Vaughan, E. & B. Clancy. 'Small corpora and pragmatics,' *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

Small corpora and pragmatics

Elaine Vaughan (University of Limerick, Ireland)

Brian Clancy (Mary Immaculate College, Limerick)

Abstract

Corpus linguistics is more often than not associated with large-scale collections of spoken or written data, representing genres, varieties or contexts of use. Many of these have been successfully exploited for pragmatics research, producing generalised findings that hold across a range of texts. However, it may be argued that rather than stopping at generalised findings that note the frequency of pragmatic phenomena in large corpora, an important research agenda now foregrounds a focus on small corpora and local pragmatic patterns. This paper will argue that smaller, carefully collected, context-specific corpora, both spoken and written, are of great import in pragmatics research. Many pragmatic features of language such as deixis or pragmatic markers play a fundamental role in communication, and, in these cases, are linguistically realised in the type of 'small' linguistic items that tend to be frequent in all corpora. Therefore, smaller corpora provide a platform for not only establishing the range and frequency of these items but the role of different genres or contexts in characterising their use. We will provide evidence for this in the form of two corpus case studies in order to illustrate how small corpora have created a practical and empirical route for the study of pragmatics, and how this synergy of small corpora and pragmatic research provides rich and contextualised findings.

1. Introduction

In this paper we argue for the benefits of using small, domain-specific corpora in pragmatic research, and this position presupposes a number of questions. The first of these questions relates to the establishment of what we mean by 'small' corpora, in what context this characterisation developed, and how this is relevant to the type of studies we review and present. The second regards to what extent corpus methodology can assist research on pragmatic phenomena, and what type of insights this empirical orientation can generate. Below we attempt to answer the questions above and frame them in general and in relation to two studies which use small corpora to investigate the pragmatics of how identities are indexed in two different speech contexts. With regard to our first question, any discussion of small corpora raises the question 'what do we mean by 'small'?', and this is worth pondering for two reasons: firstly, and instrumentally, answering this question will define our parameters in talking about 'small corpora and pragmatics' in general. More importantly, it raises some issues in connection with corpus linguistics as it has developed in the last few decades that prompts our position as to why 'small' corpora can be of benefit to pragmaticists.

The emergence of modern corpus linguistics is primarily associated with lexicography and the pioneering work of researchers such as John Sinclair. This research was predicated on creating the largest possible corpora, which in the 1960s and 1970s, as Sinclair (2001: viii) points out, 'were simultaneously the largest and smallest of their type being the only ones'. Early COBUILD corpora contained tens of millions of words and, as technology advanced, so too did the size of these corpora (Sinclair, 2001; McCarthy and O'Keeffe, 2010; Tognini-Bonelli, 2010). For example, the Collins corpus, which incorporates COBUILD, now contains 2.5 billion words of running text, with the Oxford English Corpus approximately 2 billion words; the Cambridge English Corpus, which comprises samples of British, American and Learner English consists of many billions of words. There does not appear to be any upper limit on language corpora; indeed, some discussions of 'corpus' and 'corpus linguistics' have explicitly (e.g. Biber *et al.* 1998) integrated 'large' as a defining feature, and the prevailing philosophy for corpora such as those mentioned above seems best summed up by the motto of the American-based Linguistic Data Consortium, there is 'no data like more data' (Sinclair, 2001). As corpus linguistics has developed, it has come to be associated with many aspects of language study, such as language variation studies, historical linguistics or language pedagogy and it is now possible to access a range of large corpora designed for these purposes. Corpora such as the American National Corpus (ANC) and the British National Corpus (BNC) are designed to represent the language varieties of American and British English respectively and are also designed to be comparable across genres. The BNC contains 100 million words, of which 10 million are spoken.[1] The International Corpus of English (ICE) brings together one-million-word samples from eighteen countries which have English as their first or official language, with 60% of each sample consisting of spoken texts, although some of these texts are scripted and/or monologues (see http://ice-corpora.net/ice/index.htm). The Corpus of Contemporary American English (COCA), the largest freely available corpus, is made up of over 450 million words in more than 175,000 texts, including 20 million words from each year from 1990-2011 (see corpus.byu.edu/coca/). The picture in terms of what are glossed as 'historical corpora' is no less impressive in terms of size. The Oxford Text

[1] Almost 15 million words of the ANC are currently available. This is divided into approximately 11.5 million words of written language and 3.5 million words of spoken language (see www.anc.org).

Archive houses a number of these corpora (http://ota.ahds.ac.uk/). Table 1.1 below gives a brief overview of some of these large corpora:

Table 1.1: Examples of large corpora

Corpus	**Number of words (approx.)**	**Overview of composition**
Collins Corpus and the Bank of English™	2.5 billion	Written: e.g. websites, magazines, newspapers, books Spoken: e.g. radio, TV, everyday conversations
Oxford English Corpus	2 billion+	Mainly written material from World Wide Web, e.g. academic papers, technical manuals, corporate websites, personal websites, blogs
Cambridge English Corpus	2 billion+	Written and spoken English from a range of domains, e.g. books, newspapers, letters, e-mails, websites, conversations, meetings, radio
British National Corpus	100 million	Written (90%): e.g. newspapers, books (fiction/non-fiction), letters, school/university essays Spoken (10%): e.g. informal conversations, business and government meetings
International corpus of English	1-million-word samples	Different varieties of English (e.g. British, Irish, Hong Kong, Singapore, East African English) Written: e.g. letters, academic writing, newspaper reports Spoken: e.g. conversations, meetings, radio
Corpus of Contemporary American English	450 million	Written: e.g. fiction, popular magazines, newspapers, and academic texts Spoken: e.g. television and radio programmes

In terms of language pedagogy, the Cambridge Learner Corpus (part of the Cambridge English Corpus mentioned above) contains 43 million words of written and spoken learner English across the proficiency levels and the International Corpus of Learner English is a 3.7 million word corpus of English as a Foreign Language writing from learners from 16 different mother tongue backgrounds (see http://www.uclouvain.be/en-cecl-icle.html).

1.1. Small corpora in corpus linguistics

It is hard to imagine describing any of the corpora mentioned above as 'small', and, in fact, defining our terms here requires the caveat that 'small' is relative, related to modality (the term 'modality' is used here in its loosest sense as occupying some point on the speech to writing continuum, cf. Biber, e.g. 1988), and is, inevitably, 'frequently reinterpreted' (Sinclair, 2001: xiii). Beside the behemoths of the major publishing houses, the corpora of national varieties mentioned above appear small. While it seems to be accepted that the upper limit of a small corpus is approximately 200,000-250,000 words (see Aston, 1997; Flowerdew, 2004), one- to five-million-word samples have also been described as 'small' (McCarthy 1998; Sinclair, 2001). Aston (1997) notes that small corpora exist in the 20,000-200,000 word range, and are more specialized in terms of topic and/or genre than large corpora. In terms of modality, of relevance to corpus size is the type of corpus in question. Spoken corpora – the principle focus of this paper – are often, by necessity, smaller than written corpora. There are a great number of reasons for this, not least of which is the fact that spoken data still need to be manually transcribed to adequately represent the speech event, and even manual transcription does not completely represent the complexities of spoken interaction. Multi-modal corpora are still very much in the minority, although great strides have been made in this regard (see, for example, Knight *et al.* (2009)). A major factor behind the development of small corpora has not necessarily been the corpus linguistic research agenda *per se*, but something else entirely: the emergence of small corpora is directly related to technological developments (Sinclair, 2001). In the past, assembling a large amount of data was associated with high costs because of the difficulties involved in recording, transcribing and coding the data. Data can now be easily collected, assembled, stored and analysed on a PC, thereby 'democratising' the notion of corpus building and corpus linguistics (cf. Rundell, 2008: 26).

What we are implying is that it has not always been a given that corpora considered 'small' had full legitimacy in the field of corpus linguistics. A major reason for this reluctance to fully admit small corpora to the fold was rooted in, as previously mentioned, the predominant research agenda in corpus linguistics in its 'early modern' period, lexicography, and the remediation of concerns in relation to 'representativeness' and 'balance' in commercial corpus building. Corpora used for lexicographical research need to be as large as possible in order to generate sufficient occurrences which reflect how lexical items are used, and, as previously mentioned, these large corpora, such as the Bank of English, dominated research publications representing the 'output' of corpus linguistics. Representativeness, or 'or the extent to which a sample includes the full range of variability in a population' (Biber, 1993: 243) has been a challenge in relation to language data, and, as Clear (1992: 21) points out, it is difficult to interpret the statistical notion of 'population' in relation to a phenomenon like language. One response to this difficulty, which has now become standard, has been to approach the sampling of language data in a different way. Biber (1993) proposes strata and sampling frames for representative corpus design based on register, or situationally defined text categories such as 'fiction', 'news article' etc., and linguistically defined text types, such as various written or spoken modes. In terms of the balance of a corpus, Sinclair (2005) refers to it as a rather vague notion but important nonetheless. Balance appears to rely heavily on intuition and best estimates (Atkins *et al.*, 1992; Sinclair, 2005; McEnery *et al.*, 2006). In terms of a large corpus, the Longman Spoken and Written English Corpus (LSWE) is considered 'balanced'. According to Biber *et al.*, (1999: 25), the registers contained within the corpus were selected on the basis of balance in that they 'include a manageable number of distinctions while covering much of the range of variation in English.' For example, conversation is the register most commonly encountered by native speakers whereas academic prose is a highly specialised register that native speakers encounter infrequently. Between these two extremes are the popular registers of newspapers and fiction. For a more specialised corpus, balance is reliant on the corpus containing a range of texts typical of what the corpus is said to represent. In terms of small corpus compiling, a small corpus should be approached with as much caution as building a large corpus, and issues of balance and representativeness are salient no matter the size of the corpus. A small corpus builder can address issues of representativeness by ensuring that the samples collected are typical of the speech domain represented by the corpus. For example, the corpus of family discourse discussed in section 3 features members of that family engaged in eating a meal, putting up the Christmas tree, talking about being a student in university and providing information about a city one of them is going to visit, speech situations typical of most families and, therefore, considered 'representative' (Clancy, 2010). McEnery *et al.* (2006: 5) maintain that if specialised corpora were discounted on the basis of sampling techniques used, then 'corpus linguistics would have contributed significantly less to language studies' and this is an enlightened and crucial point to keep in mind.

Sociolinguistic studies have shown that relatively small samples that could be considered technically unrepresentative are sufficient to account for language variation in large cities (see Sankoff, 1988; Tagliamonte, 2006). McEnery *et al.* (2006: 73) claim that although representativeness and balance are features that must be considered in relation to corpus design, they often depend on the ease with which the data can be collected (and, of course, the nature of the data itself) and, therefore, 'must be interpreted in relative terms i.e., a corpus should only be as representative as possible of the language variety under consideration.' They believe that corpus building is 'of necessity a marriage of perfection and pragmatism' (*ibid.*), echoing Stubbs' (2004) contention that corpus size tends to be 'a compromise between the desirable and the feasible' (p. 113). Flowerdew (2002: 96) maintains that now 'the field [of corpus linguistics] has widened considerably to include the recognition of much smaller, specialised genre-based corpora’. Small corpora have been instrumental in pushing the boundaries of corpus linguistics as a field of enquiry, and have been similarly so in prompting a shift towards empiricism in the realm of pragmatics research (cf. Romero-Trillo, 2008). The review below does not purport to, nor would it be possible to, represent the totality of the literature available on small corpora in relation to pragmatics; instead it is intended to be selective and illustrative, for our purposes, of what working empirically with small corpora and a pragmatic agenda can uncover.

2. The use of small corpora in pragmatic research: A selective review

The primary benefit of small corpora to the study of pragmatics is a fundamental one: they can enable the researcher to access authentic, naturally occurring language and to maintain a close connection between language and context. Indeed, in relation to contextual links and small corpora, Koester (2010) points out that small corpora have a clear advantage over larger ones. She maintains that large corpora are sampled from such a variety of different contexts that it is ‘very difficult, if not impossible, to say anything about the original contexts of use of the utterances’ (*ibid*: 66-67; see also Flowerdew, 2004). While it is certainly possible to investigate phenomena such as hedging using large corpora, this can be a major challenge due to the variety of (para)linguistic selections available for use as hedges. Using a small, context-specific corpus offers significant advantages. These phenomena can not only be investigated in their original context of use, it is also usually possible to investigate virtually all occurrences and essay a refined analysis which takes the polysemous nature of many pragmatic features into account. Therefore, we can move from quantitative observations regarding frequency of items with pragmatic potential, which only tell part of the story. The studies summarised below have turned up contextualised findings in relation to the pragmatic significance of linguistic and extra-linguistic strategies as diverse as question forms, modality, small talk, humour and (evaluative) speech acts.

In the public sphere of media discourse, O’Keeffe (2005) used a 55,000 word corpus from radio phone-in to focus on question forms as they are used in this context, which from other analytical perspectives – for example, conversation analysis – displays a fairly typical (and canonical) turn-taking structure with the presenter holding the discursive power. However, although many asymmetrical norms of institutional discourse do apply to this context, there is widespread downtoning of power at a lexico-grammatical level. In addition to using pragmatic markers to hedge, the presenter of the radio show employs a variety of features such as first name vocatives, latching and reflexive pronouns, as in *you’ve a daughter yourself?*, to create a ‘pseudo-intimate’ (p. 340) environment between speaker and caller. Also in the public sphere, but in a more difficult to access ‘occluded genre’ (Swales, 1996; Loudermilk, 2007), Koester (2006) created a 34,000 word corpus of American and British office talk and demonstrated the influence of local contexts on frequency and use of various phenomena, such as hedging and modality. She identifies a number of genres within the workplace discourse she investigates, and finds that modal verbs of obligation are more frequent in collaborative genres (for example, decision making or planning) than in unidirectional genres (for example, giving instructions). The boundaries between the genres she identifies are, however, fluid. She notes that there is no easy distinction between ‘on-task’ transactional talk and small or relational talk essential for building speaker relationships (p. 161) due to the complex nature of speakers’ interactional goals. Vaughan (2007, 2008) employs a 40,000 word corpus of meetings of English language teachers (C-MELT, see section 3 below) to explore particular linguistic features characteristic of this community of practice (Wenger, 1998). Part of this study involved exploring how the community managed and maintained itself, and looked at how power and solidarity are negotiated, for example through humour. The size of C-MELT allowed specific instances of humour to be isolated in order that they might be assigned a function. Vaughan (2007: 186) found that teachers ‘use [humour] to establish the social space they share, and implicitly define who they are, and what their attitude is to the work they do’; humour in this context is in fact a highly salient, ‘powerful, polyvalent pragmatic resource’ (Vaughan and Clancy, 2011: 51). Finally, in a study that is also situated in the institutional domain, Farr (2007) demonstrates how in teacher education, ‘a spoken language corpus can be a valuable instrument in the toolbox of professional development’ (p. 254) and her 80,000 word professional talk corpus allowed the identification of areas for development and, equally, good professional practice. She explores the use of relational strategies present in the data to demonstrate how trainers work to lessen asymmetrical speech relationships and claims that small talk, in particular talk about health issues, is a typical way of establishing solidarity between speakers in this context (Farr, 2005). Furthermore, she demonstrates how shared socio-cultural references such as *muinteóir*, the Gaelic word for *teacher*, are a method of diluting institutional power on the part of the teacher trainer in interaction with the trainee.

At this juncture, it is important to note that for all the studies mentioned here, the researchers were also the corpus compilers (and often participants also), and this close relationship between corpus and researcher further strengthens the advantage of small corpus research for pragmatics. As Koester (2010: 67) points out a feature of small corpus research is that the researchers themselves often has a high degree of familiarity with the context and this ensures that the quantitative corpus results generated can be ‘balanced and complemented with qualitative findings’ such as information about setting, participants and purpose. Cutting (2001) investigated the evaluative speech acts of six students as they became members of an academic discourse community, on a taught Master’s course in Applied Linguistics. Cutting isolated and tagged each of these speech acts and found that positive acts increase as the course progresses and participants build solidarity. She also found that negative speech acts are most common in conversations about the course. Cutting explicitly states that she deliberately limited the corpus used to 26,000 words so that she ‘could become familiar enough with each one’s [participants] linguistic idiosyncrasies, personalities and attitudes to interpret the findings’ (p. 1208-1209), an approach that would be very difficult with a larger corpus. This is not to say that a similar level of familiarity cannot not be attained by researchers who are not the corpus compilers in these cases. The cogent point is that the smaller sizes of the corpora facilitate ease of familiarisation.

A significant advantage of using small corpora for this type of research, as previously touched upon, is that frequency information, while interesting, is insufficient for pragmatic characterisation or categorisation. In relation to the study of the epistemic function of modal markers in English for Academic Purposes (EAP), Holmes (1988) notes that there was little corpus frequency information in relation to the occurrence of modal markers for this specific context. This, she claims, is unsurprising given that a million-word corpus, even if it contains data from EAP, when searched will provide the analyst with approximately 3,000-4,000 tokens of modal forms such as *would*, and each of these tokens requires detailed contextual analysis in order to assign function. She maintains that with a smaller domain-specific corpus, however, ‘it is possible to establish both the range and the frequency of modal verbs expressing epistemic modality’ (p. 28). Within this wider issue of a surfeit of data is a connected and rather human one: as Orpin (2005: 39) suggests, ‘an attendant danger in using a large corpus is that the researcher may feel swamped by the huge amount of data s/he is faced with.’ In order to overcome this analytical barrier of large frequency count results, researchers seek to ‘manage’ the data, primarily through the process of normalisation. For example, Torgersen *et al.* (2011) analysed the use of pragmatic markers in the Linguistic Innovators Corpus and the Corpus of London Teenage Language. The 2,000 instances of *yeah* they examined in the study comprise only 10.7% of the total number of instances of the marker (18,693). Similarly, Clancy and Vaughan (2012), faced with 4,860 instances of the item *now* in the Limerick Corpus of Irish English (LCIE), provide a detailed analysis of 500 random instances (for both of these studies, it must therefore be acknowledged that the normalized frequency information presented in the discussion of the results is based on extrapolated figures).

Researchers using large corpora for pragmatic research have also used an iterative approach and smaller and larger corpora in order to fully interpret the initial frequency information the larger corpora generate. For example, O'Keeffe and Adolphs (2008) investigate the form and function of response tokens in two one-million-word spoken corpora: the Limerick Corpus of Irish English and a one-million-word sample from the Cambridge and Nottingham Corpus of Discourse in English. To put the sort of data generated in context, response tokens tend to be very high frequency items in spoken corpora. They examine the form taken by response tokens, a largely quantitative enterprise, in Irish and British English using the one-million-word samples and found that British speakers in general use a broader range of single and two-word response token forms than their Irish counterparts. However, in order to investigate response token *function* across the two corpora, a more qualitative and detailed process, they constructed two parallel 20,000-word corpora taken from the private sphere. These corpora were comprised of the speech of Irish and British females, all around the age of 20. The female participants were students and close friends who, in most cases, shared accommodation. They found that, again, in these smaller corpora, response tokens are more frequent in British English speech. However, they found no real variation at the level of the response tokens' pragmatic functions. In other words, drilling down into quantitative results using qualitative methodologies uncovers the subtleties of the pragmatic profile of particular items which extends beyond the limited, albeit interesting, information frequency provides.

3. A case study: 'we' in small corpora

Many of the studies above have in common a methodological approach that moves from general frequency counts to investigating items in context. What they also have in common is a focus on particular locations of discourse, for example, classroom discourse, family discourse or workplace discourse, and the linguistic features that characterise them. In many cases, the research explicitly details the pragmatic norms of the contexts or communities they study. This idea of being able to use a small domain-specific corpus to characterise the discourse of a particular community is intriguing, and, with this in mind, we show below two approaches to identifying the pragmatic function of the personal pronoun *we*. What we are looking at in a broad sense is indexicality, a central notion in pragmatics. It is axiomatic that for the study of pragmatics language and context are inseparable, and it has been argued that the '*single* most obvious way in which the relationship between language and context is reflected in the structures of the languages themselves, is through the phenomenon of deixis' (Levinson, 1983: 54, our italics). The phenomenon of deixis, therefore, serves as a constant reminder to us that language can only be interpreted within its context of use, moreover, as Hanks (1992: 48) observes, '…deixis links language to context in distinguishable ways, the better we understand it, the more we know about context’. A significant criticism of corpus linguistics in the past was its abstraction of language from its original context, and to an appreciable extent the fact that most small corpora contain samples of ‘complete’ texts mitigates this quite valid point: small corpus based pragmatic research is often conducted iteratively, with quantitative observations investigated in qualitative detail to account for frequency/infrequency. For the case studies presented below, while corpus methodologies dictated the research agenda and highlighted pragmatic areas of focus, the fact they were based on small corpora allowed us to investigate the phenomena in context, and thus reanimate the disembodied data returned by corpus searches.

The principal purpose of these two case studies was to investigate how identity is expressed by two quite different communities in two quite different contexts. The first case study uses a small corpus of family discourse recorded in Limerick, a city in the south of the Republic of Ireland. The second study uses a corpus of the meetings of English language teachers (C-MELT) compiled by recording meetings in two different geographical locations, México and Ireland. Table 3.1 provides more detail on the two different small corpora consulted for the study described below.

Table 3.1: Description of the two corpora

	C-MELT	**Family corpus**
Length of recording	3.5 hours	1 hour
Number of speakers	33	6
Number of words	39,975	12,531

A point of confluence for both studies was the contention that if pragmatics is about exploring how context and speaker relationship impact on language, then, as a corollary, control (or not) of pragmatic norms is also about demonstrating membership of a community.[2] Uncovering the pragmatic norms around identity work in these particular communities was hence the primary research focus for both studies. Identities are not monolithic however (De Fina *et al.*, 2006), but mutable, dynamic and situated (Tracy, 2002). We propose to look at how identities are expressed though a detailed examination of linguistic

[2] There are various frameworks and conceptualisations of ‘community’, such as the ‘speech community’ (e.g. Patrick, 2002), ‘discourse community’ (e.g. Swales, 1990), or ‘community of practice’ (Lave and Wenger, 1991; Wenger 1998). Both of the studies reported on in section 3 operationalise the notion of community of practice.

proxies for identity, personal pronouns, in order to get at the social relationships being indexed in talk, and the pragmatic management engendered in this process.

Rees (1983) posits pronominal use on a scale of ‘distance from the self’, where ‘I’ is closest to the self, and ‘they’ is the most distant:

0	1	2	3	4	5	6	7	8
I	we	you	one	you	it	she	he	they

The complexity of reference encoded in any one of these pronouns has been the subject of much linguistic, though not necessarily always pragmatic, research. If only because, as Mühlhäusler and Harré (1990) have argued, ‘any pronoun can be used for any person’. Complexities in what aspect of identity or what speech act a speaker invokes with ‘I’ are not immediately obvious, and it may appear one of the ‘least ambiguous’ pronouns (Fasulo and Zucchermaglio, 2002), though this is only at first sight. ‘I’ may not always index the speaker only, as in reporting direct speech, for example. In addition, say in the case of ventriloquising (Tannen, 2007), ‘I’ may not refer to the ‘animator’ of the statement, but to a postulated ‘author’ (Goffman, 1981), in the case of Tannen’s research (2007) the family dog.[3] Fasulo and Zucchermaglio (2002) investigate the multiple identities speakers invoke with ‘I’ in Italian work-place meetings. They present how various role identities are enacted, and show how these identities are situated, highlighting how the meanings of pronouns (for this research ‘I’) are layered according to the context in which they are invoked. ‘You’, which has a singular and plural reference in English (plural ‘you’ is positioned in the middle of Rees’ scale above), has an obvious addressee referent. However, it can also can be used in a generalised, ‘generic’ or ‘impersonal’ (e.g. Whitley, 1978) way, for example, to create a sense of distance or objectivity, or, alternatively, to emphasise or recruit involvement (O’Connor, 1994; Stirling and Manderson, 2011). *I* and *you* are prominent features on most spoken corpus frequency lists reflecting the canonical conversational dyad. Their high frequency is also due to features of online speech production such as repetition and reduplication, as well as their frequency in fixed pragmatic clusters such as *I think*, *I mean*, *you know* and so on.

[3] In this research, Tannen examines how speakers in family discourse use the family pet to interact with one another, allowing them ‘to distance themselves figuratively from their own utterances’ (2007: 417), for example, to defuse a potential conflict.

Pennycook (1994: 176) observes of the pronoun *we* that ‘depending on the speaker’s intention, “we” is the only personal pronoun that can (a) be inclusive and exclusive and (b) claim authority and communality at the same time’. While we could argue that it is not the ‘only’ pronoun to display this type of complexity (see above), it does present an interesting case. As previously mentioned, Mühlhäusler and Harré (1990) have shown that *we* is sufficiently flexible and multifunctional to encode any of the six persons that are usually referred to in English. Biber *et al.* (1999: 329) assert that ‘the meaning of the first person plural pronoun [*we*] is often vague: *we* usually refers to the speaker/writer and the addressee (inclusive *we*), or to the speaker/writer and some other person or persons associated with him/her (exclusive *we*). The intended reference can even vary in the same context.’ Inclusive and exclusive *we* can be used to create a perspective of: *I* the speaker + *you* the addressee(s) in the immediate context (‘inclusive *we*’) and *I* the speaker + someone else not in the immediate context (‘exclusive *we*’). An investigation of *we* allows us to examine how different speaker relationships *and* identities are negotiated locally and what this negotiation reveals and entails. In this sense, the pragmatics of personal pronoun usage and invocation of identity becomes critical to conceptions of community, with their natural and appropriate use about demonstrating membership of the community. Understanding speaker identity is crucial to understanding context and it has been shown in research on intercultural pragmatics that inability to inhabit appropriate identities in context can lead to pragmatic ‘failure’ (Thomas 1983). The first step in the analysis will be to see what looking at frequency information using corpus linguistic methodology can tell us about the pronoun *we*.

3.1 Frequency

Table 3.2 illustrates that *we* features prominently in the top 25 words of the C-MELT (position 11), family (position 18) and the British National Corpus (BNC) (position 13) corpora; however, interestingly, it does not feature in the top 25 words of the Limerick Corpus of Irish English (LCIE) corpus. *We* lies just outside the top 25 words in the LCIE, in position 28 (this is a potentially interesting anomaly which it is outside the scope of the current research to investigate). If there was no other agenda, the basis of frequency alone would make this item deserving of attention.

Table 3.2: Top 25 word frequency counts for four spoken corpora (*we* is shaded)

	C-MELT	**Family Corpus**	**LCIE**	**BNC (Spoken)**
1	the	the	the	the
2	to	you	I	I
3	I	it	and	you
4	and	I	you	and

5	yeah	to	to	it
6	that	a	it	that
7	of	and	a	a
8	you	of	that	’s
9	a	that	of	to
10	it	in	yeah	of
11	**we**	is	in	n’t
12	they	yeah	was	in
13	in	no	is	**we**
14	so	it’s	like	is
15	is	on	know	do
16	but	what	he	they
17	have	do	on	er
18	do	**we**	they	was
19	think	now	have	yeah
20	be	was	there	have
21	know	have	no	what
22	if	there	but	he
23	just	like	for	to
24	what	all	be	but
25	for	not	what	for

Obviously, C-MELT, the family corpus, LCIE and the BNC are different sizes and represent different types of talk. As already detailed in Table 3.1, C-MELT is comprised of c.40,000 words and the family corpus, c.12,500 words. As we know, the Limerick Corpus of Irish English is a one-million-word corpus, whereas the spoken component of the British National Corpus is comprised of 10 million words. Therefore, in order to properly compare the frequency of *we* across the four corpora it is necessary to normalise the frequency counts (in this case, *we* is normalised per million words). Additionally, in order to provide a more accurate picture of *we* across the four corpora, Figure 3.1 presents the normalised frequency per million words for the lemmatised WE, where WE includes *we*, *we’d*, *we’ll*, *we’re*, *we’ve* and *us*:

Figure 3.1: Distribution of WE across the four corpora (normalised per million words)

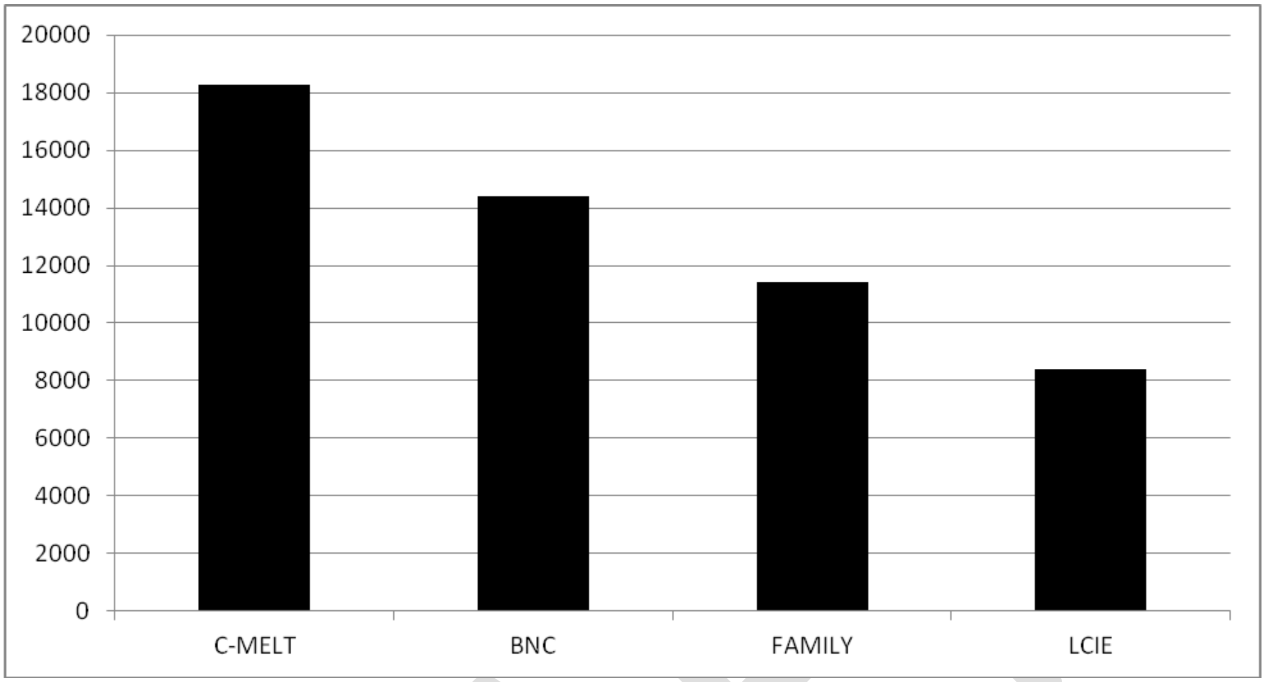


Figure 3.1 demonstrates that the lemmatised WE is most frequent per million words in C-MELT, followed by the BNC, the family corpus and LCIE. Of note here is that WE in C-MELT is more than twice as frequent as in LCIE. The reason for this is more than likely related directly to context. As already detailed, C-MELT is a corpus of workplace meetings, a professional context-type. LCIE is predominantly a corpus of informal spoken Irish English, where the casual conversation component accounts for over 70% of the corpus. This may indicate the importance of WE as a pragmatic item in professional discourse in comparison to informal discourse, such as that between family and friends. This appears to be supported by the fact that WE is 1.6 times more frequent in C-MELT than in the family corpus. In addition, C-MELT is composed of many speakers of different nationalities – Irish, British, American, Jamaican and Ugandan, for example, whereas in LCIE and the family corpus, the speakers are all Irish. Our speculations on why WE occurs with such frequency are inevitably linked to context, but also to intuitions regarding how and why certain identities are indexed in specific communities. Understanding how these identities are indexed is crucial to interpreting how these communities are supported, created and realised pragmatically. In the sections that follow, we will analyse the pragmatics of WE in the context of the family in relation to their inclusive and exclusive WE references, and for the workplace in relation to the indexical ground of WE (cf. Hanks, 1992).

3.2 Family discourse: Inclusive and exclusive WE

As has been mentioned, person reference, manifest in personal pronouns, is concerned with the orientation to identity of participants in the communicative situation. In order to investigate the identity orientation of family members, it is interesting to examine WE. There were 143 occurrences of the WE lemma in the family corpus, and in order to categorise how WE was being used, each of the 143 occurrences were tagged pragmatically as either referring inclusively to the family itself (inclusive WE) or to some other community external to the family to which the family member speaking belonged (exclusive WE). Thus tagged, it was possible to generate quantitative information on this functional difference in the use of WE. Inclusive WE was found to be notably more frequent than exclusive WE, accounting for 88% of the occurrences. This, it can be argued, indicates that this family primarily utilise WE to create a perspective of *I*, the speaker + *you*, the addressee(s) in the immediate context. This use of inclusive WE is evident from the following extract (1) from the family corpus. The siblings are in the living room discussing the origins of the name of their dog, Goldie:

(1)

<Son 1>	But Goldie's a girl's name like.
<Daughter 1>	Yeah b= **we** didn't give her the name.
<Son 1>	What?
<Daughter>	<$O> **We** didn't give her the name <\$O>.
<Son 2>	<$O> **We** didn't give her the name <\$O>. Although she was so young she wouldn't notice it.
<Son 1>	She wouldn't have a clue shur.
<Son 2>	**We** could've changed it. **We** could call her am Alex.
<Son 1>	Shit for brains.
<Daughter>	Alex.

Earlier in the conversation, son 1 has been complaining about the dog's name, and suggesting different names for her. The other siblings use *we* (marked in bold) in the repeated utterance *We didn't give her the name* as a form of 'safety in numbers' defence to deflect the criticism of the dog's name from themselves. Mühlhäusler and Harré (1990: 174) claim that in this integrative use of *we*, 'the social bonding aspect and the establishment of solidarity is of importance.' The siblings create an in-group, 'we the family', in opposition to the person who originally named the dog. Further to this, son 2 adds *We could've changed* *it. We could call her am Alex,* reaffirms this solidarity by invoking the power that the family had, and still have, to change the name of the dog should they choose to do so. In contrast, exclusive WE accounts for just 17 of 143 instances in the family corpus, or 12% of instances. Exclusive WE in the family corpus refers to a range of out-groups (marked in bold and underlined to the left) and these are illustrated in extracts (2) – (5):

(2)

Friends

<Son> Yeah but the= or they often say members and regulars. But a bouncer would just turn around to you if you said anything like that and go they're members.

<Daughter> Mm. Because one night **we** were goin right and **we** got stopped. Another two got in in front of us and **we** said what oh they're gold cards.

(3)

Workplace

<Daughter> **We** have them outside too the eighty mini bulbs. Is that what they are? Eighty mini bulbs <$G3> yeah **we**'ve them too.

(4)

University

<Son> Are you doin corpus stuff?

<Daughter> Ah **we** hit at it last semester like.

(5)

Limerick

<Son> +aren't **we** already twinned with Quimper?

<Daughter> It's in France.

<Son> Yeah.

Exclusive WE demonstrates that the family, in addition to identifying themselves as members of their family community, also identify themselves as members of a wider Irish society. This finding may indicate the nature of the different identities around which members of the families can construct their reference system. The family in this study have several 'pivots', around which to organise reference such as other communities to which they belong, for example, family, friends, the workplace or education. By invoking inclusive WE, the family is simultaneously defining its identity. The fact that the members of the family are involved in 'we' identities external to the family indicates interaction with a broader society. In findings reported elsewhere (see Clancy, 2011a, 2011b), where family discourse representing a different ethnic and socio-economic grouping in the Republic of Ireland was compared with the family discourse described above, the use or non-use of pragmatic items has been shown to have implications in terms of access or non-access to the dominant culture in Irish society.

3.3 Workplace discourse: The indexical ground of WE

Moving now to a very different speech context, the workplace. WE is again tagged pragmatically in terms of reference. This time, relying on a distinction such as 'inclusive' and 'exclusive' WE does not cover the plethora of referents within the discourse. While it is absolutely true to say that WE operates inclusively and exclusively, there are multiple inclusive and exclusive WE identities indexed, and therefore further classification and categorisation was necessary. This was done in order to trace the interactional 'footing' (Goffman, 1979) displayed by participants, get at their various roles in the discourse (Wortham, 1996), and thus delineate the 'participant framework' (Goffman, 1979). As Wortham (1996: 332) points out, 'acculturated individuals' come to expect standard participation frameworks in given situations, and this, obviously, has resonance in terms of understanding how the individuals in the workplace operate as a community. Borthen (2010: 1809), in quite a different study admittedly, has noted that '...the capacity of human beings to pragmatically enrich utterances with a seemingly sparse semantics should not be underestimated' and this certainly holds true for this data set. Bargiela-Chiappini and Harris observe a very instrumental function of *we* in the institutional context which makes it an interesting item to study: 'in a professional business setting, negotiating between "I" as an individual and some form of collective identity "we" is an everyday matter involving tactical choices, whether conscious or unconscious" (1997: 175). As part of a more general exploration of pronominal reference, the referents contained in the lemma WE were investigated in context. While multiple referents were identified, it was possible to apply a generic taxonomy for quantitative purposes, and distinguish and tag the following referents in WE:

1) Professional ([PROF]): WE as professionals, for example, in the classroom with our students; this use of WE related specifically to language teaching and its practices;
2) Departmental/Subgroup ([DEPT]/[SUB]): A local, situated WE ([DEPT]) which referred to the group of teachers in the department as part of, or as distinct from, the university as an institution. This superordinate WE [DEPT] was subdivided ([SUB]) where teachers referred to themselves in

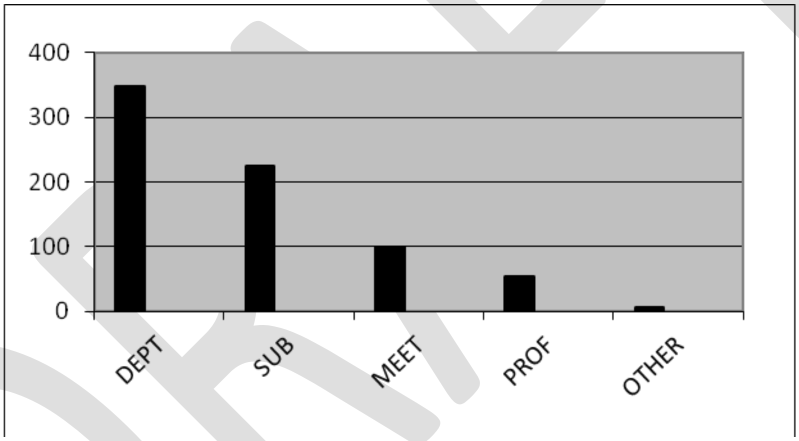
relation to particular subgroups they were part of, such as subgroups teaching different proficiency levels, or working groups set up for other purposes;

3) Procedural ([MEET]): A procedural use emerging from the speech situation itself, the meeting, and referring to everyone in the room at that point in time as a participant in the meeting;

4) Other ([OTHER]): The 'other' category held occurrences such as fixed phrases, e.g., *a bit of both as we say in Ireland*, which in this case also indexes an exclusive use of WE which refers to a national grouping not shared by all of the speakers present.

By tagging the reference of WE, it was possible to generate some data about how frequently each identity was indexed, and, on the basis of these results, note patterns and problematise why these patterns might occur.

Figure 3.2: WE by reference in context



An interesting pattern here is for WE to refer primarily to the institutional context – to the fact that the teachers are members of a department or engaged in a meeting – rather than the broader professional context of the enterprise they are engaged in (being English language teachers). As other personal pronouns, including YOU, were also analysed, a potential explanation for this can also be offered. A similar process of reference retrieval was conducted for the lemma YOU (this lemma includes singular and plural reference, but for the purpose of the study reported here excluded fixed pragmatic clusters such as *you know* and *you see* which were analysed separately). Obviously, singular and plural entity reference was implicated, as well as specific addressee and generic reference. The generic use was interesting for its strong tendency to signal a generic, impersonal reference to the professional/teacher, e.g. in the classroom with students. In other words, in a way that mirrored the [PROF] category of WE, but is somehow qualitatively different in context. This extract (6) from the C-MELT corpus gives a brief view of the professional YOU in context:

(6)

[Kate is reporting on a pilot course she taught on the previous semester, specifically how she put together a syllabus for the class in the absence of a specific textbook]

Kate: But em in that in that kind of respect there was no focus. So the classes developed according to what the students wanted to do and what they needed to do and what as the classes went by what **you [PROF]** could perceive that they needed to do and what they asked for themselves. Basically so they they the course kind of grew as opposed to was there initially.

We can argue that by invoking the YOU [PROF] reference here Kate is inviting engagement and alignment with her process in teaching this class, and suggesting that any professional would recognise this type of organic, responsive syllabus planning for a pilot course. In a broader sense, this YOU can also more safely stake out and reference professional common ground where WE might be slightly more face-threatening. The use of YOU facilitates both an invitation to render subjective judgements shared knowledge, but also, crucially in this case, a way of providing a sufficiently distant ‘professional footing’ (Vaughan, 2009). Additionally, the speaker may be pre-emptively staving off any criticism which might be made of not starting the course with a pre-determined syllabus. The potential of WE and YOU (and, indeed, all the other personal pronouns) to do this sort of complex pragmatic work makes them a rich area for investigation.

We contend that small, domain-specific corpora provide a rich resource for investigating the pragmatic systems of different communities in detail, and here a corpus-based investigation revealed the high frequency of personal pronouns in general. Our broader focus on the idea of ‘community’ and ‘identity’, with the attendant questions about how these are manifested linguistically, led us to a focus on isolating and categorising instances of *we* as (arguably) a linguistic proxy for both. What is striking is how the complexity of reference in a potentially loaded item such as *we* can resolve itself when investigated in context. This reflects the canonical concerns of pragmatics as a discipline, and through the use of a small corpus that can be tagged, in this case in terms of sphere of reference, the pragmatic nuances of how an item can be explored quantitatively as well. The broad framework used – that of identifying and isolating an ‘inclusive’ and ‘exclusive’ WE – held across both corpora, though required more elaboration in relation to the workplace context, which raises the question of why this may be the case. We suggest that this is related to the nature of the communities: the family community’s use of WE operates to define itself and its own identity, through explicitly identifying the out-groups that contrast to the core in-group.

In the case of the workplace community, the pragmatic work that WE does becomes ever more complicated: in a context where the members of the community do not share the same closeness as the family, WE is required not only as an expression of the community's identity, defining its in-groups and out-groups (and hence the parameters of the community), but must also to perform more complex functions in relation to politeness. Clancy (2011b) has demonstrated how the family represents a kind of politeness 'ground zero' (after Levinson 2004). The findings for the case studies reported above in relation to WE would appear to bear this out, showing how the referential potential of a single item is complex within the family itself, and how this complexity multiplies in another, different, context, the workplace.

4. Summary and Conclusions

In the two case studies summarised above, we have focussed on WE in relation to its intriguing 'complexity with regard to personal, social and other deixis' (Mühlhäusler and Harré, 1990: 47), and a number of points can be made now by way of summing up our observations, and underlining the case for using small corpora for investigating pragmatic phenomena. Firstly, although there are various conceptions of what 'small' might mean, as we have shown, many small corpora successfully used in the analysis of pragmatic features appear to be in the 20,000-50,000 word range. Their stance on what constitutes 'balance' and/or 'representativeness' differ from how these were traditionally perceived; however, these corpora can be argued to be both balanced and representative in terms of the speech situations they are designed to characterise. As the literature reported above illustrates, small corpora are eminently suitable for investigating phenomena in context given the constant interpretative dialectic between features of texts and the contexts in which they are produced. Another benefit of using small corpora to do empirical pragmatic research is that the results produced are manageable. In the two case studies reported, it was feasible to isolate each instance of the feature under investigation and assign it a pragmatic tag, which was in turn used to generate quantitative results. This is possible because the small corpus researchers had access to comprehensive metadata and other background knowledge of the context.

That is not to say that corpus-based research in pragmatics is without its difficulties. It is relatively straightforward to search a corpus for an item with pragmatic potential if we can connect that potential with a linguistic form or forms (as in section 3 we connect personal pronouns and the pragmatic management of identity within communities). Research has shown that investigating speech acts, such as apologies, thanks or requests, is a more difficult process. As Archer *et al.* (2012) point out, a weakness associated with using corpora for speech act research lies in the difficulties in automatically retrieving all the linguistic manifestations of a particular speech act, or identifying an appropriate ‘lexical hook’, to use Rühlemann’s phrase (2010: 290), for extracting quantitative information. As Jautz (2008: 147) observes in relation to the speech act of thanking ‘it is difficult...to investigate phenomena above the level of the word or phrase in corpora…since corpora are not (yet) tagged for speech acts, it is not possible to search for all instances of gratitude in a speech act theoretical sense.’ To an extent, these reported limitations can be mitigated by using a small corpus: a speech situation in which these acts are likely to occur can be identified, data collected, and a corpus it is possible to manually tag compiled. In fact, given that pragmatic phenomena are extremely context-sensitive and occasionally completely resistant to automatic retrieval, we should accept that larger corpora are simply not suitable for some of our purposes, despite the desirability of adding an empirical slant to pragmatic research. The middle ground lies in the design and exploitation of small corpora for pragmatic research.

References

Archer, D., K. Aijmer and A. Wichmann, 2012. *Pragmatics: An Advanced Resource Book for Students*. London: Routledge.

Aston, G., 1997. Large and small corpora in language learning. In: B. Lewandowska-Tomaszczyk and P.J. Melia (eds.), *PALC97: Practical Applications in Language Corpora*. Łodz: Łodz University Press, 51-62.

Atkins, S., J. Clear and N. Ostler, 1992. ‘Corpus design criteria.’ *Literary and Linguistic Computing*, 7(1), 1-16.

Bargiela-Chiappini, F. and S. Harris. 1997. *Managing Language: The Discourse of Corporate Meetings.* Amsterdam: John Benjamins.

Biber, D., 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Vaughan, E. & B. Clancy. 'Small corpora and pragmatics,' *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

Biber, D., 1993. 'Representativeness in Corpus Design.' *Literary and Linguistic Computing*, 8(4), 243-257.

Biber, D., S. Conrad and R. Reppen, 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan, 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Borthen, K., 2010. 'On how we interpret plural pronouns.' *Journal of Pragmatics*, 42(7), 1799-1815.

Clancy, B., 2010. Building a corpus to represent a variety of language. In: A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 80-92.

Clancy, B., 2011a. 'Complementary perspectives on hedging behaviour in family discourse: the analytical synergy of variational pragmatics and corpus linguistics.' *International Journal of Corpus Linguistics*, 16(3), 371-390.

Clancy, B., 2011b. *Do you want to do it yourself like?* Hedging in Irish traveller and settled family discourse. In: B. Davies, M. Haugh and A. Merrison (eds.), *Situated Politeness*. London: Continuum, 129-146.

Clancy, B. and E. Vaughan (in press, 2012). *It's lunacy now*: A corpus-based pragmatic analysis of the use of *now* in contemporary Irish English. In: B. Migge and M. Ní Choisáin (eds.), *New Perspectives on Irish English*. Amsterdam: John Benjamins.

Clear, J., 1992. Corpus sampling. In: G. Leitner (ed.), *New Directions in English Language Corpus Methodology*. Berlin: Mouton de Gruyter, 21-31.

Cutting, J., 2001. 'The speech acts of the in-group.' *Journal of Pragmatics*, 33, 1207-1233.

De Fina, A., D. Schiffrin and M. Bamberg (eds.), 2006. *Discourse and Identity*. Cambridge: Cambridge University Press.

Vaughan, E. & B. Clancy. ‘Small corpora and pragmatics,’ *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

Farr, F., 2005. Relational strategies in the discourse of professional performance review in an Irish academic environment: The case of language teacher education. In: A. Barron and K. Schneider (eds.), *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter, 203-234.

Farr, F., 2007. Spoken language analysis as an aid to reflective practice in language teacher education: Using a specialised corpus to establish a genetic fingerprint. In: M.C. Campoy and M. J. Luzón (eds.), *Spoken Corpora in Applied Linguistics*, Bern: Peter Lang, 235-258.

Fasulo, A. and C. Zucchermaglio, 2002. ‘My selves and I: Identity markers in work meeting talk.’ *Journal of Pragmatics*, 34(9), 1119-1144.

Flowerdew, L., 2002. Corpus-based analyses in EAP. In: J. Flowerdew (ed.), *Academic Discourse*. London: Longman, 95-114.

Flowerdew, L., 2004. The argument for using English specialised corpora to understand academic and professional settings. In: U. Connor and T. Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, 11-33.

Goffman, E., 1979. ‘Footing.’ *Semiotica*, 25, 1-29.

Goffman, E., 1981. *Forms of Talk*. Oxford: Blackwell.

Hanks, W., 1992. The indexical ground of deictic reference. In: A. Duranti and C. Goodwin (eds.), *Rethinking Context. Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press, 43-77.

Holmes, J., 1988. ‘Doubt and certainty in ESL textbooks.’ *Applied Linguistics*, 9(1), 21-44.

Knight, D., Evans, D., Carter, R. and Adolphs, S. (2009) ‘HeadTalk, HandTalk and the corpus: Towards a framework for multi-modal, multi-media corpus development.’ *Corpora* 4(1), 1-32.

Koester, A., 2006. *Investigating Workplace Discourse*, London: Routledge.

Vaughan, E. & B. Clancy. ‘Small corpora and pragmatics,’ *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

Koester, A., 2010. Building small specialised corpora. In: A. O’Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 66-79.

Lave, J. and E. Wenger, 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Levinson, S., 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Levinson, S., 2004. Deixis. In: L. Horn and G. Ward (eds.), *The Handbook of Pragmatics*, Oxford: Blackwell, 97-121.

Loudermilk, B.C. 2007. ‘Occluded academic genres: An analysis of the MBA thought essay.’ *English for Academic Purposes*, 6(3), 190-205.

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

McCarthy, M. and A. O’Keeffe, 2010. Historical perspective: What are corpora and how have they eveolved? In: A. O’Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3-13.

McEnery, T., R. Xiao and Y. Tono, 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.

Mühlhäusler, P. and R. Harré, 1990. *Pronouns and People: The Linguistic Construction of Social and Personal Identity*. Oxford: Blackwell.

O’Connor, P., 1994. ‘“You could feel it through the skin”: Agency and positioning in prisoners’ stabbing stories’, Text, 14(1), 45-75.

O’Keeffe, A., 2005. You’ve a daughter yourself? A corpus-based look at question forms in an Irish radio phone-in. In: A. Barron and K. Schneider (eds.), *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter, 339-366.

Vaughan, E. & B. Clancy. ‘Small corpora and pragmatics,’ *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

O’Keeffe, A. and S. Adolphs, 2008. Response tokens in British and Irish discourse: Corpus, context and variational pragmatics. In: K. Schneider and A. Barron (eds.), *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages*. Amsterdam: John Benjamins, 69-98.

Orpin, D., 2005. ‘Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze.’ *International Journal of Corpus Linguistics*, 10(1), 37-61.

Patrick, P., 2002. The speech community. In: J.K. Chambers, P. Trudgill and N. Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. Oxford: Blackwell, 573-597.

Pennycook, A., 1994. ‘The politics of pronouns.’ *ELT Journal*, 48(2), 173-178.

Rees, A., 1983. Pronouns of Person and Power: A Study of Personal Pronouns in Public Discourse. Unpublished MA dissertation, Sheffield University.

Romero-Trillo, J. (ed.), 2008. *Corpus Linguistics and Pragmatics: A mutualistic entente*. Berlin: Mouton de Gruyter.

Rühlemann, C., 2010. What can a corpus tell us about pragmatics? In: A. O’Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 288-301.

Rundell, M., 2008. ‘The corpus revolution revisited.’ *English Today*, 24(1), 23-27.

Sankoff, D., 1988. Problems of representativeness. In: U. Ammon, N. Dittmar and K. Mattheier (eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*. Berlin: Walter de Gruyter, 899-903.

Sinclair, J.M., 2001. Preface. In: M. Ghadessy, A. Henry and R.L. Roseberry (eds.), *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins, vii-xv.

Vaughan, E. & B. Clancy. ‘Small corpora and pragmatics,’ *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

Sinclair, J., 2005. Corpus and text: Basic principles. In: M. Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford: Oxbow Books, 1-16, available online at http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm [date accessed 25-06-2012].

Sinclair, J., 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stirling, L. and L. Manderson, 2011. ‘About you: Empathy, objectivity and authority.’ *Journal of Pragmatics*, 43(6), 1581-1602.

Stubbs, M., 2004. Language Corpora. In: A. Davies and C. Elder (eds.), *The Handbook of Applied Linguistics*. Oxford: Blackwell, 106-132.

Swales, J.M. 1990. *Genre Analysis: English and Academic Research Settings*. Cambridge: Cambridge University Press.

Swales, J., 1996. Occluded genres in the academy: The case of the submission letter. In: E. Ventola and A. Mauranen (eds.), *Academic Writing: Intercultural and textual issues*. Amsterdam: John Benjamins, 45-58.

Tagliamonte, S., 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.

Tannen, D., 2007. Talking the dog: Framing pets as interactional resources in family discourse. In: D. Tannen, S. Kendall and C. Gordon, (eds.), *Family Talk: Discourse and Identity in Four American Families*. New York: Oxford University Press, 49-70.

Thomas, J., 1983. ‘Cross-cultural pragmatic failure.’ *Applied Linguistics*, 4(2), 91-112.

Tognini-Bonelli, E., 2010. Theoretical overview of the evolution of corpus linguistics. In: A. O’Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 14-27.

Torgersen, E.N., C. Gabrielatos, S. Hoffmann and S. Fox, 2011. ‘A corpus-based study of pragmatic markers in London English.’ *Corpus Linguistics and Linguistic Theory*, 7(1), 93-118.

Vaughan, E. & B. Clancy. ‘Small corpora and pragmatics,’ *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

Tracy, K. 2002. *Everyday Talk: Building and Reflecting Identities*. New York: Guilford.

Vaughan, E., 2007. ‘*I think we should just accept…our horrible lowly status*: Analysing teacher-teacher talk within the context of community of practice.’ *Language Awareness,* 16(3), 173-189.

Vaughan, E., 2008. “Got a date or something?”: An analysis of the role of humour and laughter in the workplace meetings of English language teachers. In: A. Ädel and R. Reppen (eds.), *Corpora and Discourse: The Challenge of Different Settings*, Amsterdam: John Benjamins, 95-115.

Vaughan, E., 2009. *Just say something and we can all argue then*: Community and identity in the workplace talk of English language teachers. Unpublished PhD thesis, University of Limerick.

Vaughan, E. and B. Clancy, 2011. ‘The pragmatics of Irish English.’ *English Today*, 27(2), 47-52.

Wenger, E., 1998. *Communities of Practice. Learning, Meaning and Identity*. Cambridge: Cambridge University Press.

Whitley, M.S., 1978. ‘Person and number in the use of WE, YOU, and THEY.’ *American Speech*, 53(1), 18-39.

Wortham, S., 1996. ‘Mapping participant deictics: A technique for discovering speakers’ footing.’ *Journal of Pragmatics*, 25(3), 331-348.