7
Building a corpus to represent a variety of a language

Brian Clancy

1. What is a variety of a language?

In the literature, a *variety* of a language is, to say the least, broadly defined. Crystal (2001: 6–7) maintains that in its most general sense, the notion of a variety includes 'speech and writing, regional and class dialects, occupational genres (such as legal and scientific language), creative linguistic expression (as in literature), and a wide range of other styles of expression.' Similarly, McEnery *et al.* (2006: 90) suggest that varieties of a language are equally expansive, covering, for example, 'the standard language (standardised for the purposes of education and public performance), dialects (geographically defined), sociolects (socially defined), idiolects (unique to individual speakers) and jargons (particular to specific domains)'. This very broad definition of variety is in itself problematic, especially for the corpus builder(s). Nevertheless, according to Crystal (2001: 6–7), 'a variety of language is a system of expression whose use is governed by situational factors … varieties are, in principle, systematic and predictable.' This view is echoed by McEnery *et al.* (2006: 90) who maintain that 'a language variety can be broadly defined as a variant of a language that differs from another variant of the same language systematically and coherently.' Therefore, as Crystal (2001: 7) observes,

> it is possible to say, with some degree of certainty in a given language, how people from a particular region will speak, how lawyers will write, or how television commentators will present a type of sport. Notions such as 'British English', 'Liverpool English', 'legal French' and 'sports commentary' are the result.

Quirk (1995) refers to this profusion of linguistic varieties and the confusion that these cause. He cites the example of the word *English* preceded by a specific adjective or noun to designate a specific variety. This list of varieties of English includes, but is certainly not limited to, *American English*, *legal English*, *BBC English*, *working-class English*, *Chicano English* and *South African English*. He claims that although each is referred to as a variety, they are all formed on 'desperately different taxonomic bases' (p. 22). For example, academic English is a variety that may be used equally by speakers of both American English and British English, and speakers of other languages such as Spanish. This in turn raises

the question of whether or not there exists a variety of American Academic English as opposed to one of British Academic English as opposed to Spanish Academic English. One of the reasons for the variety of varieties listed by Quirk (ibid.) is that corpus linguists have a different method of conceptualising language variation from sociolinguists. In terms of corpus linguistics, varieties are generally explored according to *register variation* or *genre variation*. Biber *et al.* (1999: 15) use *register* and *variety* interchangeably, where register is used as a cover term for 'varieties relating to different circumstances and purposes'. These registers are delimited in non-linguistic terms, with respect to situational characteristics such as mode, interactiveness, domain, communicative purpose or topic. This results in varieties being classified in terms of registers such as academic English, legal English, crime fiction, etc. An example of a corpus constructed in this way is the Longman Spoken and Written English Corpus (LSWE, see Biber *et al.* 1999) which consists of forty million words across four core registers: conversation, fiction, news and academic prose. It was built to describe the main grammatical features of English and the actual use of each major feature, thereby allowing the study of how language varies according to the context in which it occurs. The LSWE also samples two national varieties – American English and British English. Biber *et al.* (ibid.) refer to differences within American English and British English as *dialectal* differences, and it is to this distinction that we now turn.

Although the term *register* is also widely used in sociolinguistics to refer to 'varieties according to use' (Hudson 1980: 48), the primary focus of this discipline is on *dialect*, that is 'varieties according to user' (ibid.). McEnery *et al.* (2006) define language variety geographically. They refer to national variants of a language such as American, British or Irish English as language varieties whereas regional variants (the English used in New York, Norwich or Belfast, for example) are referred to as dialects. Biber *et al.* (1999: 17) define dialects as varieties associated with different groups of speakers, 'distinguished primarily by pronunciation, and to a lesser extent by lexical and grammatical differences'. Although pronunciation has received some attention in corpus linguistics (see, for example, Knowles 1990; Cheng *et al.* 2008), sociolinguistics has long been characterised by a study of dialectal variation that concentrates primarily, though by no means exclusively, on pronunciation features (see, for example, Labov 1966, 1972; Trudgill 1974; Milroy 1987; Wolfram and Schilling-Estes 2006). In sociolinguistics, the primary focus is how sociolinguistic variables such as age, gender and social class affect the way that individuals use language. These studies in turn give rise to the varieties such as BBC English, working-class English and Chicano English in Quirk's (1995) list. According to Meyer (2002), the main reason that there are not more corpora used to study sociolinguistic variation is that 'it is tremendously difficult to collect samples of speech, for instance, that are balanced for gender, age and ethnicity' (p. 18). Most corpora, for example the BNC, the Corpus of London Teenage English (COLT), CANCODE or LCIE, do contain information on sociolinguistic variables; however, corpus linguists appear, in the main, to be primarily concerned with what the speakers are doing rather than who they are.

Accordingly, the starting point for the building of a corpus for a variety of a language could usefully be based on a fundamental decision: is the proposed corpus being built to represent a *Variety* of a language, such as American English or British English, or is it representing a *variety* of a language such as legal English or academic English. A *Variety* is defined geographically and is 'user-related' (Quirk 1995: 23), where an individual is in a sense 'tied' to, and identified by, the Variety. Therefore, Irish people speak Irish English, and this includes its corresponding dialects. On the other hand, a *variety* is defined

situationally and is 'use-related' (ibid.); therefore, it involves the discourse activity the individual is involved in or the purpose for which he/she is using language. Therefore, a conversation between two academics could feature two language Varieties, say American English and Irish English, but one language variety, academic English. Indeed, many recent corpora constructed to represent a Variety of a language are built using a range of varieties of that language (see, for example, CANCODE or ICE). The decision made to choose between Variety and variety will be largely based on the research questions the corpus is expected to answer. This fundamental choice also has defining repercussions in relation to issues of corpus design such as the construction of the corpus sampling frame, which in turn has implications on corpus size, diversity of texts selected and corpus representativeness.

2. Issues of corpus design for a variety of a language

Building a corpus for either a Variety or a variety of a language involves building something that is representative of a whole; therefore, the design of the corpus is of particular importance to the corpus builders (see Reppen, this volume). Many of the decisions made by the corpus builder(s) in the design stage are based on the proposed uses of the corpus and on the research questions that these entail (see Koester, this volume). However, as McEnery *et al.* (2006: 73) caution, 'corpus building is of necessity a marriage of perfection and pragmatism'. Although the corpus builder(s) should always strive to build the perfectly representative corpus, issues such as corpus size, text diversity and number and length of texts, as outlined in this section, may result in the corpus builder(s) making decisions based on factors that are outside their control.

Issue 1: Address corpus size

In general, the primary issue connected to corpus size is that of resources, and it is here that the corpus builder(s) may have to 'cut their coat according to their cloth'. In the design stage, the corpus builder has to consider the issue of the amount of time it will take to collect, computerise, annotate and, if required, tag and parse the corpus. One of the fundamental decisions that must also be made is whether the corpus will consist of written texts or spoken texts or both. Chafe *et al.* (1991) observe that it takes six person-hours to transcribe one minute of speech for the Santa Barbara Corpus. McCarthy (1998: 12) maintains that it takes, on average twenty hours to transcribe one hour of recorded spoken data, and, 'even then, there will inevitably be inaudible segments and segments undecipherable even to the original speakers'. Estimates for the American component of ICE range from ten hours to transcribe a 2,000-word carefully prepared monologue to twenty hours for a dialogue containing numerous speaker overlaps (Meyer 2002). For this reason, corpora such as CANCODE (exclusively spoken texts) and the BNC (10 per cent spoken text) have required considerable funding both from universities and major publishing houses in Britain. Written texts can also prove problematic when building a corpus, especially when issues of copyright are considered (see Atkins *et al.* 1992: 4; McEnery *et al.* 2006: 77–9).

Biber (1990) maintains that the underlying parameters of linguistic variation can be replicated in a relatively small corpus, if that corpus represents the full range of variation. In contrast, larger corpora are not adequate for overall analyses of textual variation if they

fail to represent the range of variation. Biber (1993) examines statistical formulae for determining sample size based on a normal distribution of grammatical features such as nouns in 481 spoken and written texts (taken from Biber 1988: 77–8). He found that, for nouns, a sample of 59.8 × 2,000 word texts (approximately 120,000 words) would be required for representativeness; however, for less common grammatical features such as conditional clauses, a sample of 1,190 × 2,000 texts would be required (approximately 2.4 million words). Meyer (2002) points out that, in general, the lengthier the corpus, the better. Similarly, Biber (1993) claims that the most conservative approach to designing a corpus is to design one that represents the most widely varying feature (in this case, the conditional clause); see also Handford, this volume.

The answer to how big a corpus should be in order to represent a language Variety, or indeed a language variety, is also strongly linked to the purpose of the corpus. For example, the Bank of English is a 450-million-word corpus of 'standard' British English designed for lexicographic purposes and is, therefore, by necessity a 'mega-corpus' (see Walter, this volume). However, in terms of a Variety of language, this corpus makes no attempt to account for regional or social variability in Britain. The BNC has 100 million words and the spoken component is demographically sampled (see Crowdy 1993 for a full outline of the process of demographic sampling undertaken by the BNC). This makes the BNC a useful resource for a wide range of research purposes. In contrast, LCIE is a one-million-word spoken corpus of Irish English designed to describe the lexico-grammatical features of the Variety which, as Section 4 will show, is a task that can be accomplished using a much smaller corpus. In terms of corpus size and corpora constructed to represent a variety of a language, the Michigan Corpus of Spoken Academic English (MICASE), designed to examine the characteristics of contemporary American academic speech, has approximately 1.8 million words. In addition to this, two of the original corpora built to represent a written Variety of a language, the Brown Corpus and the LOB corpus, are one million words each in size.

Therefore, when building the corpus, the corpus builder(s) must carefully consider issues of purpose and resources. A comprehensive examination of the lexicon of a given Variety of a language would require a large corpus such as the BNC and this corpus would also suffice to explore rarer grammatical features. A corpus used to explore a spoken Variety (or variety) is generally smaller, because of the difficulties associated with data collection and transcription. Similarly, a corpus used to account for lexico-grammatical features can be as small as one million words. The primary issue is that the corpus be as representative as possible within the allocated resources. Atkins *et al.* (1992) argue that overambition could turn out to be unsustainable, and this is particularly relevant to building a corpus to represent a Variety of a language. According to them, 'experience teaches us that it is better to aim to record initially an essential set of attributes and values which may be expanded if resources permit' (p. 6). Meyer (2002: 34) echoes this view, suggesting that, ultimately, the size of a corpus might be better determined 'not by focusing too intently on the overall length of the corpus but by focusing more on the internal structure of the corpus'. The internal structure of a corpus refers to matters such as diversity of texts, length of texts and number of texts to include.

Issue 2: Consider the diversity of texts to include

Many corpora representative of a Variety of a language have been, in essence, multi-purpose. They can, for example, be used to describe the lexical and grammatical features

of the Variety they represent, to study the differences between them and other national Varieties or to study variation within the different registers/genres that comprise the corpus. Therefore, a corpus of this type necessarily requires a wide range of texts. The Brown Corpus' sampling frame was derived from the collection of books and periodicals in the Brown University Library and Providence Athenaeum in 1961. The LOB corpus chose two sampling frames; for books, the publications listed for 1961 in *The British National Bibliographic Cumulated Subject Index, 1960–1964*, and for periodicals and newspapers, those listed in *Willing's Press Guide* (1961) (see Johansson *et al.* 1978). In terms of diversity, the BNC consists of 90 per cent written texts and 10 per cent spoken texts. The written texts were collected under three criteria: *domain*, *time* and *medium*. *Domain* refers to the context-type of the text (the BNC identified nine different context-types: for example, leisure, applied science, world affairs), *time* refers to when the texts were produced (the BNC sampled texts in the period 1960–93) and *medium* refers to the type of text publication (book, journal, newspaper, etc.). One part of the spoken part of the corpus was collected by a process of demographic sampling. Texts were collected from individuals and demographic information such as name, age, occupation, sex and social class was noted. This was further subdivided into region and interaction type (monologue or dialogue). The demographically sampled corpus was complemented by texts collected on context-governed criteria. These texts related to more formal speech contexts such as those encountered in educational or business settings (see Aston and Burnard 1998 for a full description of the design of the BNC).

The ICE corpus, which is composed of 60 per cent spoken texts and 40 per cent written, contains a genre range similar in scope to that of the BNC; however, the genres are much more specifically delineated in ICE than in the BNC (see Nelson 1996; Meyer 2002: 30–8). The written segment of the ICE corpus contains both printed and non-printed (for example, student essays, social letters) material, although the printed material accounts for 75 per cent of the written corpus. From a spoken viewpoint, similar to the BNC, ICE contains 60 per cent dialogic material and 40 per cent monologic; again, these are more thoroughly specified in ICE, with dialogues divided into public and private and monologues into scripted and unscripted. In the ICE corpus, the speakers chosen were adults of eighteen years of age or older who had received a formal education through the medium of English to at least secondary school level (however, this design proved to be flexible in the case of well-known, established political leaders and radio or television broadcasters whose public status made their inclusion appropriate). Information was also recorded about sex, ethnic group, region, occupation and status in occupation and role in relation to other participants (Greenbaum 1991). MICASE also employed context-governed criteria in collecting the data. The corpus contains speech events across the major academic disciplines in a university, for example biological and health sciences, physical sciences and engineering, and humanities and the arts. However, the professional disciplines of law, medicine and dentistry were excluded. Demographic information such as age, gender, academic role and first language were also recorded.

In relation to exclusively spoken corpora that represent a Variety, in their initial corpus design phase the CANCODE team developed a set of spoken text-types to correspond to existing text typologies for the written language. They adopted what McCarthy (1998) terms a 'genre-based' approach where not only is a population of speakers targeted, but the context and environment in which the speech is produced is also taken into consideration. The framework used for CANCODE sought to combine the nature

of speaker relationship with goal-types prevalent in everyday, spoken interaction. The nature of the speaker relationship was divided into five broad contexts: *transactional*, *professional*, *pedagogical*, *socialising* and *intimate*. For each of these contexts, three goal-types were identified; *information provision*, *collaborative task* and *collaborative idea* (see McCarthy 1998: 9–10 for a definition of the terms). Therefore, for example, a university lecture would take place in a pedagogical context with an information provision goal-type, whereas a family cooking together would be an example of an intimate collaborative task. This, according to McCarthy (ibid.: 9) 'offers the possibility of linking their [the data] contextual and social features directly with the lexico-grammatical "nuts and bolts" of their step-by-step creation'.

Issue 3: Address text length and number

In determining how 'long' a text should be in order to warrant inclusion in the corpus, the issue of corpus size must be returned to. Both spoken and written texts range dramatically in size from a few words (for example, a quick note to a friend) to millions of words (for example, a long novel), and therefore a relatively small corpus can be skewed by a relatively long text. This raises the question as to whether whole texts or parts of texts should be included in the corpus. Sinclair (2005) maintains that the best answer to this dilemma is to build a corpus large enough to dilute even the lengthiest text. However, if this is not practical, which it may not be through a range of factors such as resources and permission, then it is necessary to select a portion of the text. Meyer (2002: 40) maintains that 'corpus compilers should strive to include more different kinds of texts in corpora rather than lengthier text samples'. However, in selecting samples to be included in a corpus, attention must also be paid to ensure that text initial, middle and end samples are balanced (McEnery *et al.* 2006). Biber (1990) demonstrates a high level of stability across a range of linguistic features, for example pronouns, contractions, present and past tenses across 1,000-word samples of texts from the Brown and London–Lund corpora. He concludes that given this stability between 1,000-word samples, it seems safe to conclude that the 2,000-word and 5,000-word samples are reliable representatives of their respective text categories for analyses of this type. He also used three ten-text samples from five genres across the LOB and London–Lund corpora – conversations, public speeches, press reportage, academic prose and general fiction – and found that these ten-text sub-samples accurately represent the linguistic characteristics of genre categories, including both the central tendency and range of variation. He concludes that anywhere between eight and eighty texts within a given category is adequate for an analysis of linguistic variation (see also Biber 1993).

Where corpora have been constructed to represent a Variety of a language, for example in the ICE corpus, each text contains approximately 2,000 words with the ending occurring at a suitable discourse break (Greenbaum 1991). In addition to this, the ICE compilers decided that each regional corpus would be one million words; therefore, each one is comprised of 500 texts. They also decided on ten texts (20,000 words) as the minimum for each text category. Texts in the Brown and LOB corpora are also 2,000 words long, and therefore each corpus contains 500 texts. Both text length and number of texts differ across the spoken and written components of the BNC. For example, the demographically sampled part of the spoken corpus consists of 153 texts and approximately 4.2 million words, giving an average text length of approximately 27,500 words. The context-governed portion of the corpus consists of 708 texts and approximately

5.4 million words, giving an average of approximately 7,600 words per text. The CANCODE matrix of speech-genres (see McCarthy 1998: 9–10) yields fifteen cells and the initial target was to gather approximately 65,000 words per cell. Corpora built to represent a variety of a language show a similar diversity in terms of text number and length. For example, MICASE contains a total of 152 speech events ranging in type from lectures to meetings to dissertation defences to service encounters and, therefore, seeks to cover all speech which occurs in an academic setting. These speech events range in length from 19 to 178 minutes and in word count from 2,805 to 30,328 words (see Simpson *et al.* 2007).

3. Assessing the representativeness and balance of a corpus

Leech (1991: 27) maintains that a corpus is representative if 'findings based on its contents can be generalised to a larger hypothetical corpus'. Therefore, in the case of a corpus said to represent a language variety, it is in fact representative if its findings can be generalised to the said language variety (or Variety). Sinclair (2005: 4) outlines six defining steps towards achieving as representative a corpus as possible. The first four of these steps relate to the overall corpus design, such as the construction of the proposed corpus sampling frame, steps that can be dealt with in the pre-corpus building stage:

1 Decide on the structural criteria that you will use to build the corpus, and apply them to create a framework for the principal corpus components.
2 For each component draw up a comprehensive inventory of text types that are found there.
3 Put the text types in a priority order, taking into account all the factors that you think might increase or decrease the importance of a text type.
4 Estimate a target size for each text type, relating together (i) the overall target size for the component, (ii) the number of text types, (iii) the importance of each and (iv) the practicality of gathering quantities of it.
5 As the corpus takes shape, maintain comparison between the actual dimensions of the material and the original plan.
6 (Most important of all) document these steps so that users can have a reference point if they get unexpected results, and that improvements can be made on the basis of experience.

The fifth and sixth steps here are concerned with the *balance* of the corpus, something that is difficult to account for in the planning stages of the corpus but that may be done after a pilot or provisional corpus has been built. A balanced corpus relies heavily on intuition and best estimates (Atkins *et al.* 1992; Sinclair 2005; McEnery *et al.* 2006). This has led Sinclair (2005) to refer to balance as a rather vague notion but important nonetheless. However, in relation to corpora built to represent a language Variety or variety, when assessing the balance of a corpus it is useful to examine other corpora, and it is becoming increasingly popular, 'for good or ill' (McEnery *et al.* 2006: 17), to adopt an existing corpus model and, in doing so, to assume that issues of balance have been addressed. Written corpora like the Brown Corpus and the LOB are generally accepted as balanced written corpora. The BNC, despite the imbalance between the spoken and written components, is generally accepted to be a balanced corpus, the spoken

component all the more so given that it was collected using both demographic and context-governed approaches. This corpus design has been used by the American National Corpus, the Korean National Corpus and the Polish National Corpus. ICE could be considered a better example of a balanced corpus given that it is more heavily weighted in favour of spoken texts. However, a sixty–forty split like that in ICE is probably still not sufficient to represent the everyday linguistic experience of most people, who would experience much more speech than writing in their day-to-day lives. The LSWE contains four core registers (or varieties): *conversation*, *newspaper language*, *fiction* and *academic prose*. According to Biber *et al.* (1999: 25), these four were selected on the basis of balance in that they 'include a manageable number of distinctions while covering much of the range of variation in English'. For example, conversation is the register most commonly encountered by native speakers, whereas academic prose is a highly specialised register that native speakers encounter infrequently. Between these two extremes are the popular registers of newspapers and fiction. The corpus was designed to contain 5,000,000 words per register.

CANCODE, whose genre-based design was successfully adapted in the creation of LCIE (see Farr *et al.* 2004), is also considered a balanced corpus; however, this notion of balance was arrived at in a slightly different way from corpora such as the BNC. As already mentioned, the initial target for the CANCODE team was a figure of 65,000 words per cell. It was found that certain data, for example intimate conversation and business meetings, were more difficult to collect than other types because of their sensitive nature. Therefore, some cells were found to be more 'full' than others. The progress from the initial one million words to the final target of five million addressed these imbalances and attempted, where possible, to equally cover all the context types in the corpus. McCarthy (1998: 11) maintains that a fluid corpus design like that of CANCODE is essential as 'in the past, corpora have tended to become fossilised either because the initial design is rigidly and uncompromisingly held to, or because a particular numerical target has been achieved'. This notion of corpus design as fluid or organic in order to maintain balance is echoed in what Biber (1993: 255) calls the 'bottom-line' in corpus design. According to Biber, 'the parameters of a fully representative corpus cannot be determined at the outset. Rather, corpus work proceeds in a cyclical fashion' (ibid.: 255–6). Approaching corpus design in this way allows researchers to explore language change over a period of time, an important aspect of the study of any language Variety. However, in corpus linguistics in general, there exists a relative paucity of diachronic corpora, especially in the area of spoken language (see, however, Cutting 2001 for an example of a spoken diachronic corpus).

Hunston (2002: 30) contends that the real question as regards representativeness is how the balance of a corpus should be taken into account when interpreting data from that corpus. Any corpus that is built to represent a Variety and/or a variety of a language is by its nature a multi-purpose corpus, therefore, the builder(s) cannot predict all the queries that may be made of it. Thus, according to Sinclair (2005), it is necessary to document all decisions made in regard to the criteria decided upon in building the corpus. The analyst can then check this documentation to ensure that the corpus is suitable for the proposed purpose. Hunston (2002), Meyer (2002), Sinclair (2005) and McEnery *et al.* (2006) all maintain that the responsibility for corpus analysis is a shared one. Moreover, as Hunston (2002: 23) notes, 'a statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample'.

4. What can a corpus tell us about a language Variety? The case of LCIE

LCIE is a one-million-word corpus of naturally occurring spoken Irish English built to allow the description of Irish English as a Variety in itself rather than how it is similar to or different from other Varieties of English such as British English (for a full description of the design of LCIE see Farr *et al.* 2004). Because of the size of the corpus, much of the research done to date using LCIE is not simply quantitative in nature, a feature of much of the analysis relating to the 'mega-corpora', but also features a large degree of qualitative analysis. In addition to this, much of the research centred on the corpus has focused on the realm of pragmatics. This has allowed researchers working with the corpus to provide some very interesting insights into lexico-grammatical representations of socio-cultural norms in Irish society.

One area that has received a lot of attention is the use of hedging as a politeness strategy in Irish English. From a quantitative viewpoint, on a Varietal level, Farr and O'Keeffe (2002) found that the hedges *I would say* and *I'd say* are used more frequently by Irish speakers than by British or American speakers. Indeed, they discovered that Irish speakers are twice as tentative as their American counterparts. They label this initial finding 'restrictive in its insightfulness' (p. 29) because of the fact that geographically constrained frameworks do not further an understanding of how or why hedges are used in face-to-face interaction. In reaction to this, they analyse two varietal sub-corpora from LCIE, a 55,000-word corpus of radio phone-ins and a 52,000-word corpus of post-observation teacher training interaction (POTTI), in order to more thoroughly explore the use of *would* as a hedging device in an Irish institutional setting. In addition to confirming that Irish speakers soften face-threatening acts such as disagreement or giving advice, they also found that speakers would very often downtone when speaking about themselves, even where the propositional content is undisputed (*My hair would be brown* or *I'd be from Clare*, for example). They propose that, in order to fully understand why speakers hedge, it is necessary to consider the Irish socio-cultural context. They maintain that 'in Irish society, directness is very often avoided … "forwardness", which ranges from being direct to being self-promoting, is not valued' (p. 42). Therefore, Irish speakers may feel added pressure to hedge in situations where British or American speakers may think it unnecessary.

This research also points towards another socio-cultural element of Irish English, in that speakers appear to be acutely aware of asymmetrical speech relationships and often the speaker with the most power will seek to facilitate a more symmetric interaction. Hedging is one strategy that Irish people frequently use to overcome this asymmetry. Farr *et al.* (2004) analysed the occurrence of hedging across five contexts in LCIE: family discourse, teaching training feedback, service encounters, female friends chatting and radio phone-ins. They found the lowest instance of hedging occurred in service encounters where 'there is an existing social schema for the interaction within exogenous roles' (pp. 16–17), which simultaneously allows maximum transactional efficiency and minimum threat to face. The next least hedged context was the family, where although the speaker relationships are asymmetric in nature, the context of family discourse acts as a 'meta-hedge' (see Section 5 below). They also demonstrate that hedging is far more frequent in the more formal contexts of radio phone-ins and in teacher training feedback. O'Keeffe (2005) focuses on question forms in radio phone-ins and illustrates that, although many asymmetrical norms of institutional discourse apply to this context, there

is widespread downtoning of power at a lexico-grammatical level. In addition to using hedges, the presenter of the radio show employs a variety of features such as first name vocatives, latching and reflexive pronouns, as in *What are you doing with yourself nowadays?*, to create a 'pseudo-intimate' (p. 340) environment between speaker and caller. Similarly, Farr (2005) explores the use of three relational strategies present in her POTTI corpus to demonstrate how cultural features of Irish discourse serve to lessen asymmetrical speech relationships. She claims that small talk, in particular talk about health issues, is a socially typical way of establishing solidarity between speakers in this context. Furthermore, she demonstrates how shared socio-cultural references such as *muinteóir*, the Gaelic word for teacher, are a method of diluting institutional power on the part of the teacher trainer in interaction with the trainee.

Recently, LCIE has also been utilised in the area of *variational pragmatics* (see Schneider and Barron 2005). LCIE was designed as a comparable corpus to CANCODE. This allows researchers working with the corpus to address questions at both a Varietal level (Irish English versus British English) and a varietal level (variation within Irish English itself in different contexts of use). O'Keeffe and Adolphs (2008) analyse two 20,000-word corpora of casual conversation taken from LCIE and CANCODE in order to examine the differences between the use of listener response tokens by British and Irish females around the age of twenty. They found that although the discourse and pragmatic functions of these tokens remain constant across the two datasets, there are marked differences between the form and frequency of the tokens. In relation to form, tokens which have religious reference or are swear words, for example *Jesus*, are more common in Irish English. Farr (2005) has shown that these tokens occur even in the formal context of teacher training. O'Keeffe and Adolphs link the higher occurrence of tokens with religious reference to different socio-cultural norms that exist between the two societies. The higher frequency of these tokens is attributed to the continuing importance of religion in Irish society. They also demonstrate that, in terms of overall frequency, listener response tokens are 59 per cent more frequent in the British English data than in Irish English. The authors raise a number of interesting questions concerning discourse norms in both societies – for example: Are British people better listeners? Do Irish people talk more and respond less? Do Irish people yield turns less and interrupt more? – which will hopefully form the basis of much of the future work on LCIE.

5. What can a corpus tell us about a language variety? The case of Irish family discourse

As already mentioned, modern corpora built to represent a Variety of a language, such as LCIE, are constructed using a range of varieties of the language in question. Therefore, in the same way that LCIE enables researchers to describe a language Variety, smaller sub-corpora of this can also allow for descriptions of situational variation in varieties or registers. Register variation is generally associated with the work of Biber throughout the years (for example, Biber 1988, 1995). Biber *et al.* (1999) developed a matrix of situational characteristics that distinguish one register from another and this is applied to family discourse in Table 7.1. The characteristic *participant roles* (adapted from Ventola 1979) has been added to account for the unique speaker relationships that exist in this context; there exist pre-established speaker roles wherein the speakers are bound in an asymmetrical power relationship in an intimate and informal register.

Table 7.1 The situational characteristics of family discourse

The family
Register
• *Mode* – spoken: face-to-face
• *Interactive online production* – spontaneous, no advanced planning
• *Shared immediate situation* – the family home
• *Main communicative purpose/content* – personal communication
• *Audience* – private, immediate family members only
• *Participant roles* – hierarchic/asymmetrical – parents – children, sibling – sibling Fixed/stable and pre-established speaker relationship – family – father, mother, brothers, sisters
• *Dialect domain* – local: base-level dialect (Crystal 2000).

It has previously been noted that researchers working with LCIE have demonstrated that family discourse is markedly less hedged than discourse in other context types such as female friends chatting and radio phone-ins (Farr *et al.* 2004). Building on this work, the present author (Clancy 2005) used a corpus of c.12,500 words of casual conversation recorded in the home/family environment to compare the occurrences of eight hedges prominent in Irish English across two distinct context-types – family discourse and radio phone-ins. It was found that hedges occur more than twice as frequently in radio phone-ins than in family discourse, and this was again attributed to the unique nature of family discourse. For example, some hedges, such as *kind of/sort of*, function to reduce the social distance between speakers and also to indicate the speaker's desire for a relaxed relationship with the addressee (Holmes 1993: 101), something that has to be worked at in contexts such as radio phone-ins in order to create the pseudo-intimacy crucial to the success of the interaction, but that is unnecessary in the family as the speakers perceive social distance as being negligible. Furthermore, it has been shown (Clancy, in preparation) that it appears that all utterances in family discourse are 'meta-hedged' by the context itself, thereby eliminating the need for lexical realisations of the strategy. The present author contends that it may be hypothesised that the more intimate the context-type, the more direct a speaker can be and the less chance there is of participants perceiving an attack to their face. Therefore, it could be proposed that the more intimate the data the less need there is to hedge or soften utterances.

From the perspective of variational pragmatics within Irish society itself, the present author (Clancy, in preparation) has employed two datasets representing spoken language collected in the home/family environment, one from a middle-class Irish family and one from a family belonging to the Irish Travelling community, to illustrate how hedging is far more frequent in a settled family than in a Traveller family. This can be attributed to socio-cultural factors such as the primacy of the family in Traveller culture and the differing educational profiles of the two communities. It is argued that hedges such as *I think*, *like*, *you know*, *actually* and *just* represent those that are critical to politeness in 'mainstream' Irish culture. They are the absolute minimum needed for polite interaction among participants in Irish society and ensure a smooth transition from the family community of practice to the wider social world. They are in a sense 'redundant' in the Travelling community, given that they rarely move into the realm of mainstream society.

Further reading

Aston, G. and Burnard, L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Meyer, C. (2002) *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press. (This book provides an accessible introduction to corpus linguistics in addition to a step-by-step guide to corpus design, construction and analysis. Meyer draws heavily on corpora representing different Varieties of English such as the BNC and ICE in order to illustrate each stage.)

Simpson, R., Lee, D., Leicher, S. and Ädel, A. (2007) *MICASE Manual*. Available at http://lw.lsa.umich.edu/eli/micase/MICASE_MANUAL.pdf (accessed 7 November 2008). (These represent two essential guides for any researcher wishing to construct a corpus that represents either a language Variety (the BNC) or variety (MICASE)).

Schneider, K. and Barron, A. (eds) (2008) *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages*, Amsterdam: John Benjamins. (Not a corpus publication *per se*; however, there are three chapters that illustrate how corpora can be used to examine the nuances that exist between different language Varieties in a variety of contexts. O'Keeffe and Adolphs examine differences between Irish (LCIE) and British (CANCODE) English. Jautz compares British (BNC) and New Zealand (Wellington Corpus of Spoken New Zealand English) English and Plevoets *et al*. explore Netherlandic and Belgian Dutch (*Corpus Gesproken Nederlands*)).

References

Aston, G. and Burnard, L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, S., Clear, J. and Ostler, N. (1992) 'Corpus Design Criteria', *Literary and Linguistic Computing* 7(1): 1–16.

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

——(1990) 'Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation', *Literary and Linguistic Computing* 5(4): 257–69.

——(1993) 'Representativeness in Corpus Design', *Literary and Linguistic Computing* 8(4): 243–57.

——(1995) *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finnegan, E. (1999) *The Longman Grammar of Spoken and Written English*. London: Longman.

Chafe, W., Du Bois, J. and Thompson, S. (1991) 'Towards a New Corpus of American English', in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics*. London: Longman, pp. 64–82.

Cheng, W., Greaves, C. and Warren, M. (2008) *A Corpus-driven Study of Discourse Intonation*. Amsterdam: John Benjamins.

Clancy, B. (2005) '*You're Fat. You'll Eat Them All*: Politeness Strategies in Family Discourse', in A. Barron and K. Schneider (eds) *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter, pp. 177–99.

Crowdy, S. (1993) 'Spoken Corpus Design', *Literary and Linguistic Computing* 8(4): 259–65.

Crystal, D. (2000) 'Emerging Englishes', *English Teaching Professional* 14: 3–6.

——(2001) *Language and the Internet*. Cambridge: Cambridge University Press.

Cutting, J. (2001) 'The Speech Acts of the In-group', *Journal of Pragmatics* 33: 1207–33.

Farr, F. (2005) 'Relational Strategies in the Discourse of Professional Performance Review in an Irish Academic Environment: The Case of Language Teacher Education', in A. Barron and K. Schneider (eds) *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter, pp. 203–34.

Farr, F. and O'Keeffe, A. (2002) '*Would* as a Hedging Device in an Irish Context: An Intra-varietal Comparison of Institutionalised Spoken Interaction', in R. Reppen, S. Fitzmaurice and D. Biber (eds) *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins, pp. 25–48.

Farr, F., Murphy, B. and O'Keeffe, A. (2004) 'The Limerick Corpus of Irish English: Design, Description and Application', *Teanga* 21: 5–29.

Greenbaum, S. (1991) 'The Development of the International Corpus of English', in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics*. London: Longman, pp. 83–91.

Holmes, J. (1993) '"New Zealand Women are Good to Talk to": An Analysis of Politeness Strategies in Interaction', *Journal of Pragmatics* 20: 91–116.

Hudson, R. (1980) *Sociolinguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johansson, S., Leech, G. and Goodluck, H. (1978) *Manual of Information to Accompany the Lancaster/Oslo-Bergen Corpus of British English, for Use with Digital Computers*. Oslo: Department of English, University of Oslo.

Knowles, G. (1990) 'The Use of Spoken and Written Corpora in the Teaching of Linguistics', *Literary and Linguistic Computing* 5(1): 45–48.

Labov, W. (1966) *The Social Stratification of English in New York City*. Washington, DC: Centre for Applied Linguistics.

——(1972) *Sociolinguistic Patterns*. Philadelphia, PA: University of Pennsylvania Press.

Leech, G. (1991) 'The State of the Art in Corpus Linguistics', in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics*. London: Longman, pp. 8–30.

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

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

Meyer, C. (2002) *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Milroy, L. (1987) *Language and Social Networks*. Oxford: Blackwell.

Nelson, G. (1996) 'The Design of the Corpus', in S. Greenbaum (ed.) *Comparing English Worldwide: The International Corpus of English*. Oxford: Oxford University Press, pp. 27–36.

O'Keeffe, A. (2005) '*You've a Daughter Yourself?* A Corpus-based Look at Question Forms in an Irish Radio Phone-in', in A. Barron and K. Schneider (eds) *The Pragmatics of Irish English*. Berlin: Mouton de Gruyter, pp. 339–66.

O'Keeffe, A. and Adolphs, S. (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, pp. 69–98.

Quirk, R. (1995) *Grammatical and Lexical Variance in English*. London: Longman.

Schneider, K. and Barron, A. (2005) 'Variational Pragmatics: Contours of a New Discipline', unpublished paper presented at the 9th International Pragmatics Conference, Riva del Garda, 10–15 July.

Simpson, R., Lee, D., Leicher, S. and Ädel, A. (2007) *MICASE Manual*. Available at http://lw.lsa.umich.edu/eli/micase/MICASE_MANUAL.pdf (accessed 7 November 2008).

Sinclair, J. (2005) 'Corpus and Text – Basic Principles', in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, pp. 1–16; available at http://ahds.ac.uk/linguistic-corpora/ (accessed 28 October 2008).

Trudgill, W. (1974) *The Social Differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Ventola, E. (1979) 'The Structure of Casual Conversation in English', *Journal of Pragmatics* 3: 267–98.

Wolfram, W. and Schilling-Estes, N. (2006) *American English: Dialects and Variation*. Oxford: Blackwell.