Sociolinguistic information and Irish English corpora

Elaine Vaughan, University of Limerick

Brian Clancy, Mary Immaculate College, ~University of Limerick~

1. Introduction

The central part of this chapter presents the sort of sociolinguistic information that is retrievable from some corpora of Irish English (IrE) that currently exist. However, in order to fully explore and contextualise the research possibilities that corpora of IrE offer the sociolinguist, we probe the relationship – emergent, developing or with the potential to develop – between the core concerns of sociolinguistic research and contemporary corpus linguistics. Hence, the nature of language corpora and the fundamental aspects of the sort of analytical tools commonly used to mine them become relevant. An emergent consensus in most recent work on corpus linguistics and sociolinguistics (e.g. Friginal and Hardy 2014) is to take the view that as a methodological approach, corpus linguistics has much to offer sociolinguistics (and vice versa, though this is not as frequently discussed, see Kendall 2011). For the purpose of the present chapter, corpus linguistics is understood to be both an independent field of linguistic enquiry and a principled methodological approach to the analysis of linguistic data, one that is in the process of developing a strong, mutually beneficial research relationship with sociolinguistics, as evidenced in recent book-length treatments (e.g. Baker 2010, or Friginal and Hardy 2014). There are a number of reasons that this interdisciplinary relationship has developed, not least of which is the fact that corpus linguistics and sociolinguistics share what Baker (2010: 8-9) describes as 'fundamental tenets of best practice', *viz.*:

- Both share a focus on naturally occurring language-in-use with context recognised as critical to the production and interpretation of language;
- A quantitative orientation to data analysis is shared;
- Both use sampling techniques to capture the range and complexity of language;
- Both focus on variation across a wide range of linguistic features;

In other words, in essence, corpus linguistics and sociolinguistics overlap in their 'epistemology, focus and scope' (ibid: 9). One of the main concerns for sociolinguists in adopting a corpus-based methodology has been in relation to the type of sociolinguistic information currently available in corpora of different language varieties, a question which this chapter addresses in relation to IrE, but also the nature of what can be considered a corpus given that there is a long and strong tradition of gathering datasets of naturally occurring language-in-use – some large and some far smaller – for sociolinguistic research. This issue is considered below.

2. Corpora in Corpus Linguistics and Sociolinguistics

Corpora (sing. *corpus*) are often described quite simply as databases of naturally occurring language, amenable to automated analysis; however, this is to understate what sorts of text collections are understood as corpora in the corpus linguistic sense. In early scholarly discussions of corpus linguistics as a distinct paradigm (in the Kuhnian sense) within linguistic analysis, a number of qualifying features of what a *corpus* might be began to emerge. These are described concisely by Flowerdew (2012: 3) and can be summarised as follows, a corpus

- consists of authentic, machine-readable, naturally occurring language data;
- is designed according to coherent, principled criteria;
- is representative of a particular language or genre of language.

These criteria certainly limit the scope of which collections of texts might be described as corpora – and perhaps even exclude many datasets that sociolinguists may currently refer to as corpora. A consideration of what a corpus is or might be is therefore highly pertinent to any discussion of the relationship between corpus linguistics and sociolinguistics. If a corpus is defined loosely according

to the criteria summarised by Flowerdew above, then it could be suggested that sociolinguists have been working with corpora, albeit "unconventional" corpora for quite some time (cf. Beal et al. 2007 or D'Arcy 2011).[1] We have argued elsewhere (Vaughan and Clancy 2013) that size is not as important as research focus and design when it comes to corpus building, a position bolstered and inspired by McEnery et al.'s (2006: 2) enlightened and practical approach to the question of size in corpus construction and the retrospective, or (re)consideration of "collections of texts" as corpora:

> If specialised corpora which are built using a different sampling technique from those for balanced corpora were discounted as "non-corpora," then corpus linguistics would have contributed considerably less to language studies.

Add to this Baker's (2010) assertion of the common ground shared by both linguistic fields and the overlap between corpus-based research and sociolinguistics has serious traction.

There are manifest advantages associated with the use of corpora in the study of sociolinguistics. Firstly, since the pioneering work of Labov (1972) who used a process of demographic sampling akin to that used in modern spoken corpora such as the British National Corpus to collect his data, sociolinguistics has needed access to spoken data. Corpora provide sociolinguists with access to spoken language that is naturally occurring, real-word and spontaneous. These elements of spontaneity and naturalness may seem to contrast to the sociolinguistic interview or other methods of data collection such as discourse completion tests; however, if we are flexible in relation to what we consider a "corpus", and admit specialised, principled collections into the fold, this becomes less problematic. This sort of flexibility has challenged our conceptions of the nature of what is considered "spoken" as well as what we might consider a corpus. In diachronic corpus studies, it has been argued that where there are no audio recordings available, we need to work with data that we can reasonably describe as representative of spoken language at a particular point in time. This orientation in blended research of many kinds has expanded the boundaries of both "spoken-ness" and what can be considered corpora, or corpus-like. For example, Archer and Culpepper (2009: 288) refer to historical trial proceedings as 'speech-related data' and argue that these are as close to spoken language as we can get, for this time. Similarly, McCafferty and Amador-Moreno (2012) refer to personal letters as among the more 'oral' text types available for diachronic study, whereas Hickey (2003) centres the design of his corpus around drama – a written genre where the spoken word plays a central part. An important knock-on effect of the broadening of the sociolinguistic research paradigm, and, at least for some of it, a corollary broadening of what types of datasets and approaches might be available to analysts, has been a valuable discussion for both corpus linguists and sociolinguists regarding data and methods (e.g. Pichler 2010).

The nature of corpus design criteria and the metadata preserved in corpus databases has much to offer sociolinguistic research questions, although with obvious caveats. As will be further explored below when we look at what corpora of IrE are available for sociolinguistic research, many corpora are annotated with sociolinguistic metadata. This includes information such as age, gender, level of education, ethnicity and so on, but some corpora also feature information such as the relationship between speakers, the level of formality and information about text, as well as context or genre type, and, as we will discuss, this allows researchers to compare language variation on a number of levels. However, this is not to say that the use of corpora (existing or specially designed) is unproblematic, or that corpora of naturally occurring spoken language particularly do not come with some caveats, as previously mentioned. One of the major issues in the use of corpora for any type of research, sociolinguistic or otherwise, is that the type of research that can be carried out is dependent on, and constrained by, the design and compilation of the corpus. Spoken corpora have tended to be transcribed orthographically – that is to say according to written language conventions primarily. This

[1] Corpora are often written about just these sorts of binary terms: "conventional" versus "unconventional"; "large" ("conventional") versus "small" ("unconventional"), "written" ("conventional") versus "spoken" ("unconventional"). The reality is considerably more complex; perhaps a good translation of "unconventional" here might simply be "specialised".

raises the issue of how, and in what ways, we represent the spoken language in written form, or what spoken language written down looks like, and how authentic it can be to the nature of spoken-ness (see, for example, Du Bois, 1991). Hence, there are some problems associated with the marrying of corpus data and linguistic analysis more generally in relation to the nuanced representation of the spoken language (see section 5). The corpus-based method, using an existing corpus or building something specialised, involves a certain degree of automated analysis, and a large degree of manual, post-hoc qualitative analysis. We will return to how automated analysis may inform close, qualitative analysis in section 5, but before we move on, it is important to consider what information recorded as part of the compilation of a corpus might inform sociolinguistic research, and we do this by considering what kind of information might also qualify as a sociolinguistic variable.

3. The "sociolinguistic variable"

A distinction between two orientations within sociolinguistic research, "sociolinguistics" and "the sociology of language'" (e.g. Tagliamonte 2006), has often been suggested. The sociology of language deals with the relations between society and languages as wholes (Hudson 1996) and addresses socio-political aspects such as language maintenance and shift, language policy and planning and issues surrounding multilingualism. Sociolinguistics, on the other hand, is traditionally concerned with variation and change in language form and use – choice and selection of pronunciations, grammars or vocabularies according to categories such as male/female, socio-economic class or ethnicity, and so on. Although this distinction is somewhat contested (see Wardhaugh 2006: 13-17), corpus studies have, for the moment, allied themselves with the "sociolinguistic" orientation. This orientation involves the study of the complex interaction between two primary variables – *linguistic* and *societal* (Friginal and Hardy 2014; Holmes 2001). Table 1 collates, adapts and expands a summary of possible linguistic and societal variables critical to sociolinguistic research identified by Friginal and Hardy (2014: 4-6).

Table 1: Linguistic and societal variables investigated in sociolinguistics

Linguistic variables	**Societal variables**
Sounds, words and grammatical structures - Pronunciation, intonation, use of words and phrases **Discoursal features** - Overlap, latching, interruption, cohesive devices in writing, repair structures **Pragmatic features** - Politeness, stance, taboo language, speech acts **Communicative features** - Pauses, response tokens, greeting and leave-taking **Paralinguistic markers** - Humour, silence, gesture, body language, emoticons	**Social** - Demographic information such as gender, age, sexuality, educational background, geographical information, class, income, etc. **Situational** - Various communication contexts and registers - Speech community, social network theory, community of practice **Attitudinal and relational** - Power, solidarity, roles and relationships, perceptions and attitudes **Temporal** - Time periods, major historical events, migration patterns **Other** - Personality factors (introvert/extrovert)

As Table 1 illustrates, there are many societal variables that influence our choice of linguistic form and how we use it. Traditionally, sociolinguistic research has focussed on sounds, words and grammatical structures (for example, whether or not we use the items *hood* or *bonnet* or how we pronounce /r/) and what this says about us as speakers or writers in terms of the factors already mentioned such as gender, social class or ethnicity. This traditional focus also encompasses the social significance of the relationship between the situational context and language choice. Factors such as the setting, topic and level of (in)formality are the primary focus here. It is within this traditional remit that we find the majority of corpus-based sociolinguistic studies. However, this traditional focus has broadened significantly and sociolinguists and, indeed, corpus sociolinguists are now concerned with a range of other factors.

4. Corpora of IrE available and suitable for sociolinguistic research

Table 2 describes a number of Irish English corpora that might be used for the purposes of sociolinguistic research, including information about their size, whether they are written or spoken, the time periods they represent, the metadata their databases contain, their availability to researchers and where to find further information about them. We have included the larger-scale corpora that have been compiled but have omitted many of the smaller-scale corpora that have been collected and used for IrE sociolinguistic research by individual researchers such as TravCorp[2] (Clancy 2011a, 2011b), other small corpora used in Filppula (1999) or those created by the Bonn project on variational pragmatics (e.g. Schneider 2005). We can see that the bigger corpora are indeed primarily written in nature. Written corpora are often larger than spoken corpora due to the financial and time demands involved in constructing a spoken corpus and also ethical and permission issues. The majority of corpora that are freely available to researchers are also written corpora. Although written texts arguably do not provide the same rich vein of social variables to be mined as spoken corpora do, the majority contain only information about text type and date of publication, they do provide ample opportunity for the study of diachronic linguistic change in relation to Irish English. Indeed, McCafferty and Amador-Moreno (2012: 265) maintain that, despite a number of book-length treatments of IrE in recent years, 'there is a striking paucity of empirical research taking a long-term diachronic perspective.'

[2] See Section 6 for more information on TravCorp.

Table 2: Existing Irish English corpora

Corpus	Size*	Written or Spoken	Time period	Recorded metadata	Availability	Further information
New Corpus for Ireland (NCI)	c.30m Irish c.25m English	W	1880s – present day	Text-type; date of publication	On request	Foras na Gaeilge cconvery@forasnagaeilge.ie
Corpus of Electronic Texts (CELT)	c.16.7m	W	1200s – present day	Text-type; date of publication	Yes	University College Cork http://www.ucc.ie/celt/
The Irish-English Parallel Corpus	c.6.56m Irish c.6.45m English	W	1930s – present day	Text-type; date of publication	Yes	Fiontar (Dublin City University) http://www.gaois.ie/en/paradocs
Corpus of Irish English Correspondence (CORIECOR)	c.3m	W	c.1760s – early 1900s	Informant information; text-type; date of publication	Under construction	McCafferty and Amador-Moreno (2012)
A Corpus of Irish English	c.635,000	W	1330s – present day	Text-type; date of publication	Yes	Hickey (2003)
ICE Ireland and SPICE-Ireland	c.600,000 (S) c.400,000 (W)	S & W	1990 – 2005	Text-type; geographical information; gender; age; level of education; occupation; religion; first and other languages	Yes	Kallen and Kirk (2008) Kallen and Kirk (2012)
The Limerick Corpus of Irish English (LCIE)	c.1m	S	1998 – 2005	Context-type; goal-type; age; gender; geographical region; occupation; level of education	Restricted	Barker and O'Keeffe (1999)

The Northern Ireland Transcribed Corpus of Speech	c.300,000	S	1973 – 1980	Text-type; age; geographical region	On request	Kirk (1992)
A Corpus of Hiberno-English Speech	c.158,000	S	1970s – 1980s	Text-type; age; gender; geographical region; level of education; occupation	?	Filppula (1999)
Dialects of English: Irish English	c.28,000**	S	2008	Age; gender; geographical region; religion; level of education	Yes	Corrigan (2010)

*All word counts, where necessary, have been generated using WordSmith Tools 5.0 (Scott 2008).
**Approximate word count based on the 29 interview transcripts available on http://www.lel.ed.ac.uk/dialects/ni.html.

The spoken corpora such as the Limerick Corpus of Irish English (LCIE), in contrast to the written corpora in Table 2, provide researchers with the largest amount of demographic variables. Indeed, modern spoken corpora are characterised by their attention to database information. The two largest spoken corpora, LCIE and the Ireland component of the International Corpus of English (ICE-Ireland), contain one million words and 600,000 words of spoken IrE respectively. Both contain detailed demographic information such as age and gender and also information about where the speakers were born and where they lived at the time of recording (geographical information) and level of education. ICE-Ireland also details the religious background of the participants in the corpus which is relevant as the corpus contains speech from both the Republic of Ireland and Northern Ireland. LCIE, on the other hand, was collected exclusively in the Republic. Also included as a spoken corpus in Table 2 is Corrigan's (2010) Northern Ireland contribution to the *Dialects of English* series (Edinburgh University Press). This series allows the systematic comparison of phonological features of different dialects. Although the word count appears small, this is solely based on available interview transcripts. The dedicated website also contains sound files of a reading passage task and a sentence task designed as a resource to allow the comparison of stylistic phonological variation – the interview representing the least formal speech style and the sentences the most formal (due to the greatest amount of attention being paid to the act of speaking). The recently released SPICE-Ireland (Kallen and Kirk 2012), in addition to being tagged pragmatically, is also prosodically tagged for intonation and word stress. As Table 2 shows, many of these corpora are freely available for potential sociolinguistic studies; therefore, our attention now turns to how these might be usefully exploited for sociolinguistic gain through the use of a corpus-based method. We illustrate this through a consideration of the discourse/pragmatic item *shur* in IrE.

5. The corpus-based method: A sample and some observations

A consideration of the tools designed for use with corpora, and the type of quantification and analysis that they provide, is necessary in any discussion of what a "corpus-based" method might constitute. Corpora exist as electronic text files, and this means that they can be analysed via commercially and freely available concordancing software; we outline the automated analysis possible using concordancing software packages such as *WordSmith Tools* (Scott 2008; commercially available) or *AntConc* (Anthony 2014; freely available) below, using a sample enquiry around a corpus-based analysis of the pragmatic marker *sure*/*shur(e)* in IrE. This is presented with two ends in mind. Firstly, it presents the tools themselves, and what automated processes can yield; secondly, it is possible to also highlight some caveats relating to how corpora might be harnessed for sociolinguistic enquiry as well as some limitations of corpora as they (to a large extent) currently exist. For this sample, we have selected *sure*, an example of what have been called *discourse* or *pragmatic markers* (or even *discourse-pragmatic markers*), rendered in the corpus we use as *shur(e)* to reflect a phonological reduction in the spoken mode related to function. This exemplar is fairly basic, but is really in the service of drawing out some points relating to how 'computerised corpora form a well-prepared basis for systematic, descriptive studies of instances of actual speech, for language variation and for how social context constrains communicative practices' (Andersen 2010: 548).

For the Limerick Corpus of Irish English (LCIE; described in Table 2 above), a decision was made at transcription stage to render the particular 'Irish English *sure*' as *shur(e)*. However, in practice, it was initially transcribed in three different ways: as *sure, shure* and *shur*. As a large-scale project in spoken language corpus terms, this is not unusual but is worth noting as it points to a couple of issues that researchers using corpora need to bear in mind – especially sociolinguists, for whom this sort of variation is the motivation and focus of research (Pichler 2010). Firstly, spoken language is hugely complex, and analysis can only proceed once the phenomenon (spoken language) has been captured in some way to allow for detailed observation. This presupposes a number of removes at which a spoken sample of language can be observed: we take a vibrant and mutable phenomenon, which exists only as sound waves (more often than not), and capture it according to written language conventions. It is thus "represented", taken out of its original mode and context, though the transcriber attempts to be as faithful as possible to the original. Of course, the resultant 'static artefact' Varenne (1992: 30) is not

perfect; however, it is perhaps more pragmatic to operate sensibly within the boundaries of this imperfection than to ignore or overlook its potential.

Frequency

For the moment at least, and unless the researcher has access to the sources and resources to create the 'perfect corpus' (if such a thing exists), consulting existing corpora will mean that the analyst needs to be creative and thorough. For a discourse-level item like *sure*/*shur(e)*, this means anticipating the ways in which it might have been transcribed, and trawling a frequency list for those realisations. Frequency is one of the basic – and yet revealing and interesting – automatic processes available via corpus software. It represents "entry-level" access to the corpus (cf. Baker 2006). For a corpus-driven approach, frequency of an item or items may justify further investigation. However, sociolinguistic insight could, and frequently does, identify *a priori* elements of language variation. A characteristic of word-lists, the output view made possible by concordancing software, to point to a surfeit or dearth of an item in terms of frequency can be illuminating either way, we would argue. One characteristic of frequency lists is that they consist of mainly "small", functional items, the interactional potential of which should not be underestimated (cf. Vaughan and Clancy 2013). The corpus-based approach has by and large been a comparative enterprise. Therefore, a frequency list is often all the more interesting when compared to, for example, a list from a different variety (taking the concept of variety as a broad one, at context level).

If we compare the raw frequencies for pragmatic marker *sure/shur(e)* in LCIE (1277 occurrences), with corresponding frequency of occurrence in SPICE-Ireland (194 occurrences) and then with the spoken component of the British National Corpus (BNC; 11 occurrences), it looks like there is a fairly solid quantitative basis to claim it as a pragmatic marker typical of Irish English, but, of course, all these corpora are different sizes: LCIE contains 1 million words; SPICE-Ireland, 600,000 words of spoken IrE, and the spoken component of the BNC, 10 million words.[3] In order to make frequency information for datasets comparable, *normalisation*, a basic but informative process can be used. Normalised frequency (*nf*) can be achieved by using a simple calculation as demonstrated by Biber (1988). If we want to normalise per million words, for example, the calculation is as follows:

$$nf = \frac{\text{number of occurrences}}{\text{total number of words}} \times 1{,}000{,}000$$

When we normalise the frequencies for each corpus per million words, as in Table 3, we can see that *sure/shur(e)* is most frequent in LCIE.

Table 3: *Sure/shur(e)* in LCIE, SPICE-Ireland and the BNC normalised per million words

	LCIE	SPICE-Ireland	BNC
Raw frequency	1277	194	11
Normalised frequency	1277	310	1

Once a frequency list has been generated, another automatic procedure, the generation of keywords, can be harnessed. This creates the possibility for a different form of comparison, one that is based on *saliency*; in other words, what most strikingly frequent or infrequent in relation to a comparative baseline, usually a frequency list from another corpus, or another component of the primary corpus being used. It is usual in corpus linguistic terms to ensure that this baseline corpus is a much larger corpus, comparatively, representative of the variety that is being investigated. The saliency measure is a cross-tabulation based on a statistical test (either Chi-square or log-likelihood) to ascertain which items occur with unusually high or low frequency. Keywords are therefore not necessarily the most frequent words, but the most unusually frequent, or infrequent, words. This is a valuable measure in terms of sociolinguistic research, given that an item, or feature, of language may have been isolated

[3] These counts refer to the use of *sure/shur(e)* as a pragmatic marker (***Shur** he never goes there*), and not as an adjective (*I'm **sure** I left it here*).

for just that frequency or infrequency in a varietal context. The reference, or comparison, corpus used, will obviously have an impact on what items or terms emerge as key.

Table 4 below shows the top ten keywords for LCIE when the spoken component of the BNC is used as a reference corpus (vocalisations, such as *uh* and *hm* have been removed, as has extralinguistic information, such as laughter). In this view, *shure* is highlighted as a key.

Table 4: Top ten keywords in LCIE using BNC Spoken as reference corpus

	Keyword
1	like
2	**shure**
3	yeah
4	goin’
5	cause
6	tis
7	d’you
8	now
9	kind
10	grand

Concordance

At this point, a third function, and one that involves a significant level of human intervention – especially if it is used as we suggest, in tandem with the metadata provided in the database for the corpus – is relevant. The concordance line view involves pre-selecting the item/s for analysis. The software searches the corpus and generates concordance lines that contain the item/s, the *node*, and the five or six words that occur immediately left and right of it. Figure 1 below shows 15 concordance lines for the *shur* rendering in LCIE.

Figure 1: Sample of concordance lines for *shur* in LCIE (sorted one item to the left)

N	Concordance
1	<\$O8> the burgers over like . <$4> On his feet all day . <$1> **Shur** there's no way he'll be able for that like . <$E> speaker four
2	somewhere like . <$2> You've only thirty six cards MX . <$1> **Shur** I can't access it in U L where it is on the hard drive like .
3	<$1> Why? <$3> Look . Cos she looks at you sometimes . <$2> **Shur** half the dogs in Limerick are called Gizmo . <$4> That's the
4	first cousin . <$1> And he'll yeah he'll let you in with that accent . **Shur** didn't you get in no hassle the last time . <$3> Oh I did
5	there . So I mean that why I don't have it done . <$3> Yeah but **shur** you're not going to do anything like . <$1> I know but I still
6	| away <\$X> . Go <$X> way | away <\$X> from it . <$2> But **shur** or he could be working at the weekend you see . <$1> No
7	so young she wouldn't notice it . <$3> She wouldn't have a clue **shur.** <$4> We could've changed it . <$4> We could call her am
8	and what was the third one? <$3> Every second word is fuck **shur.** <$1> Yeah I watched a bit of it . <$3> There's fuck this
9	<$2> She's going to the Stable's drinking . <$1> Yeah I know **shur** I don't know why but she's after telling me ten times where
10	they made a lovely taglietelli vegetable thing . <$1> So FX told me **shur** FX is goin to make it some day . <$3> But I never took it
11	<$2> Definitely . <$1> Are you gutted about it? <$2> Big time . **Shur** wouldn't you be? <$1> Yeah . Why don't you go? <$2>
12	MX? <$2> No . <$3> No MX would be the wrong area all together **shur.** <$6> Did you meet the boss man? <$3> MX is just
13	. <$1> We're supposed to be makin the film this weekend **shur.** I wouldn't say that's goin to happen either now . <$3> Go
14	see if <$X> they're | they are <\$X> working . <$4> Yeah <$X> **shur** | sure <\$X> <$X> we'll | we will <\$X> plug them in there .
15	. <$3> She asked me did it mean they changed colour . <$X> **Shur** | sure <\$X> I didn't have a clue . <$4> You play with the

By examining these concordance lines, we can formulate initial hypotheses using patterning of *shur* as our starting point. Looking at Figure 1, one feature of note is that a speaker tag such as <$1>, <$2>,

etc. frequently occurs either as the first item to the left of *shur* (for example, line 1) or as the second item, for example before *but* is line 6, *yeah* in line 14 or *yeah but* in line 5. This may indicate that *shur* is often positioned as a turn initial item. Similarly, although to a lesser extent, a speaker tag occurs immediately to the right of *shur* in lines 7, 8, 12 and 13 indicating that it may also have a less frequent position as a turn final item. Previous research has shown that these positions are associated with particular discourse and pragmatic functions. Initial position is often commonly associated with discourse marking, whereas final position is associated with attention to face (Clancy and Vaughan 2012). Corpus software also makes it possible to interact with the complete original text file, as well as with the metadata. If we take the concordance line, *Shur wouldn't you be?* (line 11), it is possible to return to the original text and use the information from the original to see who is speaking. Extract 1 shows the original stretch of discourse in which *shur* is used.

Extract 1

<$1> Is Nessa going to America she is?
<$2> Definitely.
<$1> Are you gutted about it?
<$2> Big time. Shur wouldn't you be?
<$1> Yeah. Why don't you go?
**<$2>** <$E> sniffs <\$E> Cos I can't get my J one. I'll be over for holidays. That's my America fund there.
<$1> Mm hm.

A code is given to each file in LCIE, so that in the database file that contains all of the metadata. Therefore, it is possible to find out who is using the item, in this case, *shur*, the date the recording was made, the gender of the speaker, their occupation and level of education. Figure 2 shows the sort of metadata preserved with the original recordings for LCIE.

Figure 2: Sample of metadata preserved in the LCIE database

TITLE	Rec_Year	RELATIONSHIP_	S1_age	S2_age	S1_BPlace	S2_BPlace	S1_Occupation	S2_Occupation
Close family chatting	2002	intimate	21	64	Cork	Cork	Student	Housewife
Friends chatting about weight	2001	intimate	22	20	Cork	Cork	Student	Student
Friends chatting about an illness	1999	intimate	22	21	Cork	Cork	Student	Student
Friends chatting	2003	intimate	21	22	Cork	Cork	Student	Student
Family/friends chatting	2002	intimate	24	63	Cork	Cork	Student	Housewife
Friends chatting about a career	2002	intimate	23	21	Cork	Cork	Student Union officer	Student
Friends chatting	2002	intimate	25	22	Cork	Cork	Insurance Clerk	Student
Friends chatting	2002	intimate	21	20	Cork	Cork	Student	Student
Friends chatting	2002	intimate	21	20	Cork	Cork	Student	student
Friends chatting about Liz Hurley	2002	intimate	23	23	Cork	Cork	General operative	Bar waitress
Family/friends chatting	2002	intimate	22	21	Cork	Cork	Student	Student
Family/friends chatting	2002	intimate	X	21	Cork	Cork	Lecturer	Student
Family/friends chatting	2002	intimate	X	X	Cork	Cork	Student	Lecturer
Close family chatting * 3 spkrs	2002	intimate	X	21	Cork	Cork	Lecturer	Student
Close family/friends chatting	2002	intimate	63	21	Cork	Cork	Housewife	Student
Friends chatting about pop singers	2002	intimate	21	22	Cork	Cork	Student	Student
Close family/friends chatting about sick friend	2002	intimate	21	63	Cork	Cork	Student	Housewife
Close family/friends chatting	2002	intimate	32	33	West Meath	Cork	Lecturer	Lecturer
Close family/friends chatting	2002	intimate	22	21	Cork	Cork	Student	Student
Flatmates chatting	202	intimate	21	23	Cork	Cork	Student	Student Union Officer
Close family chatting	2002	intimate	68	63	Cork	Cork	Retired farmer	Housewife
Close family chatting	2002	intimate	21	22	Cork	Cork	Student	Student

Knowing that it is Speaker 2 (<$2>) who uses the item *shur* in extract 1 means that it is possible to retrieve information about that situation and that speaker: the speaker was 20 years' old in 2002 when the recording was made, male, a student (born and living in Cork). There is therefore the potential to create a highly contextualised, socially based picture of the use of *shur*. Comparing corpora designed according to the same criteria has meant that national varieties can be compared, and there is clearly great scope for this sort of comparative work, using available tools and metadata. We turn now to some relevant studies which have emerged from this tradition, and highlight some that harness corpora imaginatively and thoroughly, yielding great insight into situated linguistic phenomena.

6. Sociolinguistic case studies of IE using corpora

Thus far we have discussed the reciprocal benefits of the blend of corpus linguistics and sociolinguistics and illustrated, through a brief analysis of a pragmatic marker unique to Irish English *shur*, the potential of the automated processes the software makes available. We have also described in Table 2 the various corpora of Irish English that have been created and that can be used for the purposes of sociolinguistic research. Our attention now turns to these corpora and how they have been used to create an emerging sociolinguistic profile of the IrE from both a synchronic and diachronic persepctive. Interestingly, for the discussion of the blend of corpus linguistics and sociolinguistics, some of these corpora, for example, LCIE, were not created for an express sociolinguistic purpose, whereas others, for example, CORIECOR, were created specifically to examine sociolinguistic variation using corpus techniques. There are a number of studies that have highlighted the salience of the marker *like* in Irish English in general both in spoken and written language. Although *like* is by no means unique to IE (see, for example, Andersen 2001; Tagliamonte 2005; D'Arcy 2007; Miller 2009), the marker (in addition others such as *you know* and *now*) has emerged from corpus-based studies as a prominent item in the socio-pragmatic system of Irish English (see, for example, O'Keeffe et al. 2011; Vaughan and Clancy, 2011). This is despite the fact that, as Amador-Moreno (2010) maintains, *like* can be considered a relatively new development in IrE. Kallen (2006) demonstrates how clause- or sentence-final *like* is more frequent in ICE-Ireland than in ICE-Great Britain. Schweinberger (2012) also uses the ICE suite of corpora and found a striking difference in frequency of *like* between IrE and South-Eastern British English in his data. He demonstrates how speakers of IrE prefer clause-final position *like*, primarily associated with mitigation, while British English speakers predominantly employ the marker in clause-medial position. These differences are attributed to the social meaning and covert prestige attached to the marker, pointing to the reluctance of middle-aged or older speakers of British English to adopt a feature stigmatised as being "American". In relation to the intimate context-type in LCIE, Clancy (2005, 2011a, 2011b) has shown that although *like* is one of the most frequent hedging items in Irish family discourse. Clancy (2011a and b) has also built a corpus of family discourse from within the Irish Traveller Community, a distinct ethnic group in Irish society, in order to demonstrate how factors such as ethnicity, age and level of education play a role in people's use of pragmatic markers. He found that pragmatic markers were more frequent in the discourse of settled, middle class families than in Traveller families where it is a relatively rare feature of their discourse. Factors such as ethnicity – the Traveller Community prioritise family to such an extent that their social networks consist almost entirely of extended family – and level of education – two thirds of all Travellers in Ireland are educated to, at most, primary level – play a large part in this discrepancy.

In addition to *shur* and *like*, *now* has also emerged from corpus-based studies as a key item in Irish English. Clancy and Vaughan (2012) have shown that *now* is more frequent in the spoken Irish English represented in LCIE than in other spoken corpora such as the British National Corpus (BNC)[4], the Corpus of Contemporary American English (COCA)[5] or the Scottish Corpus of Texts and Speech (SCOTS).[6] They, in part, attribute this frequency difference to the nature of the different corpus designs. For example, in terms of situational variables, the spoken component of the BNC contains spoken language from more formal settings such as debates, interviews or commentaries than is contained in LCIE. However, previous studies on *now* have maintained that it is more frequent in these formal speech contexts than in informal ones such as the intimate discourse that we are concerned with here (see Aijmer, 2002; Defour, 2008). Clancy and Vaughan's (2012) frequency results appear to contradict this given that the data contained in LCIE is composed primarily of the intimate and socialising context-types, both of which can be classified as "informal". It has been

[4] The spoken component of the BNC (10 million words) consists of demographically sampled texts complemented by texts collected by context-governed criteria (see www.natcorp.ox.ac.uk).

[5] The spoken component of COCA contains over 90 million words of unscripted conversation from more than 150 television and radio programmes (see corpus.byu.edu/coca/).

[6] The spoken component of the SCOTS corpus contains approximately 800,000 words of Scots and Scots English collected from a range of geographical locations featuring speakers of different genders, ages, occupations etc. (see www.scottishcorpus.ac.uk).

shown that *now* is highly polysemous, functioning as a temporal adverb, a discourse marker or an intensifier, for example. Clancy and Vaughan (ibid.) maintain that it is the socio-pragmatic function of *now* that is pivotal in understanding the behaviour of the marker. This pragmatic function is markedly more frequent in informal Irish English than the sample of British English they compare it with (the spoken component of the BNC). In IE, *now* functions in final position in the utterance to soften or mitigate face threatening behaviour such as disagreement, challenge or evaluation, a function that is almost absent in the BNC data. This investigation of *now* in Irish English also highlighted an additional function of *now* as a deictic presentative. The most commonly recognised deictic presentatives such as the French *voici/voilà* or the Russian *vot/von* are examples of a linguistic item whose use is commonly accompanied by a gesture such as the presentation of food or drink (cf. Fillmore, 1975: 41; Grenoble and Riley, 1996). These addition functions performed by *now* in Irish English are essential to our understanding of sociolinguistic competence in the variety.

LCIE has also been used to explore variation in communicative features of language such as response tokens. O'Keeffe and Adolphs (2008) examined the occurrence of response tokens, verbal and non-verbal response to a speaker that indicate listenership without changing the speaker turn, for example, *yeah*, *right*, *no* or a simple head nod, in two corpora – the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) and the LCIE. On an inter-varietal level, they found that in the British English data, response tokens were both more frequent and were comprised of a wider range forms but that the Irish English data demonstrated a greater degree of informality of use. These differences can be attributed to cultural idiosyncrasies across the two national varieties. The tokens *yes* and *quite*, which can be considered as formal options for responding to something that has been said, occur in CANCODE but have no corresponding occurrence in the Irish English data. In contrast, Irish speakers favour a wider range of "taboo" religious responses: in addition to *oh God* used in both datasets, Irish speakers also frequently use *Jesus* or *Jesus Christ*. They attribute the higher frequency of religious references to, paradoxically, the washing out of their previous force but also a mark of the continuing importance of religion in Irish society. An analysis of response token function was then carried out using two sub-corpora that were controlled for the social demographic variables of gender, age and socio-economic class. They found that in the informal speech of co-habiting, middle-class, female speakers in their 20s, that there was no real variation at the level of response tokens' pragmatic function. This raises a number of interesting sociolinguistic questions regarding variation at the level of discoursal features between the two varieties – for example, do Irish people, because they use fewer response tokens, yield the turn less and interrupt more than British people?

In terms of diachronic variation in IrE, McCafferty and Amador-Moreno (2012) argue for an empirical diachronic approach to the study of IrE in the eighteenth and nineteenth centuries, a period when 'Irish English itself evolved and Anglophone settlement of North America and the southern hemisphere lead to the development of American, Canadian, Australian, New Zealand, and other colonial Englishes' (p. 282). In Late Modern English, the progressive increased in frequency and acquired new uses and diachronic corpora such as CORIECOR allow us to investigate the extent to which the spread of IrE contributed to this grammatical variation and change. In a pilot study using CORIECOR, McCafferty and Amador-Moreno (ibid.) found that in the late eighteenth century, the progressive became much more frequent in IrE and by 1840 it was four times more frequent than in 1770. At this time, it is also more frequent in the CORIECOR data than in matched British English data. They posit a number of sociophilological reasons as to why this might be the case. Firstly, the possessive may have grown in frequency in Late Modern English due to a corresponding growth in Irish emigration to other English speaking countries such as the United States. Also, the rise in the use of the progressive at the time might be due to the rise in literacy levels of the lower classes. The rise of literacy levels in Ireland occurred at the same time as the decline in the use of the Irish language and the acquisition of English. This increase in literacy could have led to a colloquialisation of the language as 'more of the linguistic usage of lower social strata will be recorded in texts produced by members of those strata' (p. 280). This increase in literacy levels, they argue, could also be responsible for the shift from first-person *shall* to *will* that occurred around the same time period (McCafferty and Amador-Moreno, 2014). Interestingly, one of the quintessential features of the

progressive in Irish English, the *after*-perfect, was found to be infrequent in the time period represented by CORIECOR.

The use of this quintessentially Irish English progressive in modern-day Irish English was the focus of O'Keeffe and Amador-Moreno's (2009) study of instances of the grammatical structure *be* + *after* + Verb-*ing* in LCIE. This progressive structure, an Irish language calque, is used to approximately convey the standard English perfect aspect. 95 occurrences of the structure were found in the one-million-word LCIE and their functions classified. In relation to function, They maintain that this structure has a range of pragmatically specialised meanings in Irish English that cannot be replicated by any standard equivalent form. For example, in the context of narrative, they argue that this progressive acts as a 'metalinguistic trigger…heralding the main event of the storyline' (p. 529). In order to further investigate the structure's sociolinguistic profile, age and gender were taken into consideration. It was found that of the 95 occurrences, 73% were used by females. In addition, this use of the progressive is particularly robust in the 18-25 year old age category. They argue that this marks the structure as core to the grammar of modern Irish English.

7. Concluding remarks

Where once the analysis of language varieties using language text corpora occupied a relatively obscure, niche position in comparison to other linguistic traditions more generally, it can be argued that it has come to prominence in the study of Irish English as a variety within the last decade at least. Much credit goes in particular to the contributions made by Kirk and Kallen and the ICE-Ireland project (see, for example, Kallen 2005 and 2006; Kallen and Kirk, 2007), as well as the more-or-less contemporaneous Limerick corpus project (Barker and O'Keeffe 1999), as well as to the work of Hickey (2003). The increasing ease of recording and transcribing spoken language data (relatively speaking) has meant that in recent years, corpora and corpus methodologies are becoming more and more widely referenced. It could be argued that significant questions remain in respect of whether or not corpora are consulted in genuinely informed ways, and to what extent we can say that available corpora are equal to the tasks currently being asked of them. At the very least, it is possible to be open-minded and creative with existing resources, bearing in mind what they can and cannot tell us, or what they might provide partial or supporting evidence of. The possibilities of corpora and corpus linguistic methodologies for access to large quantities of authentic data, and swift, automatic analyses that would be not only laborious but potentially inaccurate if attempted manually are the most often cited. However, it is arguably the emergent trend to blend corpus methodologies and other analytical and theoretical frameworks where the real value resides, though the potential is, as yet, not fully developed. As we have pointed out, it is only in recent years that book-length treatments of the beneficial relationship between corpus linguistics and sociolinguistics such as Baker (2010) and Friginal and Hardy (2014) have emerged. This situation is by no means unique – the blend of corpus linguistics and pragmatics has similarly come to the fore in recent years through the work of Romero Trillo (2008) and Aijmer and Rühlemann (2015).

There is still much work to be done in order that corpus linguistics be of further benefit to sociolinguistics (and, indeed, vice versa). The recent announcement of a new, publically available 'Spoken British National Corpus 2014' is to be welcomed as it gives sociolinguists access to a contemporary demographically balanced corpus that will allow comparison with many of the formative spoken corpora that were primarily designed and constructed in the early to mid-nineties. This new BNC is being recorded on MP3 sound files which should address one of the main criticisms of spoken corpora levelled by sociolinguists – that access to good quality sound recordings is largely restricted or unavailable. In addition, advances in the design and construction of multi-modal corpora coupled with modern technology such as voice recognition software and digital recording, both audio and visual should result is a corpus that allows sociolinguists to access sound, orthographic transcription and visual images simultaneously. The potential of these new corpora might also encourage a corresponding shift away from the dialectologically informed tradition focus in corpus-based sociolinguistic research, toward the outliers such as paralinguistic variation. In the Irish context, as we have shown, the larger spoken corpora of Irish English were finished in and around 2005, therefore, the time is ripe for an ICE-Ireland 2.0 or an LCIE 2.0 or indeed, a larger-scale corpus that is

representative of both spoken IrE and Gaeilge and designed in a way that makes it both available to and suitable for not only researchers interested in sociolinguistics, but in a range of different linguistic and non-linguistic traditions.

References

Aijmer, K., 2002. *English Discourse Particles*. Amsterdam: John Benjamins.

Aijmer, K. and C. Rühlemann (eds.), 2015. *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press.

Amador-Moreno, C., 2010. *An Introduction to Irish English*. London: Equinox.

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

Andersen, G., 2010. How to use corpus linguistics in sociolinguistics. In: A. O'Keeffe and M. McCarthy (eds.). *The Routledge handbook of Corpus Linguistics*. London: Routledge, 547-562.

Anthony, L., 2014. AntConc Version 3.4.3. Available on-line at: http://www.laurenceanthony.net/antconc_index.html (accessed 06.12.2014).

Archer, C. and J. Culpeper, 2009. 'Identifying key sociophilological usage in plays and trial proceedings (1640–1760): An empirical approach via corpus annotation.' *Journal of Historical Pragmatics*, 10(2), 286-309.

Baker, P., 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P., 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Barker, G. and A. O'Keeffe, 1999. 'A corpus of Irish English – past, present, future.' *Teanga* (Yearbook of the Irish Association for Applied Linguistics), 18, 1-11.

Beal, J., K. Corrigan and H. Moisl (eds.), 2007. *Creating and Digitizing Language Corpora, V.2: Diachronic Databases*. Basingstoke: Palgrave-Macmillan.

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

Clancy, B., 2005. 'You're fat. You'll eat them all.' Politeness strategies in family discourse. In: K. Schneider and A. Barron (eds.), *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter, 177-197.

Clancy, B., 2011a. 'Complementary perspectives on hedging behaviour in family discourse: The analytical synergy of corpus linguistics and variational pragmatics.' *International Journal of Corpus Linguistics*, 16(3): 372-391.

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, 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 Chiosáin (eds.), *New Perspectives on Irish English*. Amsterdam: John Benjamins, 225-246.

Corrigan, K., 2010. *Irish English, volume 1 – Northern Ireland*. Edinburgh: Edinburgh University Press.

D'Arcy, A., 2007. '*Like* and language ideology: disentangling the fact from fiction.' *American Speech,* 82: 386-419.

D'Arcy, A., 2011. Corpora: Capturing language in use. In: W. Maguire and A. McMahon (eds.), *Analysing Variation in English*. Cambridge: Cambridge University Press, 49-71.

Defour, T., 2008. 'The speaker's voice: A diachronic study on the use of *well* and *now* as pragmatic markers.' *English Text Construction*, 1(1): 62-82.

DuBois, J.W., 1991. 'Transcription design principles for spoken discourse research.' *Pragmatics*, 1(1): 71-106.

Filppula, M., 1999. *The Grammar of Irish English: Language in Hibernian Style*. London: Routledge.

Fillmore, C., 1975. *Santa Cruz Lectures on Deixis*. Bloomington, IN: Indiana University Linguistics Club.

Flowerdew, L., 2012. *Corpora and Language Education*. London: Palgrave Macmillan.

Friginal, E. and J. Hardy, 2014. *Corpus-Based Sociolinguistics: A Guide for Students*. London: Routledge.

Grenoble, L. and M. Riley, 1996. 'The role of deictics in discourse coherence: French *voici/voilà* and Russian *vot/von*.' *Journal of Pragmatics*, 25, 819-838.

Hickey, R., 2003. *Corpus Presenter: Software for Language Analysis*. Amsterdam: John Benjamins.

Holmes, J., 2001. *An Introduction to Sociolinguistics*. London: Longman.

Hudson, R., 1996. *Sociolinguistics*. Cambridge: Cambridge University Press.

Kallen, J., 2005. Politeness in Ireland: 'In Ireland, it's done without being said.' In: L. Hickey and M. Stewart (eds.), *Politeness in Europe*. Clevedon: Multilingual Matters, 130-144.

Kallen, J., 2006. *Arrah, like, you know*: The dynamics of Discourse Marking in ICE-Ireland. Plenary paper presented at Sociolinguistics Symposium, July, Limerick. Available on-line: http://www.tara.tcd.ie/bitstream/handle/2262/50586/Arrah%20like%20y%27know.pdf?sequence=1 (accessed 02.12.2014).

Kallen, J. and J. Kirk, 2007. ICE-Ireland: Local variations on global standards. In: J. Beal, K. Corrigan and H. Moisl (eds.), *Creating and Digitizing Language Corpora*, Vol. 1: Synchronic Databases. London: Palgrave, pp. 121-162.

Kallen J. and J. Kirk, 2008. *ICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banríona.

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

Kendall, T., 2011. 'Corpora from a sociolinguistic perspective.' *Revista Brasileira de Linguística*, 11(2): 361-389.

Kirk, J., 1992. The Northern Ireland Transcribed Corpus of Speech. In: G. Leitner (ed.), *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter, 65-73.

Labov, W., 1972. *Langauge in the Inner City: Studies in the Black English Vernacular*. Philadelphia: University of Pennsylvania Press.

McCafferty, K. and C. Amador-Moreno, 2012. A corpus of Irish English correspondence: A tool for studying the history and evolution of Irish English. In: B. Migge and M. Ní Chiosáin (eds.), *New Perspectives on Irish English*. Amsterdam: John Benjamins, 265-287.

McCafferty, K. and C. Amador-Moreno, 2014. '[The Irish] find much difficulty in these auxiliaries…putting *will* for *shall* with the first person': The decline of first-person *shall* in Ireland, 1760-1890.' *English Language and Linguistics*, 18, 407-429.

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

Miller, J., 2009. Like and other discourse markers. In P. Peters, P. Collins and A. Smith (eds.), *Comparative Studies in Australian and New Zealand English: Grammar and Beyond*. Amsterdam: John Benjamins, 317-338.

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.

O'Keeffe, A. and C. Amador-Moreno, 2009. 'The pragmatics of the *be* + *after* + Verb-*ing* construction in Irish English.' *Intercultural Pragmatics*, 6(4), 517-534.

O'Keeffe, A., B. Clancy and S. Adolphs, 2011. *Introducing Pragmatics in Use*. London: Routledge.

Pichler, H., 2010. 'Methods in discourse variation analysis: Reflections on the way forward.' *Journal of Sociolinguistics*, 14(5): 581-608.

Romero-Trillo, J. (ed.), 2008. *Corpus Linguistics and Pragmatics: A Mutualistic Entente*. Berlin: Walter de Gruyter.

Schneider, K., 2005. *No problem, you're welcome, anytime*: Responding to thanks in Ireland, England and the USA. In: A. Barron and K. Schneider (eds.), *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter, 101-140.

Schweinberger, M., 2012. The discourse marker LIKE in Irish English. In: B. Migge and M. Ní Chiosáin (eds.), *New Perspectives on Irish English*. Amsterdam: John Benjamins, 179-202.

Schweinberger, M., forthcoming. A comparative study of the pragmatic marker LIKE in Irish English and in south-eastern varieties of British English. In: C. Amador-Moreno, K. McCafferty and E. Vaughan (eds.), *Pragmatic Markers in Irish English*. Amsterdam: John Benjamins.

Scott, M. 2008. *WordSmith Tool Version 5.0*. Liverpool: Lexical Analysis Software Ltd.

Tagliamonte, S., 2005. '*So* who? *Like* how? *Just* what? Discourse markers in the conversation of young Canadians.' *Journal of Pragmatic*, 37: 1896-1915.

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

Varenne, H., 1992. *Ambiguous Harmony: Family Talk in America*. Norwood, NJ: Ablex.

Vaughan, E. and B. Clancy, 2011. 'The pragmatics of Irish English.' *English Today*, 27(2): 49-54.

Vaughan, E. and B. Clancy, 2013. 'Small corpora and pragmatics.' *The Yearbook of Corpus Linguistics and Pragmatics*, 1: 53-73.

Wardhaugh, R., 2006. *An Introduction to Sociolinguistics*. Oxford: Blackwell.