O'Keeffe, A. (2018) "Corpus-based function-to-form approaches". In A. H. Jucker, K. P. Schneider and W. Bublitz (Eds) *Methods in Pragmatics.* Berlin: Mouton de Gruyter, 587 – 618.

22 Corpus-based function-to-form approaches

Anne O'Keeffe

Abstract

This chapter sets out to explore the options for function-to-form research in the context of corpus pragmatics. Corpus-based function-to-form research approaches are used in pragmatics research to explore speech acts and related phenomena, using the function rather than the form as the starting point. Corpus studies more commonly begin with a form and, in pragmatic studies, work towards the functional analysis of these forms (i.e. form-to-function approach). However, when looking at a particular speech act, it can be challenging to find it in a corpus using form-based searches. It is possible to look at a dataset manually so as to code all instances of the speech act in the corpus, however, there is a threshold of corpus size beyond which this becomes implausible. Other systematic options and solutions have emerged such as using Illocutionary Force Indicating Devices (IFIDs) (e.g. *sorry* for apologies), typical features (e.g. positive adjectives, such as *beautiful,* for compliments) and metacommunicative expressions (e.g. using the word *compliment* to retrieve compliments). The paper will also look at some emerging approaches based on using collocational profiles of IFIDs to identify speech acts in very large corpora.

1. Introduction

Within what is termed the "empirical turn" in linguistic research (Taavitsainen and Jucker 2015), corpus linguistics (CL) has spread its application to many sub-fields as well as remaining a robust sub-field in its own right. As Ädel and Reppen note, however, in relation to CL's paradigmatic dominance, that "some subfields are more amenable to corpus-linguistic methodology than others" (2008: 1). Pragmatics is one of the sub-fields to take on this data-driven empirical methodology even though it already had established means of collecting empirical (elicited) data, mainly through Discourse Completion Tasks (DCTs) and role-plays, which are especially widespread in the context of the study of contrastive second language pragmatic competence (Blum-Kulka et al. 1989; Sasaki 1998; Billmyer and Varghese 2000) (for more on DCTs, see chapter 9, this volume). Bringing a CL methodology to pragmatic studies is not without its challenges, as this paper will discuss. The default analytical approach inherent in CL is to move from frequencies of forms to their functions (via an inductive process). In other words, it takes a primarily form-to-function approach to analysing data (see Aijmer, this volume). For those involved in the study of pragmatics, and especially speech acts, and related phenomena, the norm is to work in the opposite direction, starting with a specific pragmatic function and, through means of carefully designed elicitation tasks, to work from the function under investigation to the forms which are

typically used. This is referred to as a function-to-form approach.

Through its inductive process, Aijmer (this volume) notes that taking a form-to-function approach means that the forms can be studied with great precision with regard to frequency, distribution, positions, and collocations with different functions. Rühlemann and Aijmer (2015) point out, however, that the form-to-function approach can be weak at identifying all of the instances of a particular function, as it is form driven. So, on one hand, while CL aligns well with the core principle of pragmatics that meaning is not a stable counterpart of linguistic form, this is also its weakness when using a form-to-function approach. This challenge is referred to by Taavitsainen and Jucker (2015: 12), who say that while pragmatics has embraced the "empirical turn", and other developments in linguistics over the years, "corpus linguistics came into pragmatics later" because, "core features of pragmatics studies, such as negotiation of meanings, speech functions, and variability of language use with momentary shifts in interpersonal relations, are harder to catch with corpus methodology than lexical or morpho-syntactic features" (see also Romero-Trillo 2008; Brinton 2012; Rühlemann and Aijmer 2015). Romero-Trillo (2008: 2) refers to CL and pragmatics as being fields that were "parallel but often mutually exclusive". However, as more CL researchers draw on pragmatics to help analyse their data, and more pragmatics research questions are addressed using corpus data, we are now at the point where we talk about "corpus pragmatics" as an emerging field (see Jucker 2013; Rühlemann and Aijmer 2015).

Within the new coinage of "corpus pragmatics", more consideration is being given to how best to use CL for pragmatics research. Rühlemann and Aijmer (2015) explain that corpus pragmatics combines the key methodologies of both fields. They point out that the traditional vertical reading of corpus data (typically in concordances) needs to be balanced with the more horizontal reading of the contextual details that are required to fully understand pragmatic phenomena (see also Rühlemann and Clancy forthcoming). However, this vertical and horizontal balance presupposes that one begins by searching for a form and that one then works towards the balanced and contextualised analysis of its function(s) (i.e. form-to-function). This paper takes as its starting point the more traditional function-to-form research route of pragmatics analysis and considers this opposite methodological route in the context of corpus pragmatics. We will consider whether CL is fit-for-purpose for this traditional approach within pragmatics research. Essentially, given the importance of continuing the functional investigation of language in use, especially through the study of speech acts, there is a need to consider how, whether, and how best, this work can be done using CL methodologies. Lutzky and Kehoe (2017a) problematize this in relation to the study of speech acts and other pragmatic phenomena in large corpora. They say that, for the most part, speech acts cannot be identified automatically due to the fact that:

1) forms may be produced in a potentially infinite number of ways and,
2) forms which are prototypically associated with a specific speech act (e.g. *sorry*) may also be attested with other functions (e.g. *a sorry state*). (Lutzky and Kehoe 2017a: 38)

As a result, according to Lutzky and Kehoe corpus studies of speech acts, and related phenomena, tend to be conducted using smaller manually annotated corpora and, tend to

"resort to manual forms of analysis, or to adopt eclectic approaches, focusing for instance on specific speech act verbs" (2017a: 38).

In recent studies, this conundrum is being addressed and solutions and workarounds are emerging, as we shall discuss below. To begin with, we shall cast a cautious eye on the use of corpus data in the study of pragmatics. Then, we shall explore possible approaches for function-to-form research within the context of corpus pragmatics for both small and large datasets.

2. Some caveats of corpus data for the study of pragmatics

On one hand, it might seem so obvious that anyone wanting to investigate a pragmatic feature nowadays would first go to a corpus and start by looking at forms and frequencies related to that feature. CL seems to offer so much more in terms of language range and distribution across speakers or writers than data elicited from role-plays or DCTs. Often these corpus data are readily (and often freely) available, in abundance. There are some caveats, however, in terms of the seeming wealth of naturally-occurring language that is available for pragmatics research.

2.1 The challenge of functional diversity and ambiguity

With the abundance of naturally-occurring language data (in electronic form) comes the downside for pragmatics research in the form of functional diversity and ambiguity. A corpus, by its nature, is a sizeable sample of language. A corpus of one-million words of language is considered "small" (O'Keeffe, McCarthy, and Carter 2007: 4). Corpora of fewer than one million words are usually individual enterprises where one researcher has gathered data of a very specific nature to address a particular research question. The boon of corpus quantity brings with it the downside of having a greater remove, as a researcher, from contextual detail and richness which is core to the analysis of pragmatics. Let us consider a brief example: if we opt to look at the speech act of apology, we could immediately look up the direct speech act by searching for a prototypical Illocutionary Force Indicating Device (IFID) for apologising, such as *sorry*, in a corpus. For the purposes of this example, I will use The Limerick Corpus of Irish English (LCIE), a one-million word corpus of spoken Irish English, mostly entailing recordings from casual conversations between family and friends (see Farr, Murphy and O'Keeffe 2004 for a detailed description).

In the sample of one million words, corpus software will instantly find 363 occurrences of *sorry.* In so doing, we have taken one form associated with a speech act and we hope that it generates, or "recalls", instances of the act. Unfortunately, this is only the beginning of the challenge. Because pragmatics takes as its starting point the notion that meaning is not "a stable counterpart of linguistic form. Rather it is dynamically generated in the process of using language" (Verschueren 1999: 10), we cannot, of course, assume a direct correlation between the form and its function as an apology. As example (1) illustrates, the IFID proves unreliable as a means of recalling all, and only, instances of apologies. In this

example, the search word, or IFID, *sorry,* is functioning not as an apology but as a request for clarification used by the listener:

(1) <$1> and <$2> mark speakers one and two, respectively. Two sisters are talking. One sister, <$2>, is telling a story about a derelict house.
<$2> In the window when I was down there.
<$1> **Sorry?**
<$2> There was these kinds of bags of sugar in the window.
<$1> Yeah yeah. (LCIE)

In order to analyse apologies further in this corpus, there is a need to find a workaround. It may mean: 1) manually sifting through all the instances of the form *sorry* to eliminate any that are not related to an apology routine, or 2) "down-sampling", that is, taking a smaller sample of the data and reading this manually to identify all instances of apologies (extended over a number of turns, possibly) and then annotating these so that they can be analysed with the aid of automated tools as well as through qualitative functional analysis. In essence, in taking a pragmatic function, in this example a speech act, as a starting point, it might seem like one has a head start with a large corpus of data (relative to traditional datasets in pragmatics), but because of the lack of a one-to-one relationship between form and function, it is far from straightforward. This challenge, as noted by Lutzky and Kehoe (2017a: 54) has meant that "scholars resorted to smaller data samples (e.g. Koester 2002; Garcia McAllister 2015)" as well as "eclectic analyses of common forms or patterns associated with a speech act (see e.g. Aijmer 1996; Deutschmann 2003; Taavitsainen and Jucker 2007; Adolphs 2008)". Alternatively, others have used metacommunicative expression analysis (see e.g. Jucker et al. 2012; Jucker and Taavitsainen 2014, who use the term "compliment" to retrieve performative instances of compliments) (see Lutzky and Kehoe 2017a: 54). These processes, Lutzky and Kehoe (2017a: 54) point out, generally demand "stages of manual microanalysis to separate unwanted hits from examples with specific pragmatic functions". As we shall detail below, Lutzky and Kehoe (2017a; 2017b), Jucker and Taavitsainen (2014) and Deutschmann (2003), among others, offer plausible solutions for analysing speech acts in large corpora. Firstly however, it is important to consider the longer established approach of eliciting speech act data in the field of pragmatics, using Discourse Completion Tasks (DCTs) and how these compare with corpus data.

2.2 Breadth of form at the expense of contextual depth

DCTs have long been the orthodox method of investigating speech acts (Flöck and Geluykens 2015). They elicit responses to given situational prompts. This methodology, moving from function to form, has been the norm in pragmatics and, as Flöck and Geluykens (2015) note, this longevity is for good reason. Using a DCT means there is no ambiguity of context because the functional scope of the instrument will have been predefined and will therefore control the context and conditions very carefully, including the gender, age, social and interpersonal relationship, and so on, of the speakers. For example, the DCT could be

streamlined to gather apologies in the context of a student apologising to a college professor for being late to class. It could say that you have never met the professor before or that you have met before and that this is not the first time that you have been late with an assignment. This gives a contextual concentration and richness that provides a narrowed range of the forms used in this specific context, with confined conditions. Some would argue that this concentration, or narrowness, of DCT data is its weakness (see Schauer and Adolphs 2006; Flöck and Geluykens 2015) and that it is in stark contrast to using a corpus where one can avail oneself of a much broader range of forms and contexts in a much larger sample of naturally-occurring data. However, despite the abundance of data usually available in a corpus, it is often at the cost of being far removed from the context unless the data has actually been collected by the researcher. In large corpora, there will be detailed metadata on each recording, but this may not be readily accessible and may not be fully completed.

In terms of illustrating the contextual challenges of looking at speech acts in a corpus, let us again take as our example the speech act of apology and look at it using the LCIE. If we use the IFID, *sorry,* as a "way in" and sift through the occurrences so as to identify all instances of *sorry* functioning as the speech act of apologising, we are at the mercy of the corpus design as to how much background data we can access about who made the apology, what their interpersonal relationship with their interlocutor(s) was, what the power semantic was between the speaker and interlocutor(s) (e.g. symmetrical or asymmetrical), what led to the apology (it may or may not be obvious from the data), plus a variety of other possible contextual data. LCIE has metadata on each of the interactions which were recorded so we know certain details such as gender, age, relationship, educational background, place of birth, place where currently living, and so on. However, in the meanderings of casual conversation, as an outside reader of a conversation, one might struggle to contextualise some instances of apologies. Extract 2 is not untypical of what one will find in a corpus of casual conversation. The researcher, as well as finding out the contextual information from the speaker-information metadata database, needs to read a lot of the preceding interactional context to work out that there is a story being told, among friends, amid the interruptions, background noises, overlapping turns, unintelligible syllables, truncated words. In extract 2, with most of the mark-up removed to aid legibility here, it is still challenging to reconstruct the context of the apology, but we can guess that the three friends were chatting. One speaker, <$1>, is trying to tell a funny tale, speaker 3, <$3>, is aiding her narrative with response tokens to show listenership (e.g. *yeah*), and speaker 4, <$4>, is distracted by something (most likely what's on the television in the background) and interrupts the telling of the story by making an aside comment in relation to her observation (*Ah look what yer one's makin aren't they lovely?*). She (speaker <$4>) then apologises for this interruption (*Sorry Joanne finish your story*) and the story continues:

(2) <$N> represents a speaker in order of appearance in the recording, + represents an interrupted utterance, = represents a truncated utterance

<$3> Yeah.

<$1> < clanging sound > <unintelligible word>

<$4> What were ya doin in Tramore?

<$1> Oh we just went down we ended up just ya know+

<$3> Goin out and havin a few drinks.
<$1> +yeah wha= ?
<$3> I said Joanne and Philip do funny things.
<$1> We do weird things.
<$4> **Ah look what yer one's makin aren't they lovely**?[1]
<$3> Wha= ?
<$4> **Sorry Joanne finish your story**.
<$1> So < laughing > so < two syllables unintelligible > went in and said can I have a breast of chicken without the bra < laughter >. (LCIE)

This example illustrates the contextual lacuna that a researcher can experience when working with corpus data while on the other hand, it clearly shows a richness. The main advantage of using a corpus is the immense breadth it can offer in terms of the range of forms that are used across so many contexts, individual language users in their different roles, their varying statuses, educational and social backgrounds, ages, genders, and so on. However, though you have ready access to the form (which you elect to search for), you may not have access to the contextual variables from whence it came.

Clearly there is a trade-off between the breadth of forms that corpus data can offer a researcher and the details of context and conditions within which these forms occurred. In contrast, by using a DCT, one can carefully define the context and its conditions, but this is at the expense of breadth of form, as Figure 1 illustrates:

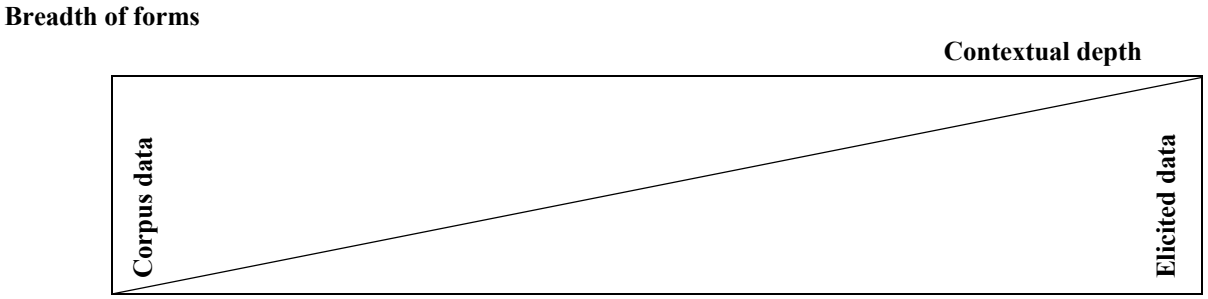

Figure 1: Form versus context in corpus versus elicited data

We will return briefly to this point below when we look at two studies that have directly compared DCT and corpus data (Schauer and Adolphs 2006 and Flöck and Geluykens 2015).

The temptation to move away from even attempting to find solutions for using function-to-form approaches is strong given the allure of big data. Let us now consider some caveats about big data options in the study of speech acts and related phenomena.

2.3 Big data caveats

Taavitsainen and Jucker (2015: 18) issue an important warning, amid the Big Data trend, "[t]his unprecedented increase in data size accentuates the problem of the right balance

[1] *Yer one* is an Irish English slang form of *your one*, meaning 'that woman', which is functioning here as a personal deictic reference.

between the amount of data and the contextualization of the data. Often the researcher has to opt for one and sacrifice the other". With such data stores at one's finger tips, it is easy to see how form-focused research, driven by the weight of data sample size, could become the preferred route for researchers interested in investigating some aspects of pragmatics. Given the importance of understanding the contextual provenance of a form in the study of pragmatics, it is crucial that the limitations of big data results be understood. Tantalisingly large databases can give immediate results across centuries of data though without the metadata that one would associate with a corpus. The best-known example, at the time of writing, is the *Google Books Ngram Viewer*, which gives instant access to the frequency of ngrams of up to five words in a corpus of over 5 million books (500 billion words), published over the last 500 years, or so (currently from 1500 – 2008).

As Taavitsainen and Jucker (2015) note, for historical pragmaticists, it offers a fascinating exploratory tool. For example, we can instantly look up the frequencies of *I apologise* and *I apologize,* between 1800 and 2000, in both American and British English books. We can see that *I apologize* has a frequency of 0.7 PMW in American and 0.25 PMW in British English (Figures 2 and 3):

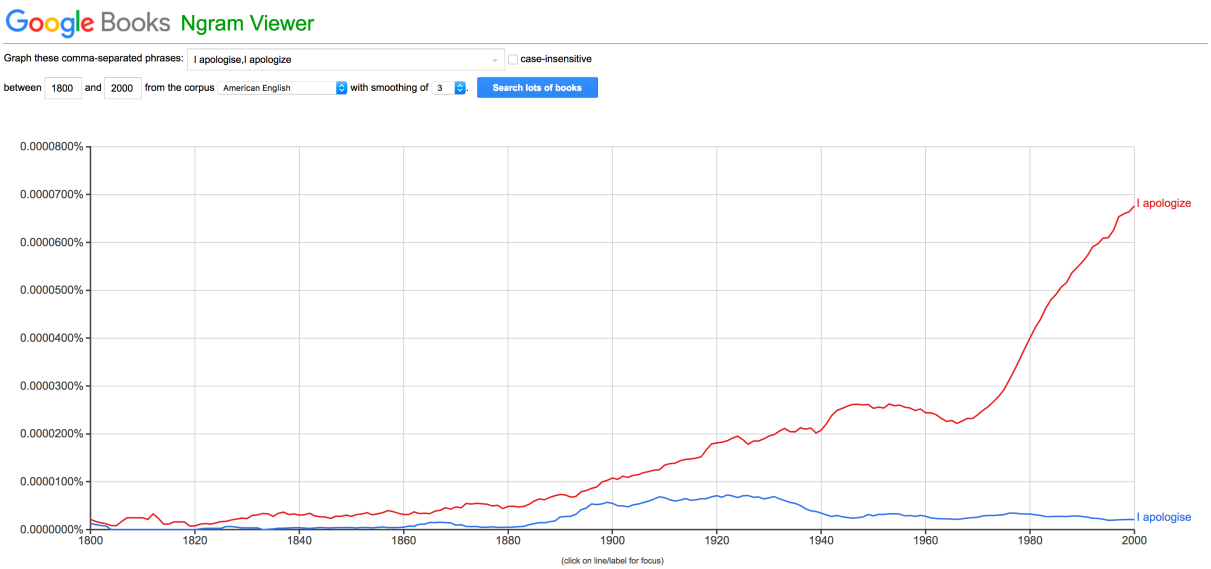

Figure 2: *I apologise* and *I apologize* in American English books 1800 – 2000 *Google Books Ngram Viewer*[2]

[2] Source code: <iframe name="ngram_chart" src="https://books.google.com/ngrams/interactive_chart?content=I+apologise%2CI+apologize&year_start=1800&year_end=2000&corpus=17&smoothing=3&share=&direct_url=t1%3B%2CI%20apologise%3B%2Cc0%3B.t1%3B%2CI%20apologize%3B%2Cc0" width=900 height=500 marginwidth=0 marginheight=0 hspace=0 vspace=0 frameborder=0 scrolling=no></iframe>

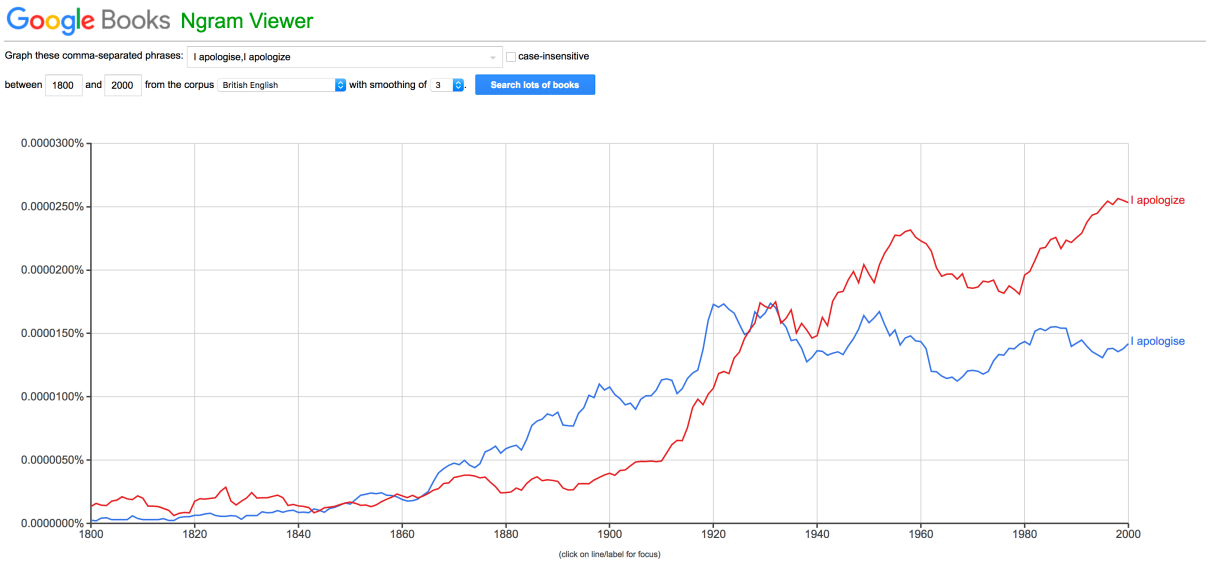

Figure 3: *I apologise* and *I apologize* in British English books 1800 – 2000 *Google Books Ngram Viewer*[3]

However, this is a database, and we must be mindful of the major limitation that it has: we are without any context for the occurrences of these forms, and so while it is interesting as an exploratory tool, it is clearly contextually devoid. It is best treated as an interesting starting point, a "ready reckoner" of forms over time but it is of little or no value to the investigation of how these forms actually function(ed).

Another corpus that is of use for diachronic analyses is the *Corpus of Historical American English* (COHA), developed by Mark Davies, Brigham Young University. Like the *Google Ngram Viewer,* it was launched in 2010 and covers data from 1500 to 2008 As Fringinal et al. (2014) note, COHA is a "smaller" mega-corpus, standing at 400 million words and its creator argues that the substantial difference in size does not affect reliability of results when these corpora are compared. The COHA comprises data from the registers of fiction, non-fiction, magazine and newspaper. It is accessible, free of charge, via the Corpus of Contemporary American English interface. As Taavitsainen and Jucker (2015: 18) point out, "This allows for detailed and fascinating information on the frequency of even extremely rare ngrams". However, in respect of pragmatics research, it comes with similar caveats in terms of the degree of contextual information the researcher has available to them.

We will now examine two studies that focus in detail on the impact of how data was collected on the output and findings in relation to function and form in the study of speech acts.

3. Evidence from studies comparing speech act data from DCTs and corpora sources

[3] Source code: <iframe name="ngram_chart" src="https://books.google.com/ngrams/interactive_chart?content=I+apologise%2CI+apologize&year_start=1800&year_end=2000&corpus=18&smoothing=3&share=&direct_url=t1%3B%2CI%20apologise%3B%2Cc0%3B.t1%3B%2CI%20apologize%3B%2Cc0" width=900 height=500 marginwidth=0 marginheight=0 hspace=0 vspace=0 frameborder=0 scrolling=no></iframe>

Schauer and Adolphs (2006) and Flöck and Geluykens (2015) are two studies which compare the benefits and challenges of using corpus data versus DCTs in the investigation of pragmatic function. These studies help us better understand the complexities of the issue.

Schauer and Adolphs (2006) investigated expressions of gratitude using a DCT of eight scenarios with 16 native speakers. They then used the forms that emerged in the DCTs as a basis for corpus searches. They envisaged the corpus data as being able to provide detailed insights into expressions of gratitude employed by "a wide part of the population in casual conversations between friends and family, while the DCT scenarios were designed to represent situations that a specific group (in this case university students) were likely to come across during a sojourn in the target environment" (123–124). They used the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a five-million-word database that was collected between 1994 and 2001 (see McCarthy 1998 for a description of this corpus). In all, nine forms emerged from the DCT for the expression of gratitude: *Thanks, Cheers, Ta, Thank you, Thanks a lot, Thanks very much, Thank you so much, Nice one,* and *Cheers sweetie* (Schauer and Adolphs 2006: 125). All but one of these forms, *Cheers sweetie,* were found in the corpus data though to differing degrees (in terms of frequencies). The three most frequent forms that appeared in the DCT were *Thanks, Cheers* and *Thank you* and these forms were also the most frequent, though in reverse order, in the corpus data where *Thank you* was by far the most frequently used form.

Schauer and Adolphs (2006) cite the length of the turn in which gratitude is being expressed as the main difference between the elicited and the corpus data. Importantly, they note that because the DCT is so focused and controlled within predetermined conditions, as discussed, it usually generates single utterances rather than stretches of interaction. The corpus data gives a broader contextual picture of the stretch of discourse that involves the act of expressing gratitude rather than a single utterance of gratitude. For example, as mentioned above, *thank you* is one of the top three DCT forms identified and the most frequent form in the corpus when compared to the other DCT-generated forms and yet, the corpus also tells us that its use can stretch over a number of turns in what Schauer and Adolphs (2006) call a "gratitude cluster".

Extract (3) from the BNC shows an example of a gratitude cluster in a conversation between a parent and a child. While it is not clear what the thanking relates to, it is interesting to observe how the thanking spreads across many speaker turns (Davies 2004):

(3) <$1> …**Thanks** very much.
<$2> **Cheers** Dad. Put away your er luggage. <$E> pause </$E>
<$1> <$E> unclear </$E>.
<$2> Er **cheers**. <$E> whispering </$E> … <$E> unclear </$E>. <$E> laugh </$E>
<$E> pause </$E>
<$1> **Cheers**. Right. Okay.
<$2> **Thanks very much**. (BNC_FM4)

Schauer and Adolphs (2006) note that because DCT data is normally based around single utterances, this can distort the overall reality of speech acts which are typically

negotiated and developed over a number of turns in a dynamic discourse event. An important point to bring in here is that Taavitsainen and Jucker (2015: 17) forecast that, in the future, speech act analyses will more consistently focus on the interaction between participants and how speech act values are jointly negotiated and established in the interaction moving from a one-dimensional focus on single utterances and their meaning to negotiated meaning within the dynamics of real-time interaction.

Schauer and Adolphs' (2006) finding that the forms generated by the DCT methodology were less complex in nature in comparison to corpus data is borne out in Bodman and Eisenstein (1988) and Yuan (2001). Additionally, studies such as Hardford and Bardovi-Harlig (1992) on rejections comparing DCTs and authentic discourse from advertising and Beebe and Cummings' (1996) study on refusals found that DCTs contained fewer semantic formulae and negotiating strategies and were overall less complex and more direct. However, Flöck and Geluykens (2015), in their study of directives, found DCT data to be more indirect and to contain more downgrading modifiers than real spoken data to which they were compared. However, Flöck and Geluykens (2015) reviewed the findings from eight comparative studies (including those cited above) and concluded that the findings were far from convergent.

Flöck and Geluykens (2015) investigate directive speech acts in three datasets:

- A sample of spoken data taken from the British component of the International Corpus of English (ICE): they manually retrieved instances of directive speech acts in the spoken component of the ICE-GB, which consists of 100 transcripts (each 2,000 words) of face-to-face and telephone conversations, between participants of mostly "low social distance".
- Elicited written data collected using DCTs: these were elicited in scenarios where fictional characters had low social status and low power relations. Flöck and Geluykens (2015) suggest that these elicited data and the corpus data are maximally comparable because they have a close genre and micro-social match-up.
- A small corpus of business letters: these are part of the Antwerp Corpus of Institutional Discourse (Geluykens and Van Rillaer 1995) and due to confidentiality constraints, there is no demographic information available.

All of Flöck and Geluykens' (2015) data are from native British English speakers and were collected within the same time span. They randomly selected 235 directive speech acts from each data set and these were then categorised according to a uniform coding system (encoding a pragmatic profile of the act). They conclude that the DCT data exhibited significant differences compared with the spontaneous data. They note the greatest degree of difference from the conversational directive speech acts in almost all aspects of their investigation (e.g. percentage of direct head acts, conventionally indirect head acts, indirect head acts, downgraded head acts, ratio of downgraders per head act, percentage of mood imperatives with *please*, number of downgrading modifiers and total number of upgrading modifiers). Interestingly and importantly, they note that spontaneous non-elicited data is far

from homogeneous. They found strong evidence of the influence of the conditions of use and genre (though further investigation was beyond the scope of their study). This led them to stress that "we should at least allow for the possibility that the type of illocution influences the production choices language users make" (Flöck and Geluykens 2015: 34). They go on to note, however, in relation to speech act variation that other speech acts, such as thanking, might be more routinized and stereotypical. They say that, "what seems clear is that corpus pragmatics in the widest sense of the word has a major role to play in unravelling some of these complex issues" (Flöck and Geluykens 2015: 34).

It is of great importance to the evolution of corpus pragmatics that we see a continued research of this nature where the output from different methodologies for data collection are closely scrutinised so as to arrive at enhanced understandings of the value and limitations of methodologies within the area of pragmatics. Leaving aside how data is collected at this point we turn now to the practicalities of how best to analyse data in a function-to-form approach.

4. Function-to-form approaches to corpus research

Ädel and Reppen (2008: 2-3), in the introduction to their edited volume, summarised four approaches to using a corpus for corpus-based form-to-function investigations of discourse (listed below). Ädel and Reppen (2008) point out that these approaches often overlap and there is iteration within any of these strategies. Nonetheless, they are useful to consider as core investigative strategies and more pertinent to the present study, we need to consider, what are the equivalent strategies or approaches that one might take if one is interested in the opposite investigative route, namely function-to-form. Ädel and Reppen's (2008: 2-3) four approaches to form-to-function corpus analysis:

- One-to-one searching: where there is a 100% match (or recall) from the search item to relevant hits; for example, you seek to investigate the use of noun phrases in a sample of data. In a Part-Of Speech (POS) tagged corpus, this will generate full recall of all noun phrases. If you wished to look at all instances of *Thank you*, again this search would result in a full recall of forms.
- Sampling: this involves using one or more search item(s) that are good examples of the linguistic phenomenon in question. In pragmatic terms, this means using IFIDs, for example. As discussed earlier, one could search for *sorry* so as to sample possible instances of apologies.
- Sifting: if you engage in sampling, you will most likely need to sift through the sample to isolate the forms/instances that you are interested in. For example, through sifting you can eliminate any instances of *sorry* that are not functioning as apologies. However, this process is limited in that you will not find instances of apologies that do not use *sorry*.
- Frequency-based listing: this is the purest corpus approach where you take a bottom up approach and start by looking at the frequencies of forms in your corpus and work from there in terms of their patterns and meanings. Many frequency-based studies of corpus data end up with pragmatic conclusions to

explain differences in frequencies and patterns across contexts of use but they set out from the baseline of frequency results of forms.

Here we will attempt to lay out the possibilities for function-to-form corpus analysis. As with the aforementioned strategies for form-to-function research, they will often overlap.

4.1 Approach 1: One-to-one searching in a pragmatically annotated corpus

In the case of function-to-form analyses, being able to conduct a one-to-one search of a pragmatic function, in a pragmatically-annotated corpus, so as to recall all of its instances of a given speech act is what O'Keeffe, Clancy and Adolphs (2011) referred as the "holy grail" for corpus pragmatic research. Now, corpus tools and annotation systems are emerging which show that this is, and will increasingly be, possible (cf. Culpeper and Archer, this volume). It ultimately means that a pragmatic function, for example a speech act such as offers, apologies and so on, could be recalled automatically because they have been annotated within the corpus and are thus retrievable, in one-to-one searches, using the appropriate tools.

As Rühlemann and Aijmer (2015) summarise, the growing body of pragmatically annotated corpora include:

- speech acts (Stiles 1992, Garcia 2007, Kallen and Kirk 2012, Kirk 2016)
- discourse markers (Kallen and Kirk 2012, Kirk 2016)
- quotatives (Kallen and Kirk 2012, Kirk 2016, Rühlemann and O'Donnell 2012)
- participation role (Rühlemann and O'Donnell 2012)
- politeness (Danescu-Niculescu-Mizil et al. 2013)

Rühlemann and Aijmer (2015) speculate that the reason why pragmatic annotation is not yet widely used is that the form-function mismatch of most pragmatic phenomena means that automative assignment of tags will often lack precision and manual laborious annotation is unavoidable. The work of Weisser (2015) offers some hope in the form of semi-automating the process of speech act identification using the Dialogue Annotation and Research Tool (DART). This tool, through carefully determined multiple syntactic structure features and mode (e.g. modals, adverbials, conditionals, etc.) as well as complex computational tagging, can identify speech acts in task-oriented dialogues from the Trains and Trainline corpora (see Weisser 2015). Weisser shows that the tool is able to generate a high number of accurately labelled speech acts, within this very defined context. These categories yielded a speech act taxonomy that included: conventionalized, dialogue-managing, information- or option-seeking, information-providing/responding, directive-seeking/providing, (dis)agreeing/acknowledging, informing, and commitment-indicating. For this tool to be further developed, Weisser stresses the need for corpora to have available more information, at transcription phase (e.g. syntactic structure, roles of the interlocutors, and prosodic description). This is borne out by the work of Kallen and Kirk (2012) on the ICE Ireland corpus, which we will look at in greater detail.

Kirk and Andersen (2016: 294-295) outline some of the challenges of pragmatic annotation, not least of all the fact that when real spoken language is transcribed, it is reduced

into a pragmatically-bereft form (as alluded to above). Kirk (2016: 300) notes that transcriptions record "the locutionary act of producing forms and constructions, but 'what is heard' (i.e. the illocutionary force or intent, and its processing as the perlocutionary effect) is only extrapolable from the transcription". These deficiencies make it even more challenging to superimpose pragmatic annotation onto existing corpora of spoken language.

What is not encoded in conventional lexico-syntactic transcriptions are indications of the pragmatics operating in an utterance: the illocutionary force or intent (the speech act status), the perlocutionary effect, the upholding or breaching of the Gricean co-operative principle, the politeness strategy invoked, the attitude of a speaker to the message of the utterance being made (pragmatic stance) or to the hearer of that utterance (face negotiation), and so to its potential impact. Much of what speakers utter is determined by a speaker's attitude towards what they are saying and towards the person(s) to whom they are saying it. (Kirk and Andersen 2016: 294-295)

Crucially, they note that understanding these deficiencies is a key to the ongoing development of pragmatic annotation: "The more linguists come to understand about those interpersonal, intersubjective, communicative ways, the more new layers may be added to the linguistic structures which have been conventionally represented hitherto" (Kirk and Andersen 2016: 295).

Of interest is the SPICE Ireland corpus because it is an example of a spoken corpus which has been pragmatically annotated and so it offers a model for how one-to-one searching can be made possible in a function-to-form approach to corpus pragmatics. SPICE Ireland is part of the International Corpus of English suite (Kirk et al. 2011; Kallen and Kirk 2012). It contains just over one million words, entailing 15 discourse situations, as well as 17 written domains. The 15 discourse situations comprise 626,597 words and all were annotated pragmatically. The annotation scheme comprises five components: the speech act status of each utterance in the corpus, based on Searle's (1976) categories of illocutionary acts, tone movements, discourse markers, utterance tags, and quotatives (see Kirk 2016: 306). Speech act status, for instance, is marked with pairs of angled brackets (based on the system used in COCOA conventions for pairs of opening and closing angle brackets for the representation of a speech act, see below). The annotation surrounds the span of an utterance which contains a speech act, i.e. with a code in angle brackets before the utterance, concluding with a backslash. An appropriate code is used to represent the type of act based on Searle's (1976) taxonomy (Kirk 2016: 302):

<rep> ... </rep> for "representatives";
<dir> ... </dir> for "directives";
<com> ... </com> for "commissives";
<exp> ... </exp> for "expressives";
<decl> ... </decl> for "declaratives"

Four other codes that were deemed necessary (Kirk 2016: 302):

<icu> ... </icu> for "indeterminate conversationally-relevant utterances"

These are used to mark a broad range of minimal responses, back-channel utterances, or "other elements of speech which are relevant to the maintenance of discourse coherence or continuity, but which lack a discernible function as a speech act" (Kirk 2016: 309).

<soc> ... </soc> for "social expressions"

This code is used for social expressions such as greetings, leave takings, and other interactive expressions fall into this category (for example the closing exchange in telephone conversation).

<xpa> ... </xpa> for utterances not analysable at a pragmatic level

Kirk (2016: 310) notes that the SPICE annotation tool requires every utterance to be glossed for pragmatic value, "yet it is inevitable in a large corpus of naturally-occurring data that many utterances will be impossible to categorise as speech acts or conversational moves of one kind or another". In such cases, this code is used to show that an utterance lies outside the pragmatic frame of analysis.

<K...> ... </K...> for "keyed" utterances.

Kirk (2016) notes that the data of ICE-Ireland provide clear examples where speakers are not being literal, but rather use the form of one type of speech act to commit an act of a different type. Kirk followed the work of Goffman (1974) on frame analysis, and devised a <K> code for such utterances, where they are treated as "keyings" of a primary speech act. He provides the following example which, Kirk (2016: 310) notes, "takes the syntactic form of a commissive (undertaking to send the listener a bill), but it is not intended as one" rather is it uttered by the judge who has just given off-the-record advice to a barrister. Kirk (2016: 310) provides the interpretation "that it has the function of a directive — an utterance made in order to provoke laughter. The humour itself derives from the speaker's intentionally anomalous use of the syntactic form of a commissive when it is understood that the commissive is not in this case genuine":

<ICE-NI-LEC-P2A-061$B> <#> **<dirK>** Yeah* <,> I 'll 1sEnd you my 2bIll% **</dirK>** <&> laughter </&> (Kirk 2016: 310)

The scale of pragmatically annotating such a substantial sample as SPICE Ireland (in terms of spoken language) seems challenging, to say the least, but there are also examples of work where researchers who are using with much smaller and more contextualised datasets have been able to engage with a similar level of pragmatic annotation for their particular purposes. A case in point is the work of Milà-Garcia (forthcoming). In this work, agreement and disagreement in spoken Catalan are the focus and the data has been annotated for this purpose. This allows the research a total recall on all stretches of discourse (which have been

coded) involving either an agreement or a disagreement. Garcia McAllister (2015) offers more interesting samples of studies where speech acts have been investigated using corpora where various workarounds have been found, especially using smaller samples which we will now consider in greater detail.

In sum, pragmatic annotation offers a possible solution for function-to-form research but it comes with limitations: 1) it is enormously time-consuming and labour intensive (and thus expensive) and, realistically, this will be a major barrier to its mainstreaming; 2) due to the inherently fuzzy and discursive nature of speech acts, decisions of interpretation are dependent on the annotator's interpretation within the bounds of his/her understanding of the contextual conditions of the speech event; and 3) because of these constraints, it is best applied in small scale studies, where the researcher is conducting the annotation and has an indepth understanding of the contextual variables and conditions.

4.2 Approach 2: Sampling, searching and sifting

As Rühlemann and Aijmer (2015) point out, pragmatics researchers are used to dealing with small amounts of text and analysing these "horizontally" (taking in all contextual factors) but, they note, "even small specialized corpora contain far more words than could possibly be read and analysed by any one researcher in the same way as the select texts which pragmaticists are used to working with" (Rühlemann and Aijmer 2015: 6).

An approach that can make function-to-form research more manageable is to randomly sample from a corpus so as to analyse pragmatic function within that smaller dataset. When the dataset has been made more manageable, the researcher can then read it qualitatively and sift through it to find all instances of a particular pragmatic phenomenon. Garcia McAllister (2015) details a study where she down-sampled a percentage of different data types from a larger corpus so as to investigate the speech act category of directives. Garcia McAllister drew down data from the spoken component (1.6 million words) of the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) (2.7 million words in total), which was collected via audiotaped recordings of conversations, business interactions, and lectures that took place in a university setting (see Biber et al. 2002, 2004). She narrowed her dataset to the following contexts of use (percentages of her down-sample are in brackets): service encounters (39.3%); office hours (32.3%), study groups (28.3%). She then had a data sample of 42,797 words, which she manually sifted (and listened to the audio recordings) to identify, code and annotate all instances of directive speech acts. Through her coding system (see Garcia McAllister 2015), she was able to then apply corpus tools to assign further linguistic and contextual information to each utterance that she had identified as relevant to her study and to provide a data set listing each utterance and its corresponding descriptors. Among other findings, she identified the role of each situational context in predicting the type of speech act used, for example, service encounters were found to be characterised by a high frequency of requests for information, services and payment, suggestions and putting interlocutors on hold. Reflecting on the process and methodology, she notes that the most difficult part was identifying speech acts in corpora and annotating them: "It took many hours of listening to audiotapes and reading transcripts to code all of the utterances analyzed in this study" (Garcia McAllister 2015: 45).

McCarthy and O'Keeffe (2003) offer another example of a study where researchers used a down-sample from a larger corpus and then sifted manually through the sample data to identify and pragmatically categorise the item which was the focus of their study. This paper sought to explore the pragmatic functions of vocatives in conversation. One of the datasets under scrutiny, a small 55,000 word corpus of radio phone-ins, was manageable enough in size to manually sift through to find and functionally classify all vocative occurrences. The other dataset was the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a five-million word corpus of spoken English (see McCarthy 1998). It is obviously implausible to look at all instances of vocatives in such an amount of data. The solution was to generate a word frequency list to find the most commonly used vocatives in the corpus. Kinship terms, *Mum(my)* and *Dad(dy)*, were also included since a good deal of the casual data was family-based. The next step was to run concordances of the high frequency names/address forms. A cut-off of a maximum of five uses of any one name/address form as vocative was set as a restriction on the corpus search (McCarthy and O'Keeffe 2003). Through this process of sifting, a total of 100 extracts involving vocatives were identified for further analysis. Among other findings, they noted a high degree of use of vocatives in the context of hedging. The vocative was neither syntactically nor semantically necessary, but it served often to build relationships ("relational"), downtone challenges, adversative comments or in disagreements. Additionally, vocatives were also often found to be a core feature of badinage (McCarthy and O'Keeffe 2003), especially in the casual conversational data. The functional results from the CANCODE data down-sample could then be compared with the radio phone-in results once the latter were normalised to percentage results (i.e. so that results were both out of 100):

Table 1 Breakdown of functional types of vocatives across a random sample of 100 from CANCODE and a percentage ratio of all vocatives in *Liveline* radio phone-in (McCarthy and O'Keeffe 2003)

Function	**CANCODE (out of a 100-vocative sample)**	**Radio Phone-in (%)**
Relational	30	7.7
Topic Management	21	9.0
Badinage	19	3.0
Mitigator	15	10.3
Turn Management	11	11.2
Summons	4	0.0
Call Management	0	58.6

The following examples, (4) to (6), of where the radio phone-in caller uses the presenter's name (Marian) illustrate the use of vocatives where they are superfluous to the transactional context of the radio phone-in but they aid the pragmatic smooth running of agreements and evaluations:

(4) <$2> Yes indeed **Marian** ah I'd I'd have to agree wholeheartedly with him. (LCIE)

(5) <$2> That's right **Marian**. (LCIE)

(6) <$2> It is indeed **Marian** because ah you know again I think that people are … (LCIE)

In the institutional data (radio phone-in), vocatives had an important call management function, which included changes in footing (Goffman 1979) from the audience to the caller. Example (7) illustrates this function. When the caller's name is used (*Austin*), this is the point at which the presenter (*Marian*) changes her footing (speaker alignment) from the audience to the caller:

(7) <$1> Now to a couple that had very very difficult Christmas this year however all's well that ends well ah **Austin** good afternoon to you.
<$2> Good afternoon **Marian**.
<$1> Your little boy went back to playschool yesterday?
<$2> Yesterday that's right. (LCIE)

In a follow-up study, Clancy and O'Keeffe (2015) used the results of the functions of vocatives identified in the 55,000 million word radio phone-in and compared these with an even smaller dataset of 12,500 words of conversations between friends and family (see Clancy 2015). This dataset was small enough to allow for the sifting and sorting of all 161 instances of vocatives in the data. Once these were classified according to their function, it allowed for the comparison of vocative use in the institutional context of radio phone-in (where pseudo-intimacy was replicated, see O'Keeffe 2006) and the intimate discourse of family and friends. Again, the results from the family and friends data, in line with McCarthy and O'Keeffe (2003), showed that vocative use was much more common in casual conversation between family and friends and that it played a key downtoning role in the context of mitigation, among other relational functions.

These case studies, along with many others, show the benefit of careful and principled sampling from existing corpora, especially where you can access metadata about the speakers and the situation. This then makes scalable the manual sifting through these data so as to pragmatically categorise all instances of your research focus.

4.3 Approach 3: Using existing research findings as "seeds"

Another important means of looking at pragmatic functions is to use existing research findings as the "seeds" or starting points. It is important to stress that there is so much research output already in existence on so many aspects of pragmatics, not least of all speech acts, from the many years of work that has preceded corpus pragmatics. These studies provide very useful starting points for search items in corpora. The first of these studies, Schauer and Adolphs (2006), has already been discussed above, in terms of its comparative findings. Here, we will focus on its methodology. Schauer and Adolphs (2006) take the DCT output from eight scenarios involving 16 native speakers as their starting point for corpus searches of the speech act of expressing of gratitude. In other words, they used the forms that

emerged in the DCTs as a basis for their corpus searches. They say that they opted to start with the DCTs as their source of corpus search items because they wanted to control the variables of the context for the scenarios. In doing so, they were able to make their output from the corpus comparable with that of the DCTs and this, as we have discussed, led them to some important methodological insights. The important methodological point here that we can draw from this study is that DCT results for a given speech act, routine or situation, can offer seeds for searching corpus data and in so doing one generates a comparable dataset and one will gain insights into how these forms are used across turns.

Another example of this use of existing research as a seed is a study by Cheng and O'Keeffe (2015) where they sought to investigate vague language approximator forms (e.g. *about seven, seven or so, at least seven*) within one corpus (inter-culturally) and also to compare the forms across with another variety of English (cross-culturally):

- Inter-cultural comparison of two sub-corpora of the Hong Kong Corpus of Spoken English (HKCSE) (a total of 216,942 words): a Native Speaker sub-corpus of 108,760 words and a Hong Kong Chinese sub-corpus of 108,182 words;
- Cross-cultural comparison of the results from the inter-cultural comparison of Hong Kong data with results from Irish English, using the one-million word Limerick Corpus of Irish English (LCIE).

Cheng and O'Keeffe were keen to investigate the degree to which these forms and their pragmatic functions were universal within and across two varieties of English. This task was too enormous to undertake for all vague language (VL) items which are not tagged in either corpus so they narrowed their focus to one type of vague language which was already described in previous research, namely Channell's (1994) approximator + number (n). They used the search items from Channell's research: *about, around, round, approximately, or, or so, at least, at most, less, more, under* and *over*. These searches had to be disambiguated through manual concordance sorting so as to arrive at only the relevant structures that contain the search items and "n" and/or "m" (where "n" refers to a number and "m" refers to a multiplier of the number, e.g. five (n) or ten (m) minutes). Following Channell's (1994) model, the HKCSE sub-corpora and LCIE were examined in detail. In summary, they found that on the surface, approximator + number (n) seemed to be a universal feature in terms of form and distribution, with no significant quantitative differences emerging either from the inter- or cross-cultural analysis. However, they note that when they looked qualitatively at what the approximators were referring to in their context of use, they found variation in terms of their distribution (e.g. approximation with time and calendar periods was the most common context). For this phase of the analysis, the researchers used a random sample of 100 items from each of the three datasets (in the manner detailed in approach 2). This close qualitative analysis also led to insights about cultural implicitness (especially within family interaction).

Reflecting on their methodology of using an existing model of forms based on existing research, Cheng and O'Keeffe say that it allowed them to work within its syntactic parameters to search through corpora for instances of one specific form of VL. They say that

"while it did involve a lot of manual sorting through concordance lines to eliminate non-VL instances, it was not by any means an insurmountable task" (2015: 374).

4.4 Approach 4: Solutions for larger corpora

Solutions proposed above are limited to smaller scale corpora or to small samples drawn down from larger datasets. Let us now showcase some studies that have used strategies to identify speech acts in larger datasets.

4.4.1 Using llocutionary Force Indicating Devices (IFIDs)

The seminal work of Deutschmann (2003) set out to examine apologies in British English using the 10 million word spoken component of the BNC. As he details, these spoken data involve a total of 4,705 speakers. From this, he isolated only those dialogues produced by speakers whose age and gender were available in the metadata. This sub-corpus comprised 5,139,082 words produced by over 1,700 speakers.

As Deutschmann (2003: 17-18) explains, the investigation was limited to explicit apologies which appeared in the form of IFIDs. Thus, his study focused on "expressions containing variants of the words *afraid, apologise, apology, excuse, forgive, pardon, regret* and *sorry*". Using the BNCweb Query System, the results were then downloaded to an Excel database for manual analysis. By sifting through the results, utterances which functioned as explicit expressions of apologies were identified. Once the data was "cleaned" of all non-apologies, each instance was analysed in the context of the conversation where it was originally uttered so as to classify it functionally and pragmatically. Where available, speaker metadata, such as gender, age, social class and the person being addressed, were also noted in the database for each apology. In addition, where possible, other contextual variables were also logged, such as the conversational setting (formality level), conversation type and the number of participants in the given interaction. Details were also entered for each apology on the power relationship and social distance of the interlocutors. With this level of meta-detail, Deutschmann was able to generate some very detailed results on how, when and by whom apologies were performed.

Deutschmann's analysis identifies three overall functional types of apologies: real (prototypical) apologies, formulaic apologies and face attack apologies. What is significant also in this study is that the author sheds light on the link between his corpus-based approach and its results in comparison to other studies which used different approaches. In other words, he explores the correlation between research design and scope of results. The scale of Deutschmann's study allows for robust correlations between apologies and variables, such as gender, age, social class, formality level, group size and genre. For instance, he was able to show that:

- Younger speakers apologised far more often than older speakers; and
- Speakers from middle class backgrounds apologised more than working class counterparts.

As Woodman (2005: 316) notes, one of Deutschmann's most novel findings was the correlation between group size and apologies:

> the more participants in an interaction, the higher the rate of apology. This meant therefore that genres such as meetings, classroom contexts, job interviews had more frequent rates of apologies than genres associated with smaller sizes, for example medical consultations and historical interviews (see Deutschmann 2003: 161).

Deutschmann's findings in relation to power relations and social distance showed that the more powerful the speaker, the higher their rate of apology and conversely, the lower the power of the speaker, the lower the rate of apology.

Woodman (2005: 316) in reviewing Deutschmann's (2003) methodology summarises the advantages of this approach by saying, "[t]he obvious advantages of using a computerized database such as the BNC are the sheer scale of the data and the fact that the language occurred naturally". Woodman (2005) continues, "[t]he disadvantages lie in the lack of crucial information in connection with the delivery of the apologies (such as body language and prosodic features), in the inevitable inaccuracies involved in the transcription process, and in the lack of any psychological contextual information about the participants (e.g. perceived gravity of offense, degree of affection between participants)". The important achievement of the painstaking work of Deutschmann is that it showed the scale of what can be done in using a large corpus for the analysis of a speech act in a systematic manner. Additionally, other studies can now "stand on the shoulders" of the work of Deutschmann because he has offered such a detailed starting point for anyone interested in looking at apologies in corpus data.

Lutzky and Kehoe (2017a) and (2017b) are examples of two studies which have built on the work of Deutschmann (2003). They both explore apologies in the diachronically-structured *Birmingham Blog Corpus* (BBC), which spans 2000-2010 and is 630 million words in total. In Lutzky and Kehoe (2017a), for example, they begin with Deutschmann's eight core apology IFIDs (see above) and their goal is to arrive at a collocational profile of these items so that they can be used for automatic attestation of apologies within their very large corpus.

Lutzky and Kehoe (2017a) begin with a sub-corpus of 95 million words of blogs, plus 86 million words of readers' comments. Using the apology IFIDs and their lemmas (e.g. *pardon/pardons/pardoned/pardoning*), they generate all occurrences in the data, without distinguishing between apology and non-apology at this stage. This meant that their initial findings included many non-apology items, for example, all instances of *afraid*, not just *I'm afraid* in the context of an apology.

Their next step was to conduct a detailed word frequency profile of the collocates of each of their initial search items. The collocate had to occur within the top 100 most frequent times and it had to be within a span of four words to the left or right of the IFID. For example, the collocates of the IFID *apologise* included items that occurred next to the search word, such as *profusely,* as well as collocates that were up to four words to the left or right of it, for example, *inconvenience* or *advance*. They used a z-score to rank significance of

collocational pairings relative to collocate frequency and corpus size. For instance, they give the example of *profusely:* this has the highest z-score though its raw frequency is relatively low. Though *profusely* is a relatively rare word (occurring 348 in the dataset), 30% of all of its occurrences are as a collocate of *apologise* and thus it has a high z-score.

By building a profile of the collocates of all of the IFIDs in this manner, Lutzky and Kehoe (2017a) were then able to aggregate the collocates across all of the IFIDs to identify the "shared collocates" (see Lutzky and Kehoe 2017a: 46-47). This showed some interesting patterns, for example, the pronoun *I* was a shared collocate of seven of the eight IFIDs (it was not a collocate of *apology* within the four word parameters set for the study) while *ignorance* only collocated with *pardon, excuse* and *forgive*. Interestingly, the reason for apologising within this genre (blogs) was reflected in this list of shared collocates, such as *spelling, typos, poor, quality* and *English*. Additionally, Lutzky and Kehoe (2017a) identified the items which were strong collocates with a given IFID but did not appear within the top 100 most frequent collocates of any other IFID. Among their findings in this set of results was the strikingly colloquial items that uniquely collocated with *sorry*, such as *oops, aww, hugs, sucks, hon/hun* (short for the endearment *honey*). They further investigated *oops* in Lutzky and Kehoe (2017b) and asserted that it could be added to the list of IFIDs for apologies in blogs.

Lutzky and Kehoe (2017a) and (2017b) offer a fascinating insight into new and evolving ways of investigating speech act phenomena in very large corpora. By profiling the similarities and differences in the collocational patterns of several IFIDs, Lutzky and Kehoe (2017a: 54) show that "functional overlaps" and "divergences can be revealed, which can in turn be used to increase the incidence of relevant examples in the search output". This ultimately means arriving at greater precision in the automated retrieval of speech acts in large corpora. The authors strongly advocate the place and merits of manual analysis, but they note, "our methodological approach allowed us to streamline the search for the fairly routinized speech act of apology in our blog data" (Lutzky and Kehoe 2017a: 54).

4.4.2 Using genre-specific search inventories from smaller samples

Kohnen (2008), in his work on directives over text and time, offers an interesting bottom-up methodology which essentially involves moving from a search inventory drawn from a micro-analysis of a representative genre sample of sermons to extracting forms from a large-scale diachronic corpus. For Kohnen, the first step was crucial. It involved manually sifting through a pilot dataset of church sermons to identify all possible forms of directives. Kohnen concedes that this "by hand" approach is "extremely labour-intensive" (2008: 296) but it generated a plausible search inventory that formed the basis of his study and which is useful for others who wish to investigate this speech act. In the initial process, there was some iteration as the pilot micro-analysis was scaled up to a broader representation of the genre – e.g. also looking at prayers, church letters, etc. Kohnen (2008: 296-7) notes, "[t]his microanalysis will probably reveal similar as well as different manifestations of the speech act, enriching the initial list of manifestations. It will also give an account of their frequencies and distribution across time".

The next step was to select manifestations of directives and their distribution in larger multi-genre corpora so as to further refine the inventory and test their frequency and distribution. This iterative process led to a principled inventory of forms in a genre-specific historic contexts. Over time, it shows the profile of a speech act within a genre and ultimately offers a robust means of moving from a micro-analysis of a speech act in a small representative sample of a genre to a largescale analysis in a diachronic dataset. Kohnen (2008: 297) reflects that by using this approach, "we could find out about genre-specific profiles and about speech-act conventions which may or may not apply in certain genres, and we could trace the development of these phenomena in the history of English". Jucker and Taavitainen (2013) observe that Kohnen's method will provide most reliable results for those patterns that are most frequent and most conventionalised. They note, however, that it is far less reliable for rare and creative patterns and that it also relies on the availability of a sufficient amount of data which is relevant and which is spread across the period of investigation. For Kohnen (2008), this approach worked well because there was a consistent sample of sermons, and related texts, over time.

4.4.3 Using searches of typical lexical or grammatical features associated with a speech act

Another interesting approach to analysing a speech act in a corpus is found in Taavitsainen and Jucker (2008). They sought to examine compliments in three historical corpora. Faced with the challenge of how to retrieve these, they used as their workaround adjectives that express positive evaluations, that is search items such as: *beautiful*, *nice*, *great*, *lovely*, and lexical strings, such as *really nice*, *really great*, *well done*, *like/love your*, *what a*, *you look/'re looking.* Reflecting on the process, Jucker and Taavitsainen (2013: 107) note that while this process did provide relevant hits, it also returned a lot of passages that were not relevant to the research focus.

The scale of their research, in terms of breadth of sources and span of time, meant that they were able to make a number of interesting statements about compliments in an historical context. For example, in the of Early Modern and Late Modern English, compliments were found to be gendered. Both male and female authors were found to use compliments in their writing but female characters received praise for their look(s) and often turned these down as flattery. On the other hand, in the case of males receiving compliments, they accepted them, by bowing. They note that this aligned with social norms of the time.

While most compliments were related to physical appearance and possessions, interestingly, Taavitsainen and Jucker (2008: 218) found only a few instances of compliments about food and they speculate that perhaps this is due to "the social norms, protagonists being mostly upper class and not directly associated with the preparation of meals".

4.4.4 Using metacommunicative expressions

Focusing the development of compliments in American English from a diachronic perspective of almost two hundred years, Jucker and Taavitsainen (2014) used both the 400-million-word *Corpus of Historical American English* (COHA) and the 425-million-word *Corpus of Contemporary American English* (COCA). Building on Taavitsainen and Jucker

(2008), they developed a systematic approach to the analysis of compliments using a "metacommunicative expression analysis" (Jucker and Taavitsainen 2014: 258). In essence, this approach entailed using the search term *compliment* to retrieve performative, descriptive and discursive instances of the act. This method was a useful starting point even though the term was mostly not used as a compliment. The process allowed the researchers to negotiate the value of a given act or to describe someone as having paid a compliment. In this process, their first step was to narrow their sample across a spread of five sub-corpora of selected decades within the sample period. These were sampled based on the number of instances of *compliment*. Subsequently, coders used the extended contexts of the node occurrence to categorise and code compliments for variables such as, type, complimenter, complimentee, object of the compliment, compliment response, as well as logging the genre in which it occurred. In their analysis, they distinguished between personal compliments and ceremonious compliments and found that in the historical data, more than 90% were personal compliments. Within that profile, they noted a steady decline in the use of ceremonious compliments, over time. Their analysis offers details on the distribution of the gender of the complimenter and complimentee and compares that with the contemporary data sample from COCA. This showed that males were in the role of complimenter between 70 and 85 percent of the time in the historical data while in the COCA sample, this statistic fell to 67.1 percent. In terms of the gender of complimentee role, there was a balance, apart from the earliest dataset (1820/1830), where males also received the majority of the compliments. The object of the compliment showed a consistent pattern across the centuries where most compliments were given on people's personality/friendship and on their ability/performance.

In looking at how compliments are responded to in contemporary American English, it has been shown that they are normally accepted (see Chen 1993). Jucker and Taavitsainen (2014) were keen to test this historically through their coding of the response to the compliment within their dataset. They found that, "acceptance of compliments remained more or less stable for the first four periods under investigation [1820/1830, 1870,1900] but it is clearly higher in the most recent period [1990/2000], in which it has reached more than 70 per cent" (Jucker and Taavitsainen 2014: 273). They speculate that this significant rise in acceptance may be connected with social and cultural changes, or perhaps a change in literary styles.

Reflecting on the metacommunicative expression analysis methodology that they deployed in this study, Jucker and Taavitsainen (2014) note that it has strengths and weaknesses:

> It allows the systematic analysis of a specific speech act in large corpora, and thus it provides a way to investigate synchronic differences or diachronic developments which would be inaccessible to other methods of investigation. On the other hand, the method mostly retrieves accounts of a particular speech act rather than the actual speech acts, and statistical results based on such accounts may be misleading. In the case of compliments, for instance, the retrieved passages may contain a disproportionate amount of problematic compliments, such as utterances whose status is unclear to the participants. Such problematic compliments may, of course, differ in systematic ways

from a large number of unproblematic compliments that are given and received in a graceful manner without any need to explicitly talk about them (Jucker and Taavitsainen 2014: 274).

5. Conclusion

There is no going back to the days before corpora; corpus pragmatics will only grow stronger amid advances in annotation models, resources and tools. It is important that within this rapid stream of progress that we are not tempted to see easily generated computations of forms as a substitute for the qualitative depth that is needed to fully understand how the meaning of these forms manifests in context. Taavitsainen (this volume) recalls the caveats of Rissanen (1989) in the early days of historic corpus linguistics. Rissanen could see that diachronic corpora offered so much to the field of philology and had so much optimism in terms of what the power and scope of CL could bring but he flagged the concern that "the corpus revolution would turn to mere number-crunching" (Rissanen 1989: 17). Though Rissanen's fears did not materialise for historic corpus linguistics, it was and still is healthy to be mindful of these words. Corpus pragmatics is at a relatively early stage and there is so much potential for both form-to-function and function-to-form approaches (and indeed a combination of both). It is important for this developing sub-field that we reflect more on how methodological approaches can be enhanced through the development of more pragmatic annotation tools, search and retrieval protocols and resources. Amid the endless growth in data size and ease of availability, we need to keep mindful of the fact that pragmatic insight often starts with small-scale scoping work, such as we have seen in the work of Kohnen (2008), Weisser (2015) and Garcia McAllister (2015). We also see that "corpus toiling" pays off. The painstaking work of Deutschmann (2003) has facilitated development in IFID collocational profiling by Lutzky and Kehoe (2017a and 2017b) or the insights which Taavitsainen and Jucker (2008) gained in the analysis of compliments using typical features of positive adjectives to aid recall led to further refinement in later work on metacognitive expressions (see Jucker and Taavitsainen 2014).

All of these "small-scale" steps, in the larger scheme of mega-corpora, are leaps in our understanding of the intricacies of building corpora that are fit for the purpose of both form-to-function and function-to-form approaches to research. An important lesson from successful examples of corpus-based function-to-form work to date is the link between the level of detail and consistent of the corpus metadata and the depth and scope of the results that have been generated about a given speech act, or related phenomenon (Deutschmann 2003 is a good example of this). The importance of gathering context-rich metadata should not be missed by those who are designing a corpus of any scale. The capturing of the subtleties of any given context will make the dataset more fruitful for pragmatics research for centuries to come.

References

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

Aijmer, Karin 1996 *Conversational routines in English: Convention and creativity*. London: Longman.

Aijmer, Karin [this volume] Corpus pragmatics: From form to function

Ädel, Annelie and Randi Reppen 2008 The challenges of different settings: An overview. In: Annelie Ädel and Randi Reppen (eds.), *Corpora and Discourse: The Challenges of Different Settings,* 1–6. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Beebe, Leslie M. and Martha C. M. Cummings 1996 Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In: Susan M. Gass and Joyce Neu (eds.), *Speech Acts across Cultures*, 65–86. Berlin: Mouton de Gruyter.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd and Maria Helt 2002 Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly* 36: 9–48.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd and Maria Helt, Victoria Clark, Viviana Cortes, Eniko Csomay and Alfredo Urzua 2004 *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. Princeton: Educational Testing Service.

Billmyer, Kristine and Manka Varghese 2000 Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests. *Applied Linguistics* 21(4): 517–552.

Blum-Kulka, Shoshana, Juliane House and Gabriele Kasper (eds) (1989) *Cross-Cultural Pragmatics: Requests and Apologies*. Norwood: Ablex.

Brinton Laurel J. 2012 Historical pragmatics and corpus linguistics: problems and strategies. *Language and Computers* 76: 101 – 131.

Bodman, Jean W. and Miriam Eisenstein 1988 May God increase your bounty: The expression of gratitude in English by native and non-native speakers. *Cross Currents* 15: 1-21.

Channell, Joanna 1994 *Vague Language*. Oxford: Oxford University Press.

Chen, Rong 1993 Responding to compliments: A contrastive study of politeness strategies between American English and Chinese speakers. *Journal of Pragmatics* 20 (1): 49–75.

Cheng, Winnie and Anne O'Keeffe 2015 Vagueness. In: Aijmer, Karin and Christoph Rühlemann (eds.), *Corpus Pragmatics: A Handbook,* 360–378. Cambridge: Cambridge University Press.

Clancy, Brian 2015 *Investigating Intimate Discourse: Exploring the Spoken Interaction of Families.* Abingdon: Longman.

Clancy, Brian and Anne O'Keeffe 2015 Pragmatics. In: Douglas Biber and Randi Reppen, (eds.), *The Cambridge Handbook on Corpus Linguistics,* 235–251. Cambridge: Cambridge University Press.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec Christopher Potts 2013 A computational approach to politeness with application to social factors. In: *Proceedings of ACL 2013*. Online at: www.mpi-sws.org/~cristian/Politeness.html (accessed January 2017).

Davies, Mark (2004) *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). Available online at https://corpus.byu.edu/bnc/.

Deutschmann, Mats 2003 *Apologising in British English*. (Skrifter från moderna språk 10). Umeå: Institutionen för moderna språk, Umeå University.

Farr, Fiona, Bróna Murphy and Anne O'Keeffe 2004 The Limerick Corpus of Irish English: Design, description and application. *Teanga* 21: 5–30

Fringinal, Eric, Marsha Walker and Janet Beth Randall 2014 Exploring mega corpora: Google Ngram Viewer and the Corpus of Historical American English. *EuroAmerican Journal of Applied Linguistics and Languages.* 1(1): 48–68.

Flöck, Ilka and Geluykens, Ronald 2105 Speech Acts in Corpus Pragmatics: A quantitative contrastive study of directives in spontaneous and elicited discourse. In: Jésus Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 201,* 7–37. London: Springer.

Garcia, Paula 2007 Pragmatics in academic contexts: A spoken corpus study. In: Mari C. Campoy and María J. Luzón (eds.), *Spoken Corpora in Applied Linguistics*, 97–128. Bern: Peter Lang.

Garcia McAllister, Paula 2015 Speech acts: A synchronic perspective. In: Aijmer, Karin and Christoph Rühlemann (eds.), *Corpus Pragmatics: A Handbook,* 29–51. Cambridge: Cambridge University Press.

Goffman, Erving 1979 *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Goffman, Erving 1974 *Frame Analysis: An Essay on the Organization of Experience*. Cambridge, MA: Harvard University Press.

Geluykens Ronald and Gert Van Rillaer 1995 Introducing ACID: The Antwerp Corpus of Institutional Discourse. *Interface. Journal of Applied Linguistics* 10(1): 83–101.

Hartford, Beverly S. and Kathleen Bardovi-Harlig 1992 Experimental and observational data in the study of interlanguage pragmatics. In: Lawrence F. Bouton and Yamuna Kachru (eds.), *Pragmatics and Language Learning 3*, 33–52. University of Illinois, Urbana-Champaign: Division of English as an International Language.

Jucker, Andreas H. 2013 Corpus pragmatics. In: Jan-Ola Östman and Jef Verschueren (eds.), *Handbook of Pragmatics*, 2–17. Amsterdam: Benjamins.

Jucker, Andreas H. and Irma Taavitsainen 2013 *English Historical Pragmatics*. Edinburgh: Edinburgh University Press.

Jucker, Andreas H. and Irma Taavitsainen 2014 Complimenting in the history of American English: A metacommunicative expression analysis. In: Irma Taavitsainen, Andreas H. Jucker and Jukka Tuominen (eds.) *Diachronic Corpus Pragmatics*. (Pragmatics & Beyond New Series 243), 257-276. Amsterdam: John Benjamins.

Jucker, Andreas H., Irma Taavitsainen and Gerold Schneider 2012 Semantic corpus trawling: Expressions of 'courtesy' and 'politeness' in the Helsinki Corpus. In: Carla Suhr and Irma Taavitsainen (eds.), *Developing corpus methodology for historical pragmatics* (*Studies in variation, contacts and change in English 11*). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/11/jucker_taavitsainen_schneider/.

Kallen, Jeffery. L. and John M. Kirk 2012 *SPICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banríona.

Kirk, John M. 2012 Beyond the Structural Levels of Language: An Introduction to the SPICE-Ireland Corpus and its Uses. In: Janet Cruickshank and Robert McColl Millar (eds.), *After the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster Triennial Meeting*, *Aberdeen 2012,* 207–232. Aberdeen: Forum for Research on the Languages of Scotland and Ireland.

Kirk, John M. 2016 The Pragmatic Annotation Scheme of the SPICE-Ireland Corpus. *International Journal of Pragmatics* 21(3): 299–322.

Kirk, John M. and Gisle Andersen 2016 Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics* 21(3): 291–298.

Kirk, John M., Jeffery L Kallen, Orla Lowry, Anne Rooney, A and Margaret Mannion 2011 The *SPICE-Ireland Corpus: Systems of Pragmatic Annotation for the Spoken Component of ICE-Ireland*. [Version 1.2.2.] Belfast: Queen's University Belfast & Dublin: Trinity College Dublin.

Koester, Almut J. 2002 The performance of speech acts in workplace conversations and the teaching of communicative functions. *System* 30: 167–184.

Kohnen, Thomas 2008 Tracing directives through text and time: Towards a methodology of a corpus-based diachronic speech-act analysis. In: Andreas H. Jucker and Irma Taavitsainen (eds.), *Speech Acts in the History of English,* (Pragmatics & Beyond New Series 176), 295–310. Amsterdam/Philadelphia: John Benjamins.

Lutzky, Ursula and Andrew Kehoe 2017a "I apologise for my poor blogging": Searching for Apologies in the Birmingham Blog Corpus. *Corpus Pragmatics* 1:37–56.

Lutzky, Ursula and Andrew Kehoe 2017b "Oops, I didn't mean to be so flippant". A corpus pragmatic analysis of apologies in blog data. *Journal of Pragmatics* 116: 27-36.

McCarthy, Michael J. 1998 *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McCarthy, Michael J. and Anne O'Keeffe 2003 "What's in a name?" – vocatives in casual conversations and radio phone in calls. In: Pepi Leistyna and Charles F. Meyer (eds.), *Corpus Analysis: Language Structure and Language Use*, 153–185. Amsterdam: Rodopi.

Milà-Garcia, Alba Forthcoming Pragmatic annotation for a multilayered analysis of speech acts: A methodological proposal. *Corpus Pragmatics*.

O'Keeffe, Anne 2006 *Investigating Media Discourse*. Abingdon: Routledge.

O'Keeffe, Anne, Michael J. McCarthy and Ronald A. Carter 2007 *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

O'Keeffe, Anne, Brian Clancy and Svenja Adolphs 2011 *Introducing Pragmatics in Use*. Abingdon: Routledge.

Rissanen, Matti 1989 Three problems connected with the use of diachronic corporal. *ICAME Journal* 13: 16–22.

Romero-Trillo, Jésus (ed.) 2008 *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Berlin/New York: Mouton de Gruyter.

Rühlemann, Christoph and Brian Clancy Forthcoming Corpus linguistics and pragmatics. In Neal Norrick and Cornelia Ilie (eds.), *Pragmatics and its Interfaces*. Amsterdam/Philadelphia: John Benjamins.

Rühlemann, Christoph and Karin Aijmer 2015 Corpus pragmatics: Laying the foundations. In: Karin Aijmer and Christoph Rühlemann, (eds.), *Corpus Pragmatics: A Handbook,* 1–26. Cambridge: Cambridge University Press.

Rühlemann, Christoph and Matthew 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.

Sasaki, Miyuki 1998 Investigating EFL students' production of speech acts: A comparison of production questionnaires and role plays. *Journal of Pragmatics* 30: 457–484.

Schauer, Gila and Svenja Adolphs 2006 Expressions of gratitude in corpus and DCT data: vocabulary, formulaic sequences, and pedagogy. *System* 34(1): 119–134.

Searle, John R. 1976 A Classification of illocutionary speech acts. *Language in Society* 5(1): 1–23.

Stiles, William B. 1992 *Describing Talk: A Taxonomy of Verbal Response Modes*. Newbury Park: Sage.

Taavitsainen, Irma [this volume] Historical corpus pragmatics.

Taavitsainen, Irma and Andreas H. Jucker 2007 Speech act verbs and speech acts in the history of English. In: Susan M. Fitzmaurice and Irma Taavitsainen (eds.), *Methods in Historical Pragmatics*, 107–137. Berlin: Mouton de Gruyter.

Taavitsainen, Irma and Andreas H. Jucker 2008 "Methinks you seem more beautiful than ever": Compliments and gender in the history of English. In: Andreas H. Jucker and Irma Taavitsainen (eds.), *Speech Acts in the History of English,* 195–228.

Amsterdam/Philadelphia: John Benjamins.

Taavitsainen, Irma and Andreas H. Jucker 2015 Twenty years of historical pragmatics: Origins, developments and changing thought styles. *Journal of Historical Pragmatics* 16(1): 1–24

Verschueren, Jef 1999 *Understanding Pragmatics*. London: Arnold.

Weisser, Martin 2015 Speech Act Annotation. In: Karin Aijmer and Christoph Rühlemann (eds.), *Corpus Pragmatics: A Handbook,* 84–110. Cambridge: Cambridge University Press.

Woodman, Gill 2005 Review of Mats Deutschmann, Apologising in British English. *Language in Society* 34: 314–317.

Yuan, Yi 2001 An inquiry into empirical pragmatics data-gathering methods: Written DCTs oral DCTs, field notes, and natural conversations. *Journal of Pragmatics* 33: 271–292.