The Limerick Corpus of Irish English: design, description, and application

Fiona Farr
University of Limerick

Bróna Murphy
University of Limerick

Anne O'Keeffe
Mary Immaculate College, University of Limerick

Abstract

This paper describes an ongoing corpus development and application project at the Mary Immaculate College and the University of Limerick, Ireland. The Limerick Corpus of Irish English is a one-million word corpus of English as it is spoken in Ireland. The corpus is genre-based and consists primarily of casual conversational data. Details of the corpus design, development, and applications, both research and pedagogic, are described. An illustrative example of the linguistic phenomenon of HEDGING is explored and some classroom activities based on the findings are developed.

Introduction

It is now possible to access many millions of words of data in a matter of seconds. For example, the British National Corpus (BNC) contains 100 million words and the Bank of English comprises over 500 million words of English data (see Sinclair 1997). However, large corpora consist mainly of British and American English in the form of written language. The dearth of spoken data is often attributed to the high cost of data collection and transcription; the quality of recorded data; the labour intensive nature of collecting, transcribing, and coding; and the difficulty of accessing representative speech data, especially in certain contexts (see McCarthy 1998). O'Keeffe and Farr (2003) argue that that such exertion and expenditure is justifiable on the grounds of needing to re-assess language interpretation and pedagogy to account for spoken as well as written

Teanga 21 (2002), pp. 5-29

norms of use, and of needing access to specific or local genres not otherwise available. There are many differences between findings from written versus spoken corpora, and indeed there are many differences *within* spoken corpora depending on the context and variety (as this paper will illustrate below).

New insights from corpus linguistics have vastly improved our dictionaries (see Fox 1998) and grammars (see Biber et al. 1999). Also, some applied linguists have used corpora to enhance our understanding of fixed expressions, collocations, and extended language patterning (see for example Sinclair 1991, Svartvik 1991, and Aston 1995). In addition, many studies have highlighted the discrepancy between the language that is presented in textbooks and pedagogic references, and the language we actually use (Holmes 1988, Baynham 1991, Boxer and Pickering 1995, Kettermann 1995, Baynham 1996, Carter 1998, Hughes and McCarthy 1998, and McCarthy 1998).

For those of us attempting to bring corpus data into the classroom, it is important to accurately represent language in use, and particularly, it is important to make a distinction between written and spoken data. Too often our classroom descriptions of the English language are based on written norms alone. Additionally, we must beware of overgeneralizations and attempt to be sensitive to specific genres and registers which are the primary determiners of language use (Biber et al. 1999 and Conrad 2000). In this paper we will describe a small spoken corpus of Irish English, the Limerick Corpus of Irish English (hereafter L-CIE), and we hope to illustrate some of its applications in advancing our description and understanding of spoken language in its context of use. We will also show how some of these data can be exploited in the language classroom in the teaching of spoken language.

Corpus design and description

While many people see a corpus as 'a helluva lot of text stored on a computer' (Leech 1992: 106), the impact of the tremendous growth in interest and activity in the area of corpus building and analysis over the past number of years has attempted to overturn this view in a bid to demonstrate that corpora are much more than just randomly selected disorganized collections of text. According to Sinclair (1995: 20), a corpus is not a random collection of texts but, in fact,

a collection of pieces of language that are selected and ordered to explicit linguistic criteria to be later used as a sample of the language in question. Nelson Francis, who co-compiled one of the earliest and most widely used computerized corpora created at Brown University defines a corpus as 'a collection of texts assumed to be representative of a given language, dialect or other subset of a language, to be used for linguistic analysis' (1982: 7), while Atkins et al. (1992: 1) understand a corpus to be a subset of an Electronic Text Library (ELT) built according to explicit design criteria for a specific purpose.

All three perspectives above are similar in defining a corpus as a collection of language, and they also concur that the design of a corpus must adhere to strict linguistic criteria if its aim is to demonstrate a real sample of language in use. This is an area that has been given much attention in recent times and it would seem from the relevant literature that the three main criteria respected in the design of the modern corpus are: (1) authenticity of the texts, (2) representativeness of language included in the corpus and (3) sampling criteria used in the selection of texts (Tognini-Bonelli 2001: 54). These issues have been discussed at great length by many corpus linguists such as Atkins et al. (1992), Biber (1993), and Crowdy (1994), and are vital in the design and compilation of corpora. They will now be discussed briefly in order to make explicit the importance of the design in the creation of corpora if they are to be regarded as accurate representations of the language, in this case spoken English.

Spoken corpora

Despite the practical difficulties of transcribing large quantities of spoken data, a number of spoken corpora have been designed and compiled to show language in use according to different principles, depending on various research foci: some represent language varieties (e.g. British, American) while others represent language genres (e.g. academic lectures, radio broadcasts) and so on (see Appendix 1 for a list of spoken corpora and their URLs). This section briefly lists and describes the more significant spoken corpora which have been constructed.

As an example of a variety-based corpus, the British National Corpus (BNC) demonstrates a snapshot of British English (see

Crowdy 1994). The corpus contains both written and spoken data. The spoken component makes up 10 million words and consists of unscripted informal conversation which was recorded by volunteers selected from different ages, regions, and social classes in a demographically balanced way.

In the case of genre-inspired projects, the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) is another relatively large spoken corpus consisting of five million words of data, mostly from Great Britain with a small amount of Irish English. It is designed to represent spoken language in different contexts of use, genres of speech, and different speaker relationships across the islands of Britain and Ireland (see McCarthy 1998). The Limerick Corpus of Irish English described below has been constructed to parallel CANCODE.

The Limerick Corpus of Irish English: design and description

L-CIE is a one million word spoken corpus of Irish English whose ongoing compilation at the University of Limerick began in the academic year 2002-03.[1] Of core concern to this project is the collection of naturally-occurring spoken data from everyday Irish contexts so as to assemble a corpus that will allow for the description of Irish English on its own terms rather than solely focusing on the extent to which it resembles or differs from other varieties of spoken English (for example, British or American English).

There were three main stages in the building of L-CIE. The first stage involved the identification of the participants in various relationships, speech genres, and settings. The second stage involved the collection and transcription of data as well as the databasing of tapes and logging of speaker information. The third stage examined the transcriptions and involved the anonymization of the transcribed data.

L-CIE currently includes conversations recorded across a wide variety of predominantly informal settings throughout Ireland (excluding Northern Ireland). The conversations were carefully collected with reference to a range of different speech genres and with an overall emphasis on casual conversation. Although 30,000 words of professional, transactional, and pedagogic Irish English are included, as Table 2 illustrates, 82% of the data is casual conversation. In the absence of any comprehensive taxonomy of spoken

genres, as discussed above, corpus compilers have to make careful decisions regarding what to include in a representative corpus. The framework adapted for L-CIE is based on the CANCODE matrix described in McCarthy (1998) and includes two axes for classification, the context type and the interaction type (see Tables 1 and 2). The interaction types outlined in Table 1 and the vertical axis of categorization on Table 3 reflect the relationship between the participants in the dyadic and multi-party conversations in the corpus. These relationships fall into five broad categories: intimate, socializing, professional, transactional, and pedagogic (see McCarthy 1998). The corpus design focuses on representing a variety of discourse contexts and speech genres across different speaker relationships with the aim of informing research and pedagogy into the fields of grammar, lexis, and discourse. Table 1 defines the five relationship types found in L-CIE.

Table 1: Interactional relationship types

Relationship	Description
Pedagogic	Pedagogic relationships are those set in contexts such as the classroom, tutorials, and lectures.
Transactional	In this relationship category, speakers do not previously know one another. Interactions usually relate to a need on the part of the speaker or hearer. The aim of the conversation is to fulfil a transactional goal.
Professional	This relationship holds between people who are interacting as part of their daily work. This only applies to interactions where all speakers are part of the professional context. Talk that is not work-related, but occurs between colleagues in work places is still classified as professional.
Socializing	This is closely related to 'intimate' and implies the voluntary interaction between speakers that seek each other's company for the sake of interaction. This relationship is usually marked by friendship and is not as close as 'intimate'. Typical venues include birthday parties and social gatherings.
Intimate	This type of relationship can be defined by the minimal distance between speakers, and often involves cohabitation. It includes partners and close family and friends.

Apart from the context-type categories described above, the corpus distinguished between talk that was predominantly collaborative and that which was non-collaborative. Within the collaborative talk, compilers distinguished between collaborative ideas (e.g. exchanging opinions) and collaborative tasks (engagement in some physical task, e.g. doing the washing up), whereas the non-collaborative texts were referred to as information provision.

Table 2: Context type categories

Collaborative Idea	Conversations in which the main goal is to exchange ideas.
Collaborative Task	Social activities which cannot be performed with language alone.
Information Provision	Explanations and information with the conversations being marked by unilinear transfer of information.

Table 3 below specifically presents the L-CIE matrix based on the classification axes of the interaction type and the context type as just described. The interaction type can be seen on the vertical axis and the context type on the horizontal axis. For each cell in Table 3, we provide an example of the type of data it contains.

Having a corpus design which is sensitive to context, speaker relationship, and speech genre allows for more refined descriptions of spoken language. In particular, it facilitates insights into how conversational conditions influence the frequency of particular discourse features. As we will illustrate later, the distribution profile of the discourse feature of 'hedging' is extremely dependent on the context of use. Where the speaker relationship is distant and the setting is institutional, the frequency of hedging items is at its highest. This kind of information advances our empirical understanding of spoken language in use and also provides us with insights which have classroom application. More importantly for us, it does so in an Irish context.

Table 3: The L-CIE data matrix with samples of data types

	% of data	Information-provision	Collaborative idea	Collaborative task
Pedagogic	7	Linguistics lecture	English poetry tutorial	One-to-one computer lesson
Professional	3	Real estate office talk	Team meeting	Waitress washing dishes
Socializing	8	Describing a new bar	Friends discussing college	Friends assembling a bed
Intimate	3	Mother storytelling	Partners making holiday plans	Family preparing dinner
Transactional	79	Product presentation	Chatting in a taxi	Eye examination

The Limerick Corpus of Irish English: pedagogic applications

From the outset, the rationale behind L-CIE was to provide a body of spoken Irish English data to be used by researchers, lecturers, and students in English Language Teaching, including English as a Foreign Language (EFL) and Teaching English as a Foreign Language (TEFL) sections of the University of Limerick and its affiliated Mary Immaculate College. The research applications of such a corpus are numerous and it has been partially built and used by those involved in final year projects, MA dissertations and theses, and PhD theses. It also serves as a useful reference corpus for those involved variously in small-scale research projects. However, the focus of this article is on pedagogic applications. To this end it is used as a general lexico-grammatical reference to inform teaching at both undergraduate and postgraduate level; some subcorpora of classroom discourse can be isolated to aid the development of pedagogic skills among trainee teachers (see O'Keeffe and Farr 2003 for details). Of course, as already mentioned, it is a primary source for the study of spoken language and specific varieties and genres, and it is these we have chosen as the focus for an illustrative example which we now describe in detail. The topic we investigate is the nature and teaching of casual conversational features in the context of Irish English. Firstly we provide a short, corpus-based

description of the general features of casual conversation. Next we hone in on and investigate one of these interactional features, namely hedging, across a selection of genres. Finally, we furnish a simple example of how these findings can be used in language learning materials for the EFL/ELT classroom.

Casual conversation features

Based on the analysis of extracts from several informal interactions between families and friends in L-CIE, the following features of casual conversational discourse were identified.

Topic management. Speakers smoothly introduce, extend, redirect, and close topics during interactions, and usually in a social and collaborative way which gives the impression of ease of construction. There is extensive use of deixis and pronominal reference devices, illustrating the high context-dependency of face-to-face casual speech. The following example comes from a conversation between two friends driving together and commenting on a house they see on their route.

(1) Female friend: *It*'s very small isn't *it*.

Such deictic items generally do not pose huge difficulties for learners of EFL as the reference point is explicit somewhere in the text. On the other hand, exophoric or context-bound references can be problematic for two reasons. They usually involve a certain cultural understanding (McCarthy 1991) and they are often encodings of shared cultural, personal, and intimate knowledge between the speakers. It is also interesting to see how the conversation moves from one speaker to the next as they co-construct the text.

(2) Female friend 1: No em, Sean O'Donoghue, no Sean+
Female friend 2: No it's, no, Sean and Sile.
Female friend 1: Yeah, it's their daughter's.

This is done at the conceptual and linguistic processing levels and can be easily identified by students when examining conversational extracts. In fact, it is likely that similar processes exist in their first languages but with slightly different conventions applying, and they can very easily identify these in an inductive way that will help

raise their awareness of what it means to be an active conversational participant.

Speaker turns and engaged listenership. Many discourse and conversational analysts, both descriptive and applied (Sinclair and Coulthard 1975, Schegloff et al. 1977, McCarthy 1991, Eggins and Slade 1997, McCarthy 1998, and others), have given due care and attention to the structure of spoken interaction and how talk moves smoothly from one speaker to the next and back again in a very seamless way. Speakers do not necessarily wait for the previous speaker to finish, they complete each other's utterances, and they intervene with appropriate responses to indicate that they are listening without attempting to take the conversational floor (Yngve 1970). The term LISTENERSHIP may be used to refer to this process whereby participants encode their (temporal) identity as listener rather than 'current speaker'. Responses which mark engaged listenership include non-verbal items such as head nods and verbalizations such as *mmm*, *yeah, oh really, you must be joking, what a pity!*, etc.

Lexico-grammar. It is a well-known fact that casual conversation does not always mirror the structural and grammatical conventions that written discourse has come to be associated with, nor does it need to. It is full of false starts, so-called incomplete sentences, and ellipted words and phrases. A number of these features are found in L-CIE. We find genuine performance slips, for example:

(3) Female friend: …who won't give me hugs when I was younger.

Also present are cases of regional language use, for example,

(4) Female friend: …herself and the husband.

This is a structure coming directly from the Irish equivalent *Í féin agus a fear* ('herself and her husband'). Ellipsis can be seen in (5) where a family is decorating a Christmas tree and one family member is offering to pass another decoration to his siblings. In its context, this utterance is pragmatically complete (and successful) without the Auxiliary + Subject *Do you*:

(5) Brother: Want another one?

Vague language is omnipresent, for example,

(6) Female friend: What have you been up to?
Male friend: Oh, this and that.

While some features only need to be recognizable to EFL students, the more typical and generic features (e.g. ellipsis and vagueness) should be explicitly observed and ideally incorporated into their active language use if they are to function in an efficient and natural way in casual conversation.

Relational talk. Casual conversation is usually part of a face-to-face encounter and therefore one of its very important, and at times exclusive, functions is simply to build and maintain good relations with the person with whom we are speaking. In L-CIE, this function is achieved through skilful negotiation of topics, appropriate use of phatic communion, questioning strategies, hedging, hesitations and restarts, complimenting, and explicit socio-cultural bonding techniques such as the use of Gaelic words or Irishisms, all of which go a long way towards meeting the relational requirements of a conversation. The quantity and quality of relational talk demanded in a conversation has a direct link with the context-related variables of the interaction. The feature of hedging will be discussed in detail below as an illustrative example.

Genre and appropriacy. Eggins and Slade (1997: 265) identified the following typical genres which occurred in their corpora of casual conversation: narratives, anecdotes, recounts, exempla, observation/comment, opinion, gossip, joke-telling, and friendly ridicule (sending up), and from L-CIE we can add to that chat, or small talk (Coupland 2000). Whichever framework we adopt there is no denying that orientation towards genres is an essential feature of any type of discourse. The difficulty for students in terms of genres is not perhaps in recognizing and appropriating them but in doing this at the challenging pace at which spoken casual language is produced. Native speakers shift from one genre to the next and back again without even thinking about it, and often certain genres are lost to the EFL learner (McCarthy 1991:138) especially irony, ridicule, and humour in casual genres such as gossip and banter. Tropes often associated with literary contexts (e.g. hyperbole; see Carter and McCarthy 2004) are part of the natural fabric of everyday talk.

Hedging across genres

One of the most pervasive features of spoken language is the use of softeners or hedges. They can be used to downtone or mitigate the force of an utterance for various reasons, e.g. politeness, indirectness, vagueness, and understatement. Hedges take many forms, most salient of which is the use of core modal verbs (*will, shall, should, would, can, could, might, may, must*) and clausal items (*I mean, I think, I suppose, you know*, etc.). Other common forms include noun-based expressions (*there is a possibility; the thing is*), degree adverbs (*quite, really, relatively, necessarily*, etc.), restrictive adverbs (*just, only*, and so on), and stance adverbs (for instance, *of course, actually, kind of, really, sort of, maybe*). Syntactic choices may also be hedged, i.e., where a speaker chooses a structure that is more indirect because of a tense or aspectual choice or the use of a modal downtoner or hypothetical clause. A radio presenter, for example, asks a lawyer: *If you think a case is spurious would you take it?*, as opposed to the more bald unhedged interrogative, *Do you take spurious cases?*. Another syntactic example from L-CIE is the use of double negatives, e.g., *It's not that I'm not afraid*, meaning 'I am afraid'.

To exemplify the context sensitivity of hedging, let us take some of the most frequent hedging items from L-CIE. By conducting a word frequency count for all of L-CIE using *Wordsmith Tools* software (Scott 1999), the 10 most frequent hedging items were isolated (see the left-hand column of Table 4 below). A two-word cluster search was conducted, again using *Wordsmith Tools* (the results are presented on the right-hand column of Table 4). This allowed for the corroboration of findings from the single word frequency search where most of the items are in fact elements of fixed expressions.

These results allowed us to formulate the following aggregate of single-word items and two-word cluster search items for analysis: *just, really, actually, probably, I think, a bit, kind of, sort of, you know, I suppose.* They were analysed across a number of contexts in L-CIE: family conversations, teaching practice feedback, calls to a radio phone-in show, conversations at the counter of a shop, and female friends chatting. Concordance lines were generated using *Wordsmith Tools* (i.e., where the software generates all the occurrences of the search word in its context and presents it in the centre of each line with around seven words at

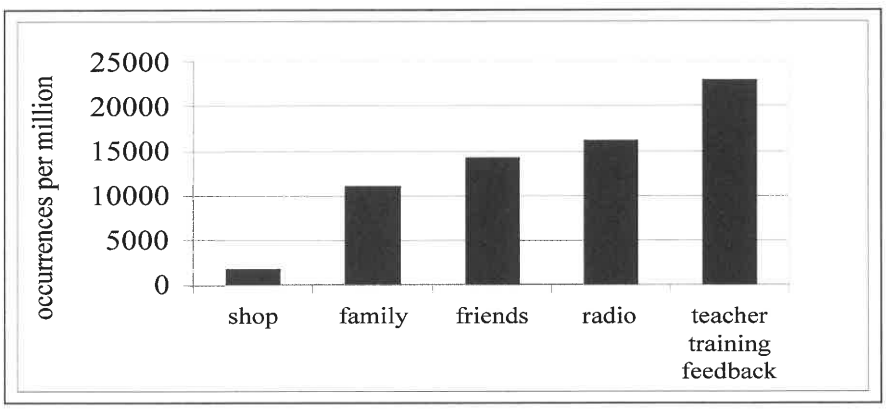
either side), and from these all non-hedging instances were eliminated.[2]

Table 4: Word and cluster frequencies for hedging items in L-CIE

Single-word frequency list		Two-word cluster frequency list	
Rank	*Hedging item*	*Rank*	*Hedging item*
13	like	1	you know
15	know	7	I think
29	just	12	kind of
50	think	62	a bit
83	really	69	I just
91	kind	77	I suppose
97	actually	80	sort of
176	probably		
200	suppose		
204	sort		

Figure 1 shows the cumulative results where total occurrences for all of the search items as hedges are presented.[3]

Figure 1: Frequency distribution of hedging items in L-CIE (per million words)



As we can see from these data, the least amount of hedging was found in shop encounters. This is perhaps explained by the lack of need to protect face in service encounters where a customer and a server do not know each other and when there is an existing social

schema for the interaction within exogenous roles. This allows for an emphasis on transactional efficiency without threat to face. In the extract below, we find an example of the use of *just* as a hedge when the customer asks for a receipt and wishes to lessen the imposition of his request on the server by using a hedge.

(7) Server: Five fifty so please.
Customer: And I'll *just* have a receipt.

However, it is normal and not at all face-threatening to make unhedged requests in stereotypical service encounters, such as in the following encounter where the server and the customer are not known to each other and there are no additional requests (note that *Carrolls* refers to a brand of cigarettes):

(8) Customer: Give us ten Carrolls there please.
Server: Ten Carrolls [pause]. One sixty two.
Customer: Thank you.
Server: Thanks.

As noted in Binchy (2000), there is a greater tendency for politeness markers such as hedging and small talk when the server and the customer are known to each other as in the following extract (note that the symbols = and + mark truncated utterances and interruptions respectively; italic and bold typeface mark hedging and small talk respectively):

(9) Customer: *I'm looking for* em *do you know* the Cadburys Fudge? Do you do it *at all*?
Server: The small little bar?
Customer: They're *kind of* long *little* finger that's what I'm looking for.
Server: Now.
Customer: *Can I have* three of those?
Server: Three now.
Customer: **I love that stuff.**
Server: **The Fudge th=**
Customer: **Yeah it's hard to get+**
Server: **Yes.**
Customer: +you don't see it everywhere you know but I love it.
Server: Sixty six *so Jane please*.

The next least hedged context was found to be in the family settings. Here, as Clancy (2000) notes, the family relationship is fixed and so needs less attention to face as a result (see also Blum-Kulka 1994, 1997a, and 1997b). In these data, we found that imperatives are relatively common but are very often hedged. In the extracts below *just* is used to hedge the imperatives:

(10) a. *Just* stick it on. *Just* tape it on to something or another
b. Cards anyone? *Just* cut them
c. Mike shut up. *Just* shut up.

Casual conversations between friends were the next most hedged dataset in the sample taken from L-CIE. In the example below a group of friends are reminiscing about a teacher they used to have in secondary school.

(11) ...he *kind of* had a thing for one of the girls he was *like kind of* friendly but not sleazy friendly, but he frightened me one day...

The radio phone-in data sampled was found to have a relatively high degree of hedging and this correlates with the speaker relationship, where the participants are not normally known to each other but where (unlike service encounters), the interaction is not just a straightforward transactional affair. In a radio phone-in, face mitigation is very important (see O'Keeffe 2003). As detailed in O'Keeffe (in press), frequently the presenter downtones the force of questions, for example by the use of a redundant reflexive pronoun as a hedge (see Figure 2 below). These appear to serve as downtoners by personalizing the question. When we examine L-CIE as a whole, we find that this feature is commonly used, as illustrated in the sample concordance lines of Figure 3.

The highest instance of hedging was found in the institutional data from the university setting involving post-observation feedback sessions reviewing teacher trainee taught language lessons. Obviously these interactions involve the highest degree of potential face threat. The speaker relationship is asymmetrical yet the power role holder (the teacher trainer) wishes to downtone her power and to seem encouraging to the trainees, while at the same time the trainee wishes to defer to and explicitly acknowledge the power of

the trainer (see Farr and O'Keeffe 2002 and Farr in press), as the following example illustrates:

(12) Trainee: It was it was helpful to see Laura first *actually now* sometimes I watched *like* it di= it didn't help *at all like* it made me feel worse *am* but *am like* watching her helped *I think.*

Figure 2: Concordance line for redundant pronoun yourself *in radio phone-in data (O'Keeffe in press)*

This is because you had side effects	**yourself**?	Umhum yeah.
always are. How are you fixed	**yourself**	for nightclubs and the
This is because you had side effects	**yourself**?	Umhum yeah.
Yeah yes I do you've a daughter	**yourself**?	I have Emm
e involved in the medical profession	**yourself**?	I am yes.
You do this	**yourself**	I take it? I h
And what are you doing with	**yourself**	nowadays? I hav
arian. You were a boarder there	**yourself**?	I was a boarder

Figure 3: Concordance lines of yourself *in the Limerick Corpus of Irish English (O'Keeffe in press)*

r why not. Are you going on holidays	**yourself**	Joe? That's |
This is because you had side effects	**yourself**?	Umhum yeah.
Okay. +do do you see that	**yourself**	that am to "build on
ng to the end now if you had to give	**yourself**	two pieces of advice
mit to it maybe. Have you kids	**yourself**	have you? I
riate? Me= have you learnt languages	**yourself**?	No.
hat something you had thought about	**yourself**	Am yeah but
this and realised	**yourself**	that this is?
And what are you doing with	**yourself**	nowadays?
Do you want to do it	**yourself**	like?

Thus far we have looked at the hedging items as a whole, however further internal variation is evident when we break down the results per item across the contextual cells as Table 5 illustrates.

Table 5: Breakdown of hedging items across contextual cells (occurrences per million words)

ITEM	CONTEXT				
	Phone	Family	Friends	Radio Phone-in	Teacher Training Feedback
really	32	1333	2641	1600	1693
just	274	3250	4769	2509	7986
you know	433	2083	1488	5764	3190
actually	90	1333	1000	1182	2452
probably	45	750	641	382	1194
a bit	256	584	356	374	1,361
I suppose	472	73	356	697	505
I think	90	1500	1615	3582	4405
TOTAL	1692	10906	13334	16090	22786

For example, we see that *just* is 53% more frequent in the conversations between friends than in radio phone-in data, while *you know* is 55% more frequent in radio phone-in interactions compared with teacher training feedback sessions, and so on. As this analysis illustrates, contextual factors have a strong influence over the type of discourse that results. Hedging, as one example of a feature common to spoken language, was found to be 13 times more frequent in an institutionalized interaction than in service encounters and almost double what one would expect to find between friends. Clearly such contextual factors of genre need to be taken into account in a more delicate description of spoken language and its application in pedagogic contexts. When we deconstruct the key contextual components that bring about generic variation, we find the following items need to be considered when analysing empirical data: cultural context, speaker relationships, speaker roles, and register (within which we must consider the goal of the interaction, the type of talk, the audience, and the setting).

Classroom activity

Appendix 2 details a series of tasks that can be used with EFL students or ELT trainee teachers to raise their formal awareness of hedging. These activities are an example of how the findings on

linguistic hedges can be converted into materials suitable for classroom use. As they are presented here they are suitable for upper intermediate to advanced level students. Every other student should initially complete either Task A1 or Task A2 individually and then Tasks B, C, and D in pairs or groups.

Conclusion

Using language corpora, in conjunction with other language learning and teaching materials, can give access to more refined and specific language information. They also provide access to spoken language data from specific varieties, registers, and genres. They can be used in an investigative or more directive way in the classroom, but in all cases judicious editorial decisions on the part of tutors may be required, and mediation of the raw data is usually vital. Although we are a long way from full integration of computerized corpora, we are now beginning to explore the issues associated with corpus use in the classroom, and many methodological and procedural advances have been made in recent years (see O'Keeffe and Farr 2003). For us, building the Limerick Corpus of Irish English has provided a local context within which to conduct our empirical research and on which to base our classroom materials. Most of all, it has brought to light the fact that while there are differences between Irish English and other varieties, internal variation across contexts of use is common to all varieties and is perhaps a more relevant preoccupation for pedagogy than an exclusive focus on dialectal features.

Notes

1. Funding for the compilation and electronic packaging of L-CIE came from the Higher Education Authority Targeted Initiatives Programme.

2. Note that the hedge *like* which is the highest ranking word in the frequency list has been omitted from this analysis for practical purposes. Due to its multifunctionality as a non-hedging verb, a hedge, and an interactional marker, it would was not plausible to manually sort this item in this amount of data.

3. As is standard practice, all results presented here have been 'normalized', that is, they have been converted to occurrences per million words so as to assist comparability.

Appendix 1: Spoken corpora websites

— Limerick Corpus of Irish English (L-CIE)
www.ul.ie/~lcie/homepage.htm
— The Wellington Spoken Corpus of New Zealand English (WSC).
www.vuw.ac.nz/lals/wgtn_crps_spkn_NZE.htm
— British National Corpus
http://info.ox.ac.uk/bnc/
— The International Corpus of English
http://www.ucl.ac.uk/english-usage/ice-gb/
— The Corpus of Spoken Professional American English (CSPAE)
http://www.athel.com/cspa.html
— The Michigan Corpus of Academic Spoken English (MICASE)
http://www.hti.umich.edu/m/micase/
— The Corpus of London Teenage Language (COLT) (available on ICAME CD-ROM)
http://www.hit.uib.no/colt/; http://www.hit.uib.no/icame.html
— ICAME Collection of English Language Corpora
http://www.hit.uib.no/icame.html

Appendix 2: Classroom activities

Task A1
a. Read the extract below and complete the rating scales in relation to the language and communication strategies used in each interaction. Circle your choices.

Formality	Very Formal 1	2	3	Informal 4
Clarity	Very clear 1	2	3	Unclear 4
Brevity	Very brief 1	2	3	Very elaborate 4
Friendliness	Very unfriendly 1	2	3	Very friendly 4
Directness	Very direct 1	2	3	Very indirect 4
Speaker Relationship	Distant 1	2	3	Close 4

b. Do your results appropriately reflect what you might expect to find in this context? Why (not)?

Extract 1 – A teacher trainer and trainee engaged in post-observation TP review and feedback.

Trainer:	Was there a big improvement?
Trainee:	Yeah. I don't know why. I didn't teach at all last week. I got out of it because Linda needed a class so it was good. I wanted a break and I'll do two in a couple of weeks time cos I know other people have done it.
Trainer:	You just needed a break cos last week was the bank holiday.
Trainee:	Yeah the Monday was skipped so it was fewer classes.
Trainer:	So a couple of you ended up not teaching last week.
Trainee:	Yeah.
Trainer:	Yeah but this was good I was+
Trainee:	It was helpful to see Brenda first. Sometimes I watched it didn't help. It made me feel worse but watching her helped.
Trainer:	Was it? And Brenda's a nice teacher.
Trainee:	Yeah.
Trainer:	She's very natural in the classroom.
Trainee:	Yeah I learned from her she's very clear. Some of my instructions weren't clear.
Trainer:	What sort of things did you learn from her? What were you watching?
Trainee:	Well her introduction was good. She didn't just launch into things cos I used to be blunt and launch right in and they're not going to be responsive if you haven't broached the subject.

Task A2

a. Read the extract below and complete the rating scales in relation to the language and communication strategies used in each interaction. Circle your choices.

Formality	Very Formal			Informal
	1	2	3	4
Clarity	Very clear			Unclear
	1	2	3	4
Brevity	Very brief			Very elaborate
	1	2	3	4
Friendliness	Very unfriendly			Very friendly
	1	2	3	4
Directness	Very direct			Very indirect
	1	2	3	4
Speaker relationship	Distant			Close
	1	2	3	4

b. Do your results appropriately reflect what you might expect to find in this context? Why (not)?

Extract 1 – A teacher trainer and trainee engaged in post-observation TP review and feedback.

Trainer: Do you think there was a big improvement?
Trainee: Am yeah I don't know why. I didn't teach at all last week. I kind of got out of it because Linda needed a class so am I think it was good. I just wanted a break yeah and I'll do two in a couple of week's time cos I know other people have done it.
Trainer: Maybe you just needed a break yeah yeah cos last week was the bank holiday wasn't it?
Trainee: Yeah the Monday was skipped so it was less classes anyway but.
Trainer: So a couple of you ended up not teaching last week wasn't that it?
Trainee: Yeah.
Trainer: Yeah but this was this was quite good I mean there was I mean I was.
Trainee: It was it was helpful to see Brenda first actually. Now sometimes I watched like it di= it didn't help at all like it made me feel worse am but am like watching her helped I think.
Trainer: Was it? And Brenda's a nice teacher.
Trainee: Yeah.
Trainer: She's very natural in the classroom.
Trainee: Yeah I learned from her. She's very clear am some of my instructions weren't too clear am.
Trainer: What sort of what sort of things did you learn from her? What were you watching?
Trainee: Well her introduction was good like. She didn't just launch into things cos I used to be a bit blunt and just start launch right in and you know they're not going to be responsive if you haven't somehow like broached the subject a little bit.

Task B
List the hedges that are present in the unedited version of the interaction above by finding a partner with a version different from your own.

Task C
In the following extract the hedges have been removed and are listed underneath. Try to replace the appropriate hedge in the gaps in the text.

Extract 2 — A recorded interview with one of the founding members of a local rural radio station in the south of Ireland. The interviewer is female and the interviewee is male.

Interviewee:	You know the community is getting more complex _____
Interviewer:	Right in what sense?
Interviewee:	Em before in a town like X everybody knew everybody else.
Interviewer:	Right yeah.
Interviewee:	Now it's ____you could say _____it's ah there's a lot of strangers in.
Interviewer:	Okay.
Interviewee:	D= that would have no affiliation to the town people that may be working here that built houses here that wouldn't have any roots in the town.
Interviewer:	People'd be commuting _____?
Interviewee:	People co= exactly commuting _____ thing.
Interviewer:	Yeah and is the station doing anything to build them in to in to the network _____?
Interviewee:	Well it is it's i= i= i= it it is it's it's it's what the station is doing ____ it's trying to cover all aspects of community life out there.
Interviewer:	Okay well.
Interviewee:	_____.
Interviewer:	Who would you see as the that you you interact with then (pause) who who are that community? Is it the town and the surrounding areas?
Interviewee:	The town and the surrounding areas _____. We try to combine them both.
Interviewer:	Okay and wo= Yeah. And _____ you target specific groups within tha= tha= that area?
Interviewee:	For spe= at at specific times.
Interviewer:	Right yeah.
Interviewee:	Like for instance now we say we'd have a sporting programme aimed at the _____the local soccer fraternity here in town.
Interviewer:	Sure.
Interviewee:	And in the country ah football areas etc you know ah they're _____ from the sporting point of view.
Interviewer:	Uh-huh.
Interviewee:	Ah different programmes then we would do we'll say from the from _____ Ait Mor Trash Mor.
Interviewer:	Uh-huh.
Interviewee:	You know that kinda thing that we would….

you know (x 2), would, it's, we'll say, actually, just, maybe, really, or, now, the likes of, kind of

Task D
Choose one of the linguistic hedges from Task C, run a corpora concordance on it and investigate any patterns or conditions of use it favours.

Task E
Taking into account insights gained from the first three tasks you have performed consider the following questions;

1. What are linguistic hedges?
2. Why do we use them in spoken language?
3. What would be the effect of not using them in spoken language?
4. Are they important items of language to learn in the English language classroom? Why/why not?
5. How might they best be learned?

References

Aston, Guy. 1995. Corpora in language pedagogy: matching theory and practice. *Principle and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson,* ed. by Guy Cook, and Barbara Seidlhofer, 257-270. Oxford: Oxford University Press.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7 (1): 1-16.

Baynham, Michael. 1991. Speech reporting as discourse strategy: some issues of acquisition and use. *Australian Review of Applied Linguistics* 14: 87-114.

Baynham, Michael. 1996. Direct speech: what's it doing in non-narrative discourse? *Journal of Pragmatics* 25: 61-81.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4): 243-257.

Biber, Douglas and Edward Finegan. 1994. *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London/New York: Longman.

Binchy, James. 2000. Relational language and politeness in southern Irish service encounters. Unpublished M.A. dissertation, University of Limerick.

Blum-Kulka, Shoshana. 1994. The dynamics of family dinner table: cultural contexts for children's passages to adult discourse. *Research on Language and Social Interaction* 27(1): 1-50.

Blum-Kulka, Shoshana. 1997a. *Dinner Table Talk: Cultural Patterns of Sociability and Socialisation in Family Discourse.* Mahwah, NJ: Lawrence Erlbaum.

Blum-Kulka, Shoshana. 1997b. Discourse pragmatics. *Discourse as Social Interaction*, ed. by Teun A. Van Dijk, 38-63. London: Sage Publications.

Boxer, Diana and Lucy Pickering. 1995. Problems in the presentation of speech acts in ELT materials: the case of complaints. *ELT Journal* 49 (1): 99-158.

Carter, Ronald. 1998. Orders of reality: CANCODE, communication and culture. *ELT Journal* 52 (1): 43-56.

Chomsky, Noam. 1964. *Current Issues in Linguistic Theory*. Paris: Mouton.

Clancy, Brian. 2000. A case study of the linguistic features of a Limerick family. Unpublished M.A. dissertation, University of Limerick.

Conrad, Susan. 2000. Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly* 34(3): 548-560.

Coupland, Justine, ed. 2000. *Small Talk*. Harlow: Pearson Education Ltd.

Crowdy, Steve. 1994. Spoken corpus transcription. *Literary and Linguistic Computing,* 9: 25-28.

Eggins, Suzanne and Diana Slade. 1997. *Analyzing Casual Conversation*. London/New York: Cassell.

Farr, Fiona. In press. Relational strategies in the discourse of professional performance review in an Irish academic environment: the case of language teacher education. *The Pragmatics of Irish English*, ed. by Anne Barron and Klaus P. Schneider. Berlin: Mouton de Gruyter.

Farr, Fiona and Anne O'Keeffe. 2002. *Would* as a hedging device in an Irish context: an intra-varietal comparison of institutionalized spoken interaction. *Using Corpora to Explore Linguistic Variation*, ed. by Randi Reppen, Suzan Fitzmaurice and Douglas Biber, 25-48. Amsterdam: John Benjamins.

Fox, Gwyneth 1998. Using corpus data in the classroom. *Materials Development in Language Teaching,* ed. by Brian Tomlinson, 25-43. Cambridge: Cambridge University Press.

Francis, Nelson. 1982. Problems of assembling and computerizing large corpora. *Computer Corpora in English Language Research*, ed. by Stig Johansson, 7-24. Bergen: Norwegian Computing Centre for the Humanities.

Holmes, Janet. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics* 9: 21-44.

Hughes, Rebecca and Michael McCarthy. 1998. From sentence to discourse: discourse grammar and English language teaching. *TESOL Quarterly* 32 (2): 263-287.

Kettermann, Bernhard. 1995. Concordancing in English language teaching. *TELL and CALL* 4: 4-15.

Leech, Geoffrey. 1991. The state of the art in corpus linguistics. *English Corpus Linguistics: In honour of Jan Svartvik*, ed. by Karin Aijmer and Bengt Altenberg, 1-10. London: Longman.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Stockholm, 4-8 August 1991,* ed. by Jan Svartvik, 105-122. Berlin: Mouton de Gruyter.

McCarthy, Michael. 1991. *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.

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

McCarthy, Michael and Ronald Carter. 2004. "There's millions of them": Hyperbole in everyday conversation. *Journal of Pragmatics* 36 (2): 149-184.

McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

O'Keeffe, Anne. In press. 'You've a daughter yourself?': A corpus-based look at question forms in an Irish radio phone-in. *The Pragmatics of Irish English*, ed. by Anne Barron and Klaus P. Schneider. Berlin: Mouton de Gruyter.

O'Keeffe, Anne. 2003. Strangers on the line: a corpus-based lexico-grammatical analysis of radio phone-in. Unpublished Ph.D. thesis, University of Limerick.

O'Keeffe, Anne and Fiona Farr. 2003. Using language corpora in language teacher education: pedagogic, linguistic and cultural insights. *TESOL Quarterly* 37(3): 389-418.

Schegloff, Emanuel, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53(2): 361-382.

Scott, Michael. 1999. *Wordsmith Tools*. Software. Oxford: Oxford University Press.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1995. Corpus typology — a framework for classification. *Studies in Anglistics,* ed. by Gunnel Melchers and Beatrice Warren, 17-34. Stockholm: Almqvist and Wiksell International.

Sinclair, John. 1997. Corpus evidence in language description. *Teaching and Language Corpora,* ed. by Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 27-39. London: Longman.

Sinclair, John and Malcolm Coulthard. 1975. *Towards an Analysis of Discourse. The English used by Teachers and Pupils.* Oxford: Oxford University Press.

Svartvik, Jan. 1991. What can real spoken data teach teachers of English? *Linguistics and Language Pedagogy: the State of the Art,* ed. by James E. Alatis, 555-565. Washington, DC: Georgetown University Press.

Tognini Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Yngve, Victor H. 1970. On getting a word in edgewize. *Papers from the 6th Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.