Beyond the Structural Levels of Language: An Introduction to the SPICE-Ireland Corpus and its Uses[1]

John M. Kirk

1 Introduction: A pragmatically - and prosodically - annotated corpus

Linguists have long been pre-occupied with the study of meaning. They have studied individual words better to understand notions like ‘sense’ and ‘reference’ and semantic relationships between words. They have studied sentences for the truth-value entailed by the syntax used in the expression of propositions. In the last quarter-century, linguists have come to realise that what is central to the conveyance of any message is also the interpersonal and attitudinal meanings which accompany spoken utterances. That study of utterance-based communicative semantics – known as pragmatics – is concerned with meaningful communication beyond the structural levels of language and ‘at or above the level of the conversational act’ (Archer et al. 2008: 614). Pragmatics deals with the entire discourse situation: the speaker, the addressee, the topic, the locus or situation of speaking, the function or purpose of the communication, the attitude of the speaker to that communication, the intention as well as the impact of the utterance, and so on. The purpose of this article is to show how a pragmatic analysis can be applied to spoken language and, more specifically, how an annotation scheme for such analysis can be devised and incorporated into transcriptions of spoken language data.

Pragmatics had its origins in the notion of ‘how to do things with words’ (Austin 1962) and later in ‘speech acts’ (Searle 1969). Speech act strategy was quickly shown to be indirect as well as direct; and, by showing how something comes to be understood in an utterance without being explicitly stated, to involve ‘implicatures’ (Searle 1975). Communication

[1] I am most grateful to the audience of the Forum conference for their questions and comments. I am also indebted for comments on an earlier draft to John Corbett and Robert Millar.

Kirk, John M. 2013. Beyond the Structural Levels of Language: An Introduction to the SPICE-Ireland Corpus and its Uses’. In Cruickshank, Janet and Robert McColl Millar (eds.) 2013. *After the Storm: Papers from the Forum for Research on the Languages of Scotland and Ulster triennial meeting, Aberdeen 2012*. Aberdeen: Forum for Research on the Languages of Scotland and Ireland, 207-32. ISBN: 978-0-9566549-3-9

was considered to be ‘co-operative’ among speakers (Grice 1975) and to involve ‘politeness’ and regard for each other’s ‘face needs’ (Brown & Levinson 1978, 1987; Leech 1983). As markers of these communicative strategies, two features have stood out: prosody and discourse markers. Prosody, especially tone movement, often regarded as a phonetic feature at the supra-segmental level, or as an accompaniment to syntax, becomes re-analysed as an instrument for conveying intent; and discourse markers became increasingly interpreted as conveyers of pragmatic intent as well. Recent handbook[2] and textbooks[3] show a considerable consensus about the nature and importance of pragmatics in the interpretation of human utterances. However, the study of pragmatics has rarely been approached through the use of corpus-linguistic techniques (but cf. Archer et al. 2008 and Rühlemann 2010).

Corpus Linguistics,[4] has also been growing in importance through its use of corpora of usually but not invariably large or very large amounts of authentic, representative and increasingly spoken (as well as written) data.[5] Corpus linguistics has also increased in popularity because its methodology makes use of replicable, verifiable quantitative analyses to reinforce any qualitative analyses which may also be undertaken.

Following the publication of *A Comprehensive Grammar of the English Language* (Quirk et al. 1985), which attended to syntactic variation in different spoken and written registers, one of its authors, Sidney Greenbaum, saw the importance of investigating registers across national varieties of English, thereby, in 1989, inaugurating the International Corpus

[2] E.g. titles in the *Handbook of Pragmatics* series and installments (gen. eds.) J.-O. Östman and J. Verschueren) (Amsterdam: Benjamins); titles in *Handbooks of Pragmatics* series (gen. eds.) W. Bublitz, A. Jucker, and K.P. Schneider) (Berlin: Mouton de Gruyter) especially Locher and Graham 2010, Andersen and Aijmer 2011, and Bublitz and Norrick 2011; titles in the *Studies in Pragmatics* series (gen. eds. M.-B. M. Hansen, K. and A. Barron) (Bingley: Emerald) especially Aijmer and Vandenbergen 2006, Fraser and Fischer 2006, and Barth-Weingarten, Dehé and Wichmann 2009; the many monographs in the *Pragmatics and Beyond New Series* (Amsterdam: John Benjamins); also Horn and Ward 2005; and Allan and Jaszczoit 2012.

[3] E.g. Levinson 1983; Blakemore 1992; Mey 1993; Thomas 1995; Yule 1996; Watts 2003; Griffiths 2006; Cutting 2007; Grundy 2008; Cruse 2010; O’Keeffe et al. 2011. Archer et al. 2012.

[4] Recent handbooks include Lüdeling and Kytö 2008; Baker 2009; and O’Keeffe and McCarthy 2010. Recent textbooks include McEnery and Wilson 1996; Biber et al. 1998; Kennedy 1998; Meyer 2002; Baker 2006; McEnery et al. 2006; O’Keeffe et al. 2007; Anderson and Corbett 2009, Lindquist 2009; Baker 2010; McEnery and Hardie 2012; and Cheng 2012.

[5] A convenient survey is to be found in O’Keffe et al. 2007: Appendix 1.

of English project, of which Jeff Kallen and I came to produce the Ireland Component. Greenbaum explains that:

> its principal aim is to provide the resources for comparative studies of the English used in countries where it is either a majority first language (for example, Canada and Australia) or an official additional language (for example, India and Nigeria). In both language situations, English serves as a means of communication between those who live in these countries. The resources that ICE is providing for comparative studies are computer corpora, collections of samples of written and spoken English from each of the countries that are participating in the project.
>
> (Greenbaum 1996: 3)

Nelson (1996: 28) further elaborates the ICE concept in describing the social characteristics of the contributors to ICE corpora:

> The authors and speakers of the texts are aged 18 or over, and have been educated through the medium of English to at least the end of secondary schooling. We use these two criteria because they are quantifiable. We do not attempt an evaluation of the language in a text as a criterion for inclusion or exclusion. Age and education can be accurately measured, and they can be applied in the same way in every country. The project, then, is not based on any prior notion of what 'educated' or 'standard' English is.

Although we were preparing ICE-Ireland in terms of the agreed protocols for transcription and mark-up set out in Greenbaum 1996,[6] Kallen and I became aware that a conventional orthographic transcription was quite insufficient to capture the pragmatics of any spoken exchange; and that much of what is conventionally considered to be purely syntactic is inherently conditioned by pragmatic intent and made interpretable only by reference to prosody (cf. Wichmann 2010; Wichmann et al. 2009). In inaugurating the *SPICE*-Ireland project ('Systems of Pragmatic Annotation

[6] The compilation of the ICE-Ireland Corpus was funded by AHRB Research Grant No. AR12375 as part of a project entitled *Sociolinguistics of Standardisation of English in Ireland*, which ran from 2001–2003, and for which due acknowledgement is made. The beta version of 2003 was released as v. 1.2 in 2007, v. 1.2.1 in 2009 and v. 1.2.2 in 2011. For their collaboration on the project, I am most grateful to Jeff Kallen, my co-director, and to Orla Lowry and Anne Rooney, our two post-doctoral research assistants, and to Margaret Mannion, for editorial assistance. The *ICE-Ireland: A User's Guide* was published in 2008.

in the Spoken Component of ICE-Ireland'), to the spoken component of ICE-Ireland which we took as our primary database, we devised and applied an annotation scheme which covered the pragmatic, prosodic and discoursal features which we considered essential and which we wished to investigate for analysis and description. Thereby, we were seeking to overcome the limitations of traditional grammar-based approaches and describe language in its wider interpersonal context of communicative use, and to examine ways in which pragmatic intent and prosodic features may function differently even in the relatively standardised registers of the ICE corpus.[7]

1.1 The SPICE-Ireland Corpus: Aims and Objectives

One aim of the SPICE-Ireland project, briefly chronologised in footnote 9, was to make use of the valuable collection of data for the ICE-Ireland project by the integration of pragmatic and prosodic information into an expanded corpus. The corpus comprises 626,597 words, from 964 educated, adult speakers over the age of 18,[8] in two geopolitical zones (Northern Ireland: NI; and the Republic of Ireland: ROI), in 15 discourse situations, with scope for investigating the differences between those discourse situations at both macro levels (180 dialogues vs 120 monologues; 100 private vs 80 public dialogues; 50 scripted v. 70 unscripted monologues), and at context-specific levels, as Table 1 sets out:

Dialogue = 180 texts
- **Private** = 100 texts
  - Face to face conversations (FTF) = 90 texts
  - Telephone conversations (TEC) = 10 texts
- **Publi**c = 80 texts
  - Classroom discussions (CLD) = 20 texts
  - Broadcast discussions (BRD) = 20 texts

[7] The compilation of the SPICE-Ireland Corpus was funded by AHRB Research Grant No. AR 16248 as part of a project entitled *Integrating Prosody, Pragmatics and Syntax in a Corpus-based Linguistic Description of Irish Standard English*, which ran from 2003–2006, and for which due acknowledgement is made. The beta version of 2005 was released as v. 1.2.1 in 2009 and v. 1.2.2 in 2011. For their collaboration on the project, I am again extremely grateful to Jeff Kallen, my co-director, and to Orla Lowry and Anne Rooney, our two post-doctoral research assistants, and to Margaret Mannion, for editorial assistance. The *SPICE-Ireland: A User's Guide* was published in 2012. Parts of this paper are based on the project's End-of-Award Report to the AHRC in 2006. I am further grateful to Jeff Kallen for his invaluable help in preparing that document.

[8] Bio-details of each speaker is provided in Kallen & Kirk (2008: §10) and, with a couple of corrections, in Kallen & Kirk (2012: §10).

Broadcast interviews (BRI) = 10 texts
Parliamentary debates (PAD) = 10 texts
Legal cross-examination (LEC) = 10 texts
Business transactions (BUT) = 10 texts

Monologue = 120 texts
Unscripted = 70 texts
Spontaneous commentaries = 20 texts
Unscripted speeches = 30 texts
Demonstrations = 10 texts
Legal presentations = 10 texts
Scripted = 50 texts
Broadcast news = 20 texts
Broadcast talks = 20 texts
Scripted speeches = 10 texts

Table 1 – SPICE text categories
(cf. Kallen and Kirk 2008: 9, 98; Kallen and Kirk 2012: 9, 120).[9]

The transcriptions and mark-up are identical in the ICE-Ireland and SPICE-Ireland Corpora, v.1.2.2, as described in Kallen and Kirk (2008: §8 and 2012: §8). Overlapping speech is marked up by two sets of pairs of brackets: <{> … </{> denote the initiation and completion of a stretch of overlapping speech; <[> … </[> denote the initiation and completion of the utterance of a particular speaker which is overlapping with another utterance at that point.

While ICE protocols follow standard practice in corpus linguistics by abstracting grammar and lexicon from the pragmatic context of use and, in spoken language, from the prosodic domain, the raw materials on which the ICE corpus is based show that the abstractions in transferring speech to writing both neglect the role of the speaker's addressee(s) in shaping the discourse of the speaker and exclude from consideration the role of intonational and other prosodic features in the segmentation of speech and in the interpretation outside of consideration, the SPICE-Ireland project aimed to go beyond the lexical-grammatical corpus and construct a triangulated corpus in which it became possible to access information at pragmatic and prosodic levels that have a bearing on the syntactic choices

[9] The text category abbreviations will be used in the examples below.

made by speakers in particular contexts. The annotations encompassing the orthographic transcription are thus available for exploitation using the well-established corpus-linguistic methodology of quantitative as well as qualitative analysis.

A second aim of the SPICE-Ireland project was to continue exploration in the comparative use of corpora, both within Ireland and with other corpora, for the establishment or confirmation of descriptive hypotheses relating to national varieties of standard English or world Englishes. From a corpus-linguistics perspective, Irish standard English may be regarded empirically as a dynamic and variable set of linguistic items and their uses in particular registers, rather than any fixed or legislated standard. The hypothesis that standard English in Ireland varies in significant ways across the political border between Northern Ireland and the Republic of Ireland can, in fact, only be partially tested using standard grammatical means (cf. Kallen and Kirk 2008: §3.2; Kirk and Kallen, 2011: §1). In the process of constructing the ICE corpus, it became evident that pragmatic and stylistic devices such as the use of non-literal speech acts (irony, understatement and overstatement, and other ways of flouting Gricean conversational maxims) varied across political borders. Yet, without a systematic accounting for the relationship between syntax and pragmatics, it had become impossible to test such a hypothesis. Moreover, there are major prosodic differences in different parts of Ireland, and these in turn interact with syntax and pragmatics: unless and until a methodology for tracking such correlations within the corpus had been developed, it would have been impossible to capture vital knowledge of language use which exists in the mind of every native speaker.

The aim of the SPICE-Ireland project was not merely methodological and analytical. A third aim was to provide an electronic database which goes beyond the purely lexical-grammatical and displays necessary pragmatic and prosodic features which condition the texts of the corpus. As a machine-readable corpus, the SPICE-Ireland Corpus provides a unique and pioneering resource in its own right, in addition to providing a valuable model for other such corpora.

1.2 The SPICE-Ireland Corpus: Issues in Compilation and Interpretation

The fundamental descriptive problem which the SPICE-Ireland project addressed arose from the distortion imposed by the transfer of speech to writing in the course of linguistic analysis. Even in registers which appear relatively standardised, such as broadcasts and other institutional discourses like education of the law, speakers do not speak according to the patterns of written language, which is usually characterised by a clear linear order of words and the unambiguous marking of clause and sentence boundaries. The unplanned discourse of speech places enormous demands of linguistic processing on both speaker and listener, and the syntax found in a corpus of speech reflects these demands far more than is commonly supposed. The use of transcripts without accompanying pragmatic information also creates a distortion, since it implies that all uses of any particular syntactic structure, collocation, or lexical item are fundamentally equal. Yet all such uses are not equal: the crucial role of the pragmatic context of utterance in conditioning syntactic and lexical choice renders unrealistic an analysis based on syntactic frequencies alone. Even the division of spoken language into sentences and clauses within a text transcript creates an analysis which is better suited to writing than to speech. Segmentation into clause and sentence is often impossible on the basis of grammar alone, nor are pauses explicit markers of grammatical boundaries: transcription often relies, if only intuitively, on prosodic features. By extension, if a corpus does not make the role of prosody explicit, it runs the risk of defining the fundamental units of analysis according to the preconceived rules of written language rather than reflecting the linguistic behaviour of the speaker.

In order to overcome these shortcomings in conventional corpora, the SPICE-Ireland project built on the spoken material of the ICE-Ireland Corpus to augment standard corpus mark-up by recognising the inter-relationships in pragmatics, syntax, and prosody and making them explicit through annotation. Given the fundamental problem of the transfer of speech to writing, the solution which the SPICE-Ireland project proposed lay in the integration of pragmatic, syntactic, and prosodic information into a single machine-readable transcription system. The ICE-Ireland Corpus data had the further advantage of comprising 15 discourse situations, as outlined above, with scope for investigating the differences between them – at both macro levels (dialogues vs monologues; private vs public dialogues; scripted v. unscripted monologues), and micro, some domain-specific, levels (cf. Kallen and Kirk 2008: 9; 98; Kallen and Kirk 2012: 9 and 120).[10] The incorporation of pragmatics and prosody into the investigation of the relationship between language and political boundaries in Ireland opens up new research questions. On the basis of material in the SPICE-Ireland

[10] The text category abbreviations will be used in the examples below.

Corpus, it is possible to demonstrate that syntactic differences across the two political jurisdictions in Ireland are in fact less salient than differences in either prosody or devices used to signal pragmatic intent. No doubt this hypothesis will be borne out by further empirical investigation.

At the time of the project's inception, in 2003, almost no work on pragmatics in Ireland had been undertaken. Since then there have been contributions by Kallen (2005a, 2005b, 2006, in press) and members of the Limerick Centre for Applied Language Studies (especially Brian Clancy, Fiona Farr, Anne O'Keeffe and Elaine Vaughan).[11] A further volume is in preparation (Amador et al. forthcoming). Data from – and papers arising from – the SPICE-Ireland Corpus will contribute further to this area. Research on pragmatics, for example, has long suggested that pragmatics and syntax are in some ways connected (as in the examination of indirect vs. direct speech acts, the use of conventional implicatures, etc.), but it has tended not to use large-scale corpora to advance the understanding of how syntax and pragmatics interact in actual cases of unplanned discourse (but cf. Miller and Weinert 1997 and Rühlemann 2007, 2010).

For pragmatics, the SPICE-Ireland project developed a corpus annotation scheme in which ICE-Ireland data was annotated with a set of tags reflecting a taxonomy of pragmatic variables. These variables cover both the illocutionary force and speech act status of utterances and their conveyance by means of direct expression, conventional implicature, conversational implicature and explicature. They also cover the role of pragmatic frames such as narration, transaction, persuasion, and direct address in conditioning the realisation of syntactic variables. Using this scheme of annotation, it became possible to understand how such phenomena as politeness, directness and humour operate in both parts of Ireland.

The prosodic and pragmatic annotations which are integrated with the lexico-syntactic transcription of ICE-Ireland to form the unique resource which is the SPICE-Ireland Corpus, have created a new single scheme with parallel tiers of representation. With this new resource, items searched for on the basis of the lexico-syntactic transcription can be displayed with their pragmatic and prosodic annotation. Conversely, it is possible to search for the exponents of pragmatic categories and display their realisations with the prosodic transcription, as the examples below show.

[11] For publication details, http://www3.ul.ie/llcc/cals/english/presentations.shtml

2 The SPICE-Ireland annotation scheme

This section introduces and explains the features of the annotation scheme devised for the SPICE-Ireland Corpus. Neither the basic transcription protocol nor the extended markup schemes for ICE attempt to indicate speakers' pragmatic intentions within the corpus (cf. Greenbaum 1996). By contrast, SPICE-Ireland encodes, in so far as possible, the speech act status of each utterance in the corpus, using a scheme that is developed from the work of Searle. Searle (1976: 10) constructs a taxonomy of what he terms 'the basic categories of illocutionary acts', paying attention especially to the ways in which these different acts reflect 'differences in the direction of fit between words and the world' (1976: 3). We often faced a stark choice in this regard: either the words are made to fit the world (as in a factual description), or the world is made to fit the words (as when the utterance of a form of words, such as 'I name this child Matilda', actually brings about a change in the non-linguistic world). Searle's taxonomy of illocutionary acts focuses on five types of speech act, labelled as *representatives*, *directives*, *commissives*, *expressives*, and *declaratives*. Searle's taxonomy is designed to illustrate systemic aspects of language, not to encode actual examples of language in use. Nevertheless, because it comprises only five main contrastive types, Searle's taxonomy provides a realistic basis on which to build a scheme of pragmatic annotation that provides for an exhaustive and explicit categorisation of the material in the SPICE-Ireland Corpus; moreover we were able to implement it across the entire corpus successfully and satisfactorily within the reasonable timeframe of the grant.

Following foundational conventions set out in Crowdy 1993; Edwards and Lampert 1993; Johansson 1995; Leech at al. 1995; Leech 1997; Greenbaum 1996; Garside et al. 1997, Thompson 2004, among others, the transcription practice in SPICE-Ireland is to mark the speech act status of an utterance with a code in angle brackets before the utterance, concluding with a backslash and the appropriate code at the end. The usual scope of an utterance for the annotation of pragmatic effect corresponds to a sentence or clause. The scope over which speech annotation applies typically begins with an utterance initiator <#> or a pause <,> that indicates a new following element, and continues to a conclusion that may be indicated by grammatical boundaries or discoursal features such as conversational overlap or self-interruption. Some grammatically-defined sentences, as with the tag questions discussed below, include more than one speech act; more rarely, a speech act may run over the course of two grammatical sentences. Though it is possible to understand some strings of words as including more than one pragmatic intention, the notation here works on a principle of exclusivity, whereby only one speech act is assigned to any given string of words. Cases which appeared ambiguous have been annotated on the most likely interpretation within the context of the conversation as a whole; utterances which cannot plausibly be linked to a particular function are so marked (see below). No simple algorithm exists for determining the speech act status of an utterance; annotation is made on the basis of detailed analysis of language in use.

The substance of our pragmatic annotation is not original, and builds on established conventions, as indicated, but its combination is unique. Searle's work has been refined by others especially Leech and Weisser (2003; also Weisser 2003) who developed a scheme for one register (telephone conversations) very thoroughly and in great detail; after experimentation, we rejected their taxonomy because it was overly detailed and cumbersome for analysis on the entire spoken ICE-Ireland subcorpus comprising 15 spoken registers and 300 texts. In any case, all proposals for a pragmatic taxonomy lead back to Searle's original taxonomy, which we concluded to be the most manageable pragmatic taxonomy over such large amounts of data not least because its simplicity gave it its unequivocal strength.[12] The five speech act types are:

<rep> … </rep>	Representatives
<dir> … </dir>	Directives
<com> … </com>	Commissives
<exp> … </exp>	Expressives
<dec> … </dec>	Declaratives

The taxonomy was refined by adding four further classifications:

<icu> … </icu>	Indeterminate conversationally-relevant unit
<soc> … </soc>	Social expression (greetings, leave-takings, etc.)
<xpa> … </xpa>	Unanalysable at pragmatic level
<…K> … </…K>	keying

<icu> relates to 'indeterminate conversationally-relevant units', such as feedback responses or signals such as *right*, *yes*, or *ok* which provide conversational coherence but are not uttered with an intended pragmatic

[12] Since the SPICE-Ireland Corpus was annotated, a notable and comprehensive overview of the pragmatic annotation schemes has appeared in Archer and Culpeper 2003 and Archer et al. 2008. In the latter volume, also relevant to the present discussion are papers by Wichmann 2008 and McCarthy & O'Keeffe 2008.

function or with any other commitments in the unfolding conversation or discourse, but which are crucial to the development of the ongoing discourse.
<xpa> classifies utterances (e.g. incomplete utterances or fragments) not included in the pragmatic analysis as they are pragmatically indecipherable.
<soc> include greetings, leavetakings, and other interactive expressions fall into this category.
<…K> marks what Goffmann (1974) has labeled 'keyings' for utterances involving humour or irony where speakers are not being literal or felicitous, and where normal conditions of language use do not apply.

2.1 The Annotation of Speech Acts

A simple but central set of research questions arising from the devising and implementation of this annotation scheme is the establishment of raw frequencies in respect of each pragmatically or prosodically encoded item, and to make comparisons between the North and South, and also between them (separately or together) and any other country for which information is available.

A primary result arising from the annotation scheme is that, among its 300 texts across 15 discourse situations, the SPICE-Ireland Corpus has a grand total of 54,612 speech acts. Details of the raw occurrences per text category, North and South, the relativized (or normalized) frequency of those occurrences per 1,000 words, again in each text category, North and South; and the percentage of each Speech Act type per text category, North and South are given in the *SPICE-Ireland User's Guide* (Kallen and Kirk 2012: Tables 5, 6 and 7), In so far as there is no such hard information hitherto, these are major, unprecedented findings.

2.1.1 Representatives <rep> … </rep>

Examples are:[13]

[13] To facilitate understanding of its corpus source, each example is prefaced with an abbreviated 'header' which has been edited at the start of each example to show, in an identifying bracket, **the geopolitical zone** (NI or ROI), **the text category** (see list above – here FTF), **the text-id** (here P1A-064) and, after the $ symbol, **the speaker id** of that particular text (here E). To re-cap, the speech act is denoted in a pair of opening and closing angle brackets – here **<rep> …. </rep>;** the symbol **<#>** denotes the start of a sentence or sentence-fragment, and <,> denotes a brief pause; the symbol **%** indicates the termination of an intonation unit, within which the vowel with a pitch change is **capitalised** and its syllable preceded by a **number**. A word suffixed by an **asterisk** is annotated as a discourse marker.

(1a) <ROI-FTF-P1A-064$E> **<rep>** We were at a badminton match the other night <{> <[> and there was this girl </[> **</rep>**

(1b) <NI-BRD-P1B-021$A> <#> **<rep>** Brian 1D’Arcy% is a regular 8brOAdcaster% and Sunday newspaper 1cOlumnist **</rep>**

(1c) <ROI-SPC-P2A-017$A> <#> **<rep>** There 1Is an 1extraOrdinary sense% <,> you ‘re 1lOOking there at the 1rOlling 1fIElds of 1nOrth County 1DUblin% <,> of 1occAsion and 1hIstory this morning Brian%@ **</rep>**

Not surprisingly, Representatives are the most frequent Speech Act type. Overall, by averaging all speech act types in each of the 15 registers, and when totaled, they amount to 65% (or almost 2 out of 3). 19% (almost 1 in 5) are Directives (Kallen and Kirk 2012: Table 7). Whereas such distributions might seem intuitively satisfying, the annotation scheme has enabled us empirically to make these calculations, it would seem, for the first time.

As each text category contains a different total number of texts, resulting in considerable variation in raw frequencies, *relativised* frequencies or ratios provide a stronger basis for comparisons. Representatives are most frequent in FTF (NI: 81 and ROI: 86 – note that these and later figures are *relativized frequencies per 1,000 words*), TEC (NI: 76 and ROI: 78), and SPC (NI: 75 and ROI: 66) and least common in DEM (NI: and ROI: 32), where it is outranked by the <dir> (NI: 42 and ROI 35); a similar pattern holds for the LEC category (figures for relativised frequencies are from Kallen & Kirk 2012: Table 6). In those categories showing the highest numbers of discrete speech acts per text, the speech acts will be relatively shorter than those in some of the other text types, thus showing that speech in formal, institutional settings have longer turns – which may be prepared – than speech in informal more personal settings, where speech is unplanned or unpremeditated and spontaneous occurs in much shorter outbursts.

In other categories where the **<rep>** is relatively less frequent, such as the LEP (NI: 33 and ROI: 32), the lack of high frequency for any other speech act indicates that the <rep> speech acts are relatively long when compared to the more interactive discourse found in the FTF and TEC categories.

Between these two extremes, there is a broad middle comprising CLD, both SCS and UNS, and the various categories of Broadcast texts. Frequencies are generally consistent between NI and ROI, with CLD, for example, showing 46 in each zone, and UNS giving frequencies of 42 (NI) and 43 (ROI).

From this range of variation, the most salient result is that the average relativized frequency overall turns out to be identical for each zone (57 per 1000 words – still from Kallen & Kirk 2012: Table 6). Bearing in mind that the ICE project is designed to investigate national varieties of English (see quotation above from Greenbaum 1996), such intra-corpus similarities as between the North and South of Ireland may indicate that the frequencies of speech act types across registers in a spoken corpus may be relatively stable. *A fortiori*, these frequencies may lend themselves as baselines for future cross-corpus comparisons.

There are three text categories with frequency scores for Representatives above the mean of 57 tokens per 1,000 words of text: FTF, SPC, and TEC.

2.1.2 Directives <dir> … </dir>

Examples are:

(2a) < ROI-DEM-P2A-056$A> <#> **<dir>** 1LOOk at your
8vEgetables% <,> **</dir>** <#> **<dir>** 1ThInk 1flAvour%
while you ’re 1Actually 1chOOsing them% and if you 1cAn
1fEEl them% and 1pIck up% and get the 1crIspest% <,>
and 1frEshest ones you 1cAn% **</dir>**

(2b) <P1B-021$D> <#> <**dir**> Now* what ’s the cause of
1thAt% </**dir**>

(2c) <ROI-BRI-P1B-048$A> <#> **<dir>** Michael 1COllins is
somebody who ’s had a 1vEry profound impact% on your
2lIfe and on your work as a 1histOrian% **</dir>**

(2d) <ROI-TEC-P1A-098$B> <#> <rep> I ’m changing 1wArds%
</rep> <#> **<dir>** You 1knOw that% **</dir>**
<$A> <#> <rep> Aye you said that <{> <[> in your text
yeah@* </[> </rep>

As for Directives, the highest frequency of the **<dir>** annotation occurs in the DEM (Demonstrations) category (NI: 42, ROI: 35 – from Kallen & Kirk 2012: Table 6), where speakers expect others to perform or undertake various tasks or activities such as baking or flower arranging either at the time of utterance or at some time in the future. In close second is the FTF category, for it is in the nature of everyday conversation to make requests, seek confirmations or clarifications, and pose questions for a wide range of purposes.

The lowest frequency for the **<dir>** falls in categories which are not interactional: BRN and BRT (with the latter scoring a ratio of only 1 in ROI and 2 in NI), UNS in NI, and in the BRN, LEP, and SPC categories in the ROI subcorpus.

It further shows that Directives vary considerably with text category: the mean of 16–17 tokens per 1,000 words is exceeded in the categories of BUT, CLD, FTF, LEC and TEC, and is especially high with DEM.

With the relatively low frequencies for the other speech act types, it is only the three categories of <rep>, <dir>, and <icu> which, in their various frequency constellations, are able to characterise different text categories.

2.1.3 Indeterminate Conversationally-relevant Unit <icu> … <icu>

Examples are:

(3a) <NI-TEC-P1A-098$A> <#> <rep> I ’m not even sure
2exActly when I ‘ll 2nEEd somebody from% </rep>
<$B> <#> **<icu>** 2Right% **</icu>**

(3b) <ROI-SCS-P2B-050$A> <#> <rep> 1MY budget target for the
E-B-2R% would not then be 1incrEAsed by making further
1pAYments% </rep> <#> <rep> 1And the assets I will
1consIder 1dispOsing of% are not in the commercial
semi-state 1bOdies% <,,> </rep>
<P2B-050$C> <#> <**icu**> Watch this space </**icu**>
<P2B-050$D> <#> **<icu**> Read my lips </**icu**>

One of the speech act categories specially created for the present annotation scheme is that of the ‘indeterminate conversationally-relevant unit’. There is considerable variation in the distribution of the <icu> annotation among the 15 spoken registers. At the top end, it is TEC which has the highest score (NI: 27; ROI: 18), where the <icu> often makes up for the absence of body language in the dislocated conversations. The FTF category ranks only third for use of the <icu> (NI: 12; ROI: 11), behind BUT (NI: 20; ROI 13 – relativised frequencies are again from Kallen & Kirk 2012: Table 6), where the urging and persuading necessary to achieve agreements or undertakings may often be accompanied by <icu> markers. However, across all 15 registers, the average ratio for the relativized frequency of <icu>s is only 6 – i.e. 6/1000 words.

2.1.4 Summary

It is only the three registers of **<rep>, <dir>, and <icu>** which, in their various frequency constellations, are able to provide significant evidence for differences of frequency among different spoken registers. Moreover, the high frequency and percentage distributions of the **<rep>** means that unless its occurrence is extremely high (as with the BRN, BRT, and SPC categories), it is the relative values of <dir> and <icu> which serve as the more discriminatory factors.

2.2 Speech Acts and Registers

Using the distributional frequencies of speech acts among spoken registers (i.e. what are here called text categories), it is possible to offer fresh profiles of each register, of which the following are indicative.

Like the other more conversational text type categories, **Face to face conversations** are largely characterised by Representatives (with relativized frequencies per 1,000 words as follows: NI: 81; ROI: 86), Directives (with relativized frequencies per 1,000 words as follows: NI: 30; ROI: 33), and the <icu> category (with relativized frequencies per 1,000 words as follows: NI: 12: ROI: 11, each set from Kallen and Kirk 2012: Table 6). The combined percentage distribution of these three speech act types accounts for 91% (NI) and 91% (ROI) (from Kallen and Kirk 2012: Table 7).

The special nature of **Demonstrations** leads to the frequent use in this text type of Directives (NI: 42; ROI: 35), which are more common than the Representatives (NI: 29; ROI: 32, each set from Kallen and Kirk 2012: Table 6) which dominate every other text type category. The combined percentage distribution of these two speech act types accounts for 93% (NI) and 95% (ROI) (from Kallen and Kirk 2012: Table 7).

Spontaneous commentaries, which largely focus on the provision of information and rarely allow for conversational interaction, show a very high frequency of Representatives, (NI: 75; ROI: 66, from Kallen and Kirk 2012: Table 6), accounting for 92% (NI) and 94% (ROI) of speech acts within the category (from Kallen and Kirk 2012: Table 7).

In line with other conversational text type categories, **Telephone conversations** are largely characterised by the presence of three speech act types: Representatives (NI: 76; ROI: 78), Directives (NI: 25; ROI: 22), and the <icu> category (NI: 27: ROI: 18) (each set from Kallen and Kirk 2012: Table 6). The combined percentage distribution of these three speech act types accounts for 89% (NI) and 90% (ROI) (from Kallen and Kirk 2012: Table 7).

These brief profiles indicate that Representatives predominate in each text category except for DEM, and constitute 90% or more of speech act annotations in the BRN, BRT, and SPC categories. Directives and the <icu> may also be prominent, to greater or lesser degrees. It is also the case that, in some text categories, there are very few texts (sometimes in each zone only 5), so that it is entirely possible for an individual speaker or group of speakers to skew such relatively small sets of figures. By contrast, the figures for FTF, with 45 texts in each zone, seem all the more robust. Because there has never been a pragmatically-annotated corpus comprising 15 spoken registers, these quantifications add a genuinely innovatory component to the characterization and profiling of spoken registers.

We refrain from speculating about these distributions on the basis of any stereo-typical text category characteristics such as spontaneity, preparedness, or scriptedness (‘written to be read’), or on whether the speech is a monologue, a genuine dialogue (literally between two people), or a polylogue (between many speakers). Nevertheless, speech act annotations open up many possibilities for the analysis of language in use. The validity of any analysis derived from ICE-Ireland rests not only on the authenticity of the data, but on standardisation measures such as the selection of text types, speakers, and text size, as set down by ICE protocols (cf. Greenbaum 1996). What emerges from the data is both consistency and variation across text categories and the speech act types.

Although some findings reported here may lend themselves to comparison with text category characteristics made on qualitative or impressionistic grounds in the past, we know of no other studies comprising such a broad range of spoken text categories (as so conveniently facilitated by an ICE-corpus) which have received a pragmatic profiling along the present quantitative lines, or with which the present results may be compared.

3 Pragmatic discourse markers

A further innovation of the SPICE-Ireland Corpus pragmatic annotation scheme is its treatment of discourse markers. Following the pioneering work of Schiffrin (1987) and later work by Stenström (1990), Aijmer (1996, 2002), and others working with the London-Lund Corpus, we have taken a broad view of the discourse marker as an element of discourse that marks the speaker's orientation towards the illocutionary core of an utterance.[14] Because of the pragmatic function of discourse markers which mark the speaker's orientation towards the predication of the utterance and towards the speaker and listener, typically signaling change of topic, seeking clarification, unpacking shared knowledge, negotiating face, indicating emphasis, etc., we found it desirable to mark those items. Discourse tags are marked by an asterisk '*****' directly after the discourse marker (such items are, of course, not marked in this way if used as a lexical item). If the discourse marker comes in sentence-final position as an utterance tag, it is marked with both **@** and *****. If it occurs at the end of an utterance/intonation unit, that symbol **%** comes first.[15] The SPICE-Ireland annotation of discourse markers is not intended as a theoretical analysis, but is designed to provide text annotation using a sufficiently constrained definition of discourse marker to facilitate the comparison of similar expressions of similar functions within ICE-Ireland and across English-language corpora more generally.

Discourse markers can be used to signal the speaker's commitment to the illocutionary core of the utterance (truth in the case of representatives, intention in the case of Searle's *commissives*, etc.), which can be hedged or emphasised in various ways. They can also be used to indicate other aspects of the unfolding discourse, such as the relationship between speaker and listener, topic change, the status of information as either shared or novel, clarification, the provision of supporting evidence, and so on. In the SPICE-Ireland annotation, items whose context of usage suggests that they are used purely as elements of vocal performance (especially fillers with no apparent relation to an utterance with illocutionary force, etc.) are not classed as discourse markers.

The fundamental question in deciding whether or not to consider a particular element as a discourse marker in SPICE-Ireland lies in its

[14] Here is not the place to rehearse the various names which have been used for 'discourse markers' in recent times or the associated definitions, but see Fraser 1993 and, most recently, Aijmer and Simon-Vandenbergen 2011.

[15] As the examples implicitly show, the annotation of discourse markers takes several forms: *just** (simple discourse marker), *just@** (discourse marker which occurs as an utterance/sentence tag), *just%** (discourse marker which occurs as final word in intonation unit), and *just%@** (discourse marker which is a tag and also occurs as final word in intonation unit). Also: discourse markers containing more than one word are hyphenated, e.g. *you-know*, *kind-of*, *I-don't-know*, *oh-no*, *yeah-yeah*, etc.

contribution to the illocutionary core (predication, directive, commitment, etc.) of the utterance. For the purposes of annotation as a discourse marker, it was considered that discourse markers do not contribute to the predication or other core function of an utterance, but express the speaker's attitude towards this core illocution within the context of emerging discourse. In short, the SPICE-Ireland annotation scheme works on the principle that the status of a word as a discourse marker is not based on any inherent quality of the word itself, but on the way in which the word is used within a stretch of discourse. This approach is problematical for a purely machine-based analysis, and the annotations of SPICE-Ireland are based on the detailed analysis of words in context. The annotation thus makes a distinction between words that are used as discourse markers and the same words when used as lexical items or in other ways.

As Table 2 shows, the SPICE-Ireland Corpus has 14,472 occurrences or tokens of discourse markers. These are subdivided between three main types: lexical (e.g. *like*), syntactic (e.g. *I don't know*) and phonological (e.g. *oh*), and between Northern Ireland (ICE-NI) and the Republic of Ireland (ICE-ROI).[16] In their aggregation, discourse markers are split 50-50 between ICE-NI and ICE-ROI. Among the three structural types, lexical discourse markers predominate (70%). The amazing exactness of a 50-50 distribution North-South is to be explained by the universal functionality of discourse markers marking attitude or stance towards proposition. In encoding interpersonal and attitudinal meaning, as with the use of speech acts, speakers in each part of Ireland appear to be behaving in remarkably identical ways.[17]

[16] Further details are contained in *SPICE-Ireland: A User's Guide* (Kallen and Kirk 2012) §11–§14, and Part D.

[17] In the space available here, it is not possible to discuss individual discourse markers, of which an example is Kirk (in press).

	ICE-NI		**ICE-ROI**		**IRL**	
	N	%	N	%	N	%
Lexical	4483	47	5122	53	9605	70%
Syntactical	2351	55	1949	45	4300	26%
Phonological	350	62	217	38	567	4%
Total	**7184**	**50**	**7288**	**50**	**14472**	**100**

Table 2 – Discourse markers in SPICE-Ireland

4 Conclusion

The study of spoken language has come a long way in the last fifty years. Nevertheless, the central methodology of an orthographic transcription whereby speech is transferred to writing using certain conventions (and typified in the references above) is under fresh challenge. Whereas a transcription may record the content of any authentic communication in terms of its lexico-grammatical expression, and also such productive markers as hesitations, pauses, false starts, incomplete words or propositions, or simultaneous and overlapping speech, transcriptions have spectacularly failed to capture the speaker's pragmatic intent, his negotiation with the face needs of the addressee as well as himself, or anything else acting at or beyond the conversational act itself.

This paper shows how it is possible to annotate pragmatically and prosodically, and to search and investigate the elements of that annotation scheme for the insights into the operation and management of human interaction and communication. The SPICE-Ireland Corpus accommodates more than qualitative analysis: its huge innovation is the provision of quantitative distributional information regarding and pragmatic and prosodic functions across the formal dimensions of order within the corpus: geopolitical zones, and text categories (registers). In addition, there is scope for unraveling the social identities of the speakers, by zone, sex, age, and also, if desired, job or profession, level of education, knowledge of languages, etc. (cf. the biodata listed in Kallen and Kirk 2012: Part C). The paper also provides some reflections over the pioneering enterprise. As the pragmatic functionality of discourse markers is universal, resulting in remarkably identical frequencies between Northern Ireland and the Republic of Ireland data, so those distributional results may serve as milestones for comparison and also as stimuli for comparative studies. If the ICE-GB Corpus were tagged and analysed in the same way, how similar would the frequencies be with the SPICE-Ireland Corpus?

In the SPICE-Ireland Corpus, the rich pragmatic and prosodic manually-coded annotation scheme goes well beyond the structural levels of language so familiar from orthographic transcriptions to establish both a methodology and a paradigm for helping analysts get a lot closer to the purpose and effect in the original communicative exchanges – as well as a set of primary analyses in themselves, now ripe for secondary analysis and systematic interpretation. In this way, the gulf between corpus linguistics and pragmatics is at last being bridged; and the hope is to build in further social and situational features into the scheme to creative a truly variationist pragmatics (cf. Schneider and Barron 2008; Barron and Schneider 2009). With similar aims in mind, a further approach is by multi-modal analysis, now becoming increasingly well established (cf. Adolphs, 2008; Adolphs and Carter 2013). Each is based on an annotation scheme applied across different types of spoken data, with the additional benefit of diagnosing register variation within and across the spoken language as a whole. Together, such developments in pragmatic annotation look certain to set the agenda for what Knight et al. (2009) call 'the third generation corpora' in the future, with the makings of that 'gold standard' for pragmatics which Archer et al (2008: 638) so fervently seek.

References

Adolphs, S. 2008. *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: John Benjamins.

Adolphs, S. and R. Carter. 2013. *Spoken Corpus Linguistics. From Monomodal to Multimodal.* London: Routledge.

Aijmer, K. 1996. Conversational Routines in English: Convention and Creativity. London: Longman.

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

Aijmer, K. 2002b. 'the "adjuster" sort of', in Aijmer 2002a: 175–209.

Aijmer K. and A.-M. Simon-Vandenbergen (eds.) 2006. *Pragmatic Markers in Contrast.* Studies in Pragmatics, vol. 2. Amsterdam: Elsevier Science.

Aijmer K. and A.-M. Simon-Vandenbergen, 2011. 'Pragmatic Markers'. In (eds.) Zienkowski, Jan, Jan-Ola Östman and Jef Verschueren. *Discursive Pragmatics*. Amsterdam: John Benjamins. 223–247.

Allan, K. and K.M. Jaszczoit. 2012. *Cambridge Handbook of Pragmatics*. Cambridge: Cambridge University Press.

Amador, C., McCafferty, K. and E. Vaughan (eds.) forthcoming. *The Pragmatics of Irish English*. Amsterdam: John Benjamins.

Andersen, G. and K. Aijmer 2011. Pragmatics of Society. Handbooks of Pragamtics series, vol. 5 (Berlin: Mouton de Gruyter).

Anderson, W. and J. Corbett. 2009. *Exploring English with Online Corpora*. Houndmills: Palgrave Macmillan.

Archer, Dawn and Jonathan Culpeper. 2003. 'Sociopragmatic Annotation'. In (eds.) Wilson, A., Rayson, P. and A.M. McEnery. *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. 2003. Frankfurt am Main: Peter Lang. 37–58.

Archer, D., Culpeper, J. and M. Davies. 2008. 'Pragmatic Annotation'. In Lüdeling and Kytö (2008): 613–642.

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

Austin, J.L. 1962. *How to do Things with Words*. (ed.) J.O. Urmson. London: Oxford University Press.

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

Baker, P. (ed.) 2009. *Contemporary Corpus Linguistics*. London: Continuum.

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

Barron, A. and K.P. Schneider. 2005. *The Pragmatics of Irish English*. Trends in Linguistics, Studies and Monographs. vol. 164. Berlin: Mouton de Gruyter.

Barron, A. and K.P. Schneider [Guest Editors]. 2009. *Intercultural Pragmatics.* 6(4). [special issue]

Barth-Weingarten, D., Dehé N. and A. Wichmann. 2009. *Where Prosody meets Pragmatics*. Studies in Pragmatics, vol. 8. Bingley: Emerald.

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

Blakemore, D. 1992. *Understanding Utterances: Introduction to Pragmatics*. Oxford: Blackwell.

Brown, P. and S. Levinson. 1978. 'Universals in Language Usage: Politeness Phenomena'. In Goody, E.N. (ed.) *Questions and Politeness*. Cambridge: Cambridge University Press. 56–311.

Brown, P. and S. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press

Bublitz, W. and N. Norrick. 2011. *Foundations of Pragmatics*. Handbooks of Pragmatics series, vol. 1 (Berlin: Mouton de Gruyter).

Cheng, W. 2012. *Exploring Corpus Linguistics: Language in Action*. London: Routledge.

Cole, P. 1975. *Syntax and Semantics*. Vol. 9. New York: Academic Press.

Crowdy, S. 1993. 'Spoken Corpus Design'. *Literary and Linguistic Computing*. 8(4): 259–265.

Cruse, A. 2010. *Meaning in Language: An Introduction to Semantics and Pragmatics*. 3rd Edition. Oxford: Oxford University Press. [1st edn. 2004; 2nd edn. 2008]

Cutting, J. 2007. *Pragmatics and Discourse: A Resource Book for Students*. London: Routledge. [1st edn. 2002]

Dahlmann, I. and S. Adolphs. 2009. 'Spoken Corpus Analysis: Multimodal Approaches to Language Description'. In Baker (2009): 125–139.

Edwards, J.A. and M.D. Lampert. 1993. *Talking Data: Transcription and Coding in Discourse Research.* Hillsdale, NJ: Lawrence Erlbaum.

Fraser, B. 1993. 'What are Discourse Markers?' *Journal of Pragmatics.* 31. 931–952.

Fraser B. and K. Fischer. 2006. *Approaches to Discourse Particles*. Studies in Pragmatics, vol. 1. Amsterdam: Elsevier Science.

Garside, R., Leech, G. and T. McEnery, (eds.) 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Harlow: Longman.

Goffmann, E. 1974. *Frame Analysis: An Essay on the Organisation of Experience*. Boston, MA: Northwestern University Press.

Greenbaum, S. (ed.) 1996. *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon Press.

Grice, H.P. 1975. 'Logic and Conversation'. In Cole (1975): 41–57.

Griffiths, P. 2006. *An Introduction to English Semantics and Pragmatics*. Edinburgh: Edinburgh University Press.

Grundy, P. 2008. *Doing Pragmatics*. 3rd edn. London: Hodder Education. [1st edn. 1995; 2nd edn. 2000]

Horn, L. and G. Ward. 2005. *Handbook of Pragmatics*. Oxford: Wiley-Blackwell.

Johanssan, S. 1995. 'The Approach of the Text Encoding Initiative to the Encoding of Spoken Discourse'. In (eds.) Leech, G., Myers G. and J. Thomas, *Spoken English on Computer*, Harlow: Longman. 82–98.

Kallen, J.L. 2005a. 'Politeness in Ireland: "*… In Ireland, it's done without being said*"'. In (eds.) Leo Hickey and Miranda Stewart. *Politeness in Europe,* Clevedon: Multilingual Matters. 130–144.

Kallen, J.L. 2005b. 'Silence and Mitigation in Irish English Discourse'. In (eds.) Barron and Schneider (2005): 47–71.

Kallen, Jeffrey L. 2006 '*Arrah, like, you know*: The Dynamics of Discourse Marking in ICE-Ireland'. Plenary address, Sociolinguistics Symposium 16, Limerick. [Available online at http://hdl.handle.net/2262/50586]

Kallen, J.L. in press. *Dialects of English: Irish English*: Vol. 2: *The Republic of Ireland*. Berlin: Mouton de Gruyter.

Kallen, J.L. and J.M. Kirk. 2007. 'ICE-Ireland: Local Variations on Global Standards'. In (eds.) Beal, J.C., Corrigan K.P. and H.L. Moisl, *Creating and Digitizing Language Corpora*, vol. 1: *Synchronic Databases*. London: Palgrave Macmillan. 121–162.

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

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

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Harlow: Longman.

Kirk, J.M., and J.L. Kallen. 2011. ‘The Cultural Context of ICE-Ireland’. In (ed.) Hickey, R. *Researching the Languages of Ireland*. Uppsala University Press, Seria Celtica Upsaliensis. 269–290.

Kirk, J.M., Kallen, J.L., Lowry, O., Rooney, R. and M. Mannion. 2011a. International Corpus of English: Ireland Component. The ICE-Ireland Corpus: Version 1.2.2. Belfast: Queen’s University Belfast and Dublin: Trinity College Dublin. [beta version completed 2003; v. 1.2 released 2007; v. 1.2.1 released December 2009]

Kirk, J.M., Kallen, J.L., Lowry, O., Rooney, R. and M. Mannion. 2011b. *The SPICE-Ireland Corpus: Systems of Pragmatic Annotation for the Spoken Component of ICE-Ireland*. Version 1.2.2. Belfast: Queen’s University Belfast and Dublin: Trinity College Dublin. [beta version completed 2005; v. 1.2 limited released 2007]

Kirk, J.M. in press. ‘*kind-of* and *sort-of* in the SPICE-Ireland Corpus’. To appear in Amador et al. forthcoming.

Knight, D., Evans, D., Carter, R. and S. Adolphs. 2009. ‘Redrafting Corpus Development Methodologies: Blueprints for 3rd Generation “multimodal, multimedia” Corpora’. *Corpora* 4(1): 1–32.

Leech, G.N. 1983. *Principles of Pragmatics*. Harlow: Longman.

Leech, G. 1997.’Introducing Corpus Annotation’. In (eds.) Garside et al. 1997. 1–18.

Leech, G. and M. Weisser. 2003. ‘Generic Speech Act Annotation for Task-Oriented Dialogues’. In (eds.) Archer, D., Rayson, P., Wilson, A. and T. McEnery. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster: University Centre for Computer Corpus Research on Language Technical Reports 16(1). 441–446.

Leech, G., Myers, G., and J. Thomas. (eds.) 1995. *Spoken English on Computer*. Harlow: Longman.

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

Lindquist, H. 2009. *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Locher, M. and S.L. Graham. 2010. *Interpersonal Pragmatics*. Handbooks of Pragmatics series, vol. 6. Berlin: Mouton de Gruyter.

Lüdeling, A. and M. Kytö. 2008. *Corpus Linguistics: An International Handbook*. 2 vols. Berlin: Mouton de Gruyter.

Marckwardt, A.H. and R. Quirk. 1954. *A Common Language*. London: BBC.

McCarthy, M. and A. O’Keeffe. 2008. ‘Corpora and Spoken Language’, in Lüdeling and Kytö (2008): 1008–1024

McEnery, T. and A. Wilson. 2001. *Corpus Linguistics: An Introduction*. 2nd edn. Edinburgh: Edinburgh University Press. [1st edn. 1996]

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

McEnery, T. and A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mey, J.L. 1993. *Pragmatics; An Introduction*. Oxford: Blackwell.

Meyer, C.F. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Miller, J. and R. Weinert. 1997. *Spontaneous Spoken Language: Syntax and Discourse*. Oxford: Clarendon Press.

O’Keeffe, A., McCarthy, M. and R. Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

O’Keeffe, A. and M. McCarthy. (eds.) 2010. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

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

Nelson, G. 1996. ‘The Design of the Corpus’. In (ed.) Greenbaum 1996: 27–35.

Quirk R., Greenbaum, S., Leech, G. and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Harlow: Longman.

Rühlemann, C. 2007. *Conversation in Context*. London: Continuum.

Rühlemann, C. 2010. ‘What Can A Corpus Tell Us About Pragmatics?’ In (eds.) O’Keeffe and McCarthy (2010): 277–301.

Schiffrin, D. 1987. Discourse Markers. Cambridge: Cambridge University Press.

Schneider, K.P. and A. Barron. 2008. *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages*. Amsterdam: John Benjamins.

Searle, J.R. 1969. *Speech Acts*. Cambridge: Cambridge University Press.

Searle, J.R. 1975. ‘Indirect Speech Acts’. In Cole. 59–82.

Stenström, A.-B. 1990. ‘Lexical Items Peculiar to Spoken Discourse’. In Svartvik (1990): 137–175.

Svartvik, J. (ed.) 1990. *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Gleerup/Liber.

Thomas, J.A. 1995. *Meaning in Interaction: An Introduction to Pragmatics*. Harlow: Longman.

Thompson, P. 2004. ‘Spoken Language Corpora’. In (ed.) Wynne, M. *Developing Linguistic Corpora: A Guide to Good Practice.* London; Arts and Humanities Data Service. [available at http://www.ahds.ac.uk/guides/linguistic-corpora/chapter5.htm]

Watts, R.J. 2003. *Politeness*. Cambridge: Cambridge University Press.

Weisser, M. 2003. ‘SPAACy: A Semi-automated Tool for Annotating Dialogue Acts’. *International Journal of Corpus Linguistics* 8(1): 63–74.

Wichmann, A. 2008. ‘Speech Corpora and Spoken Corpora’. In Lüdeling and Kytö (2008): 187–207.

Wichmann, A. 2010. ‘Prosody and Pragmatics’. In Cummings, L. (ed.) *The Pragmatics Encyclopedia,* London: Routledge.

Wichmann, A., N. Dehé and D. Barth Weingarten. 2009. ‘Where Prosody Meets Pragmatics: Research at the Interface’. In Barth-Weingarten, Dehé and Wichmann (2009): 1–20.

Yule, G. 1996. *Pragmatics*. Oxford: Oxford University Press.