal Response Modes. *Proceedings of 2006 Australian Language Technology Workshop* (ALTW 2006), 34–41.
Leech, Geoffrey and Michael H. Short
1981 *Style in Fiction.* London: Longman
Lenk, Uta
1998 *Marking discourse coherence. Functions of discourse markers in spoken English.* Tübingen: Narr.
Levinson, Stephen C.
1983 *Pragmatics.* Cambridge: Cambridge University Press.
Louw, Bill
1993 Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In: Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), *Text and Technology*, 157–92. Amsterdam: Benjamins.
Louw, Bill
2000 Contextual prosodic theory: Bringing semantic prosodies to life. In: Chris Heffer and Helen Saunston (eds.), *Words in context*. Discourse Analysis Monograph 18 [CD Rom]. Birmingham: University of Birmingham.
Lycan, William G.
2008 *Philosophy of language. A contemporary introduction.* New York: Routledge.
Lyons, John
1977 *Semantics*. Volumes I and II. Cambridge: Cambridge University Press.
Manes, Joan and Nessa Wolfson
1981 The compliment formula. In: Florian Coulmas (ed.), *Conversational Routine. Explorations in Standardized Communication Situations and Prepatterned Speech,* 115–132. The Hague: Mouton.
Maynard, Carson and Sheryl Leicher
2007 Pragmatic annotation of an academic spoken corpus for pedagogical purposes. In: Eileen Fitzpatrick (ed.), *Corpus Linguistics beyond the Word. Corpus Research from Phrase to Discourse,* ■XX?.■ Amsterdam: Rodopi.
McCarthy, Michael J.
1998 *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.
McIntyre, Dan, Carol Bellard-Thomson, John Heywood, Tony McEnery, Elena Semino and Mick Short
2004 Investigating the presentation of speech, writing and thought in spoken British English: A corpus-based approach. *ICAME Journal* 28: 49–76
Mey, Jacob L.
1991 Pragmatic gardens and their magic. *Poetics* 20: 233–245
Morley, John and Alan Partington
2009 A few *Frequently Asked Questions* about semantic – or *evaluative* – prosody. *International Journal of Corpus Linguistics* 14(2): 139–158.
Norrick, Neal
2009 Using large corpora of conversation to investigate narrative. The case of interjections in conversational storytelling performance. *International Journal of Corpus Linguistics* 13(4): 438–464.

O'Halloran, Kieran
2007 Critical discourse analysis and the corpus-informed interpretation of metaphor at register level. *Applied Linguistics* 28(1): 1–24.
Partington, Alan
2004 "Utterly content in each other's company". Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1): 131–156.
Partington, Alan
forthcomingSemantic prosody, evaluation, priming and phrasal irony. In: Karin Aijmer and Christoph Rühlemann, *Corpus Pragmatics. Exploring Speaker Meaning Computerized Corpora.*
Rimé, Bernard and Loris Schiaratura
1991 Gesture and speech. In: Robert S. Feldman and Bernard Rimé. F*undamentals of Nonverbal Behaviour,* 239–281. Cambridge: Cambridge University Press.
Rühlemann, Christoph
2006 Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics* 11(4): 385–409.
Rühlemann, Christoph
2007a *Conversation in Context. A Corpus-driven Approach.* London: Continuum.
Rühlemann, Christoph
2007b Lexical grammar: the GET-passive as a case in point. *ICAME Journal* 31: 111–127.
Russell, Bertrand
1905 On denoting. *Mind* 14: 479–493. Reprinted in *Mind* (2005) 114: 873–887.
Russell, Bertrand
1919 *Introduction to Mathematical Philosophy.* London: Routledge.
Searle, John R.
1969 *Speech acts. An essay in the philosophy of language*. New York: Cambridge University Press.
Searle, John R.
1975 Indirect Speech Acts. In: Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics III*, 59–82. New York: Academic Press.
Sinclair, John McH.
1991 *Corpus, concordance, collocation.* Oxford: Oxford University Press.
Sinclair, John McH.
1999 A way with common words. In: Hilde Hasselgard and Signe Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson*, 157–179. Amsterdam: Rodopi.
Sinclair, John McH
2004 *Trust the text.* London: Routledge.
Sperber, Dan and Deidre Wilson
1995 *Relevance. communication and cognition.* 2nd ed. Oxford: Blackwell.
Stenström, Anna-Brita
1998 From Sentence to Discourse: *Cos (because)* in Teenage Talk. In: Andreas H. Jucker and Yael Ziv (eds.), *Discourse Markers. Descriptions and Theory.* 127–146. Amsterdam: Benjamins.
Stenström, Anna-Brita, Gisle Andersen and Ingrid K. Hasund
2002 *Trends in teenage talk*. Amsterdam: Benjamins.

Stiles, William B.
1992 *Describing talk. A taxonomy of verbal response modes*. Newbury Park/CA: Sage Publications.
Strawson, Peter F.
1950 On referring. *Mind* 59: 320–344.
Stubbs, Michael
1996 *Text and corpus analysis.* Oxford: Blackwell.
Stubbs, Michael
2001 *Words and phrases. Corpus studies of lexical semantics*. Oxford: Blackwell.
Tagliamonte, Sali and Rachel Hudson
1999 *Be like* et al. beyond America: The quotative system in British and Canadian youth. *Journal of Sociolinguistics* 3(2): 147–172.
Whitsitt, Sam
2005 A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10(3): 283–305.