DE GRUYTER OPEN

ICAME Journal, Volume 41, 2017, DOI: 10.1515/icame-2017-0008

Annotating the ICE corpora pragmatically – preliminary issues & steps

Martin Weisser, Guangdong University of Foreign Studies

Abstract

Since the inception of the ICE project in 1990, ICE corpora have been used extensively in the investigation and comparison of varieties of English on different linguistic levels. These levels, however, have so far primarily been restricted to lexis and lexico-grammar, while relatively little has to date been achieved in the investigation of pragmatic strategies used by the speakers in these corpora. One of the main reasons for this shortcoming is a lack of suitable annotation that would make such a detailed pragmatic comparison possible. This paper will propose a suitable model and format for converting and enriching the ICE corpora with different levels of pragmatics-relevant information, as well as discussing the issues involved in this endeavour. And finally, to illustrate the feasibility of this aim, the paper will also include a small case study carried out on a number of files, pointing out how the resulting annotations could later be exploited in pragmatics research.

1 Introduction

Since the beginning of the ICE project in 1990, the ICE corpora have, as envisaged in their original conception, been used extensively in the investigation and comparison of varieties of English on a number of different linguistic levels. So far, however, these levels have more or less exclusively been restricted to the 'simpler' and more 'straightforward' ones, such as lexis and lexico-grammar, while relatively little has been achieved in terms of in-depth research into differences between the varieties represented in ICE in their use of pragmatic strategies. The one notable exception in this regard is perhaps the work carried out by John Kirk and others on the SPICE-Ireland corpus (Kirk 2013), which we shall return to in more detail soon.

The primary reason for this shortcoming can be found in the lack of a suitable level of annotation contained in the ICE corpora that would make possible

such a detailed pragmatic comparison of the different varieties. Corpus pragmatics, still a rather recent discipline itself, has for a long time been limited to the search for the relatively well known, and also relatively fixed, patterns of interaction that have commonly been the focus of traditional discourse analysis. These include, for example, discourse markers (henceforth DMs; Aijmer 2002 or selected papers in Fischer 2006), Initiation-Response-Feedback (IRF) sequences (cf. Stenström 1994; Tsui 1994), such as questions followed by answers and potential follow-ups, or perhaps certain (assumed) politeness phenomena, such as the use of *please* (Wichmann 2004) or *thank you*. A more comprehensive overview of recent work in the field is presented in Aijmer and Rühlemann (2015), but even there, most of the chapters focus on research investigating smaller lexico-grammatical units, rather than using annotated corpora that contain information about speech acts, which is what is ultimately needed in order to be able to analyse communicative strategies properly.

In this paper, I will propose a suitable model and format for achieving the aim of annotating all spoken parts of the currently available ICE corpora. In doing so, I shall also discuss some of the preliminary issues involved in carrying out this longer-term project, regarding the conversion of these corpora to a suitable format and carrying out the annotations. The proposed method for achieving the annotation goals revolves around the use of my Dialogue Annotation & Research Tool (DART; Weisser 2016b), which makes it possible to annotate corpora of dialogue data on the levels of syntax, speech-acts, semantico-pragmatics (akin to Searle's IFIDs), semantics (topics being talked about), and (surface) polarity fully automatically, and with a high degree of precision, so that the resulting annotations need only be corrected manually to some extent later.

To illustrate the feasibility and usefulness of the endeavour, the paper will also include a small case study carried out on a number of files, pointing out how the resulting annotations could later be exploited in pragmatics research on the communicative strategies employed in the different varieties covered. Readers interested in a more detailed example of a comparison between different speakers and speaker groups can consult Weisser (2016c) to understand the full potential better.

2 The proposed annotation format

In this section, I will first present a non-exhaustive overview of the proposed annotation format, and then, as and when required, add further details as part of the discussion of how the existing ICE-format may be adapted to conform with this. To facilitate readability, I will represent specific XML coding, especially the speech-act labels used in DART, `in this font format`.

As the ICE markup format itself is quite extensive, and covers both spoken and written data, though, it will not be possible to cover all of it in detail. Instead, I will discuss those levels that are most relevant to pragmatics-related annotation, and then summarise how the remaining tags may be dealt with.

The annotation format adopts what I have elsewhere (Weisser 2016a: 236) termed 'Simple XML', a highly readable and 'linguist-friendly' form of XML (eXtensible Markup Language; see World Wide Web Consortium). In other words, while it bears all the advantages of XML in its extensibility, i.e. the ability to define and add linguistic categories flexibly, unlike many other forms of XML currently used for linguistic annotation, it can easily be viewed, edited, and searched without the need to transform it into a more human-readable format first. The basic DART XML format, which is optimised for spoken-language analysis with a special emphasis on pragmatics-relevant features, is shown in Figure 1 and discussed below.

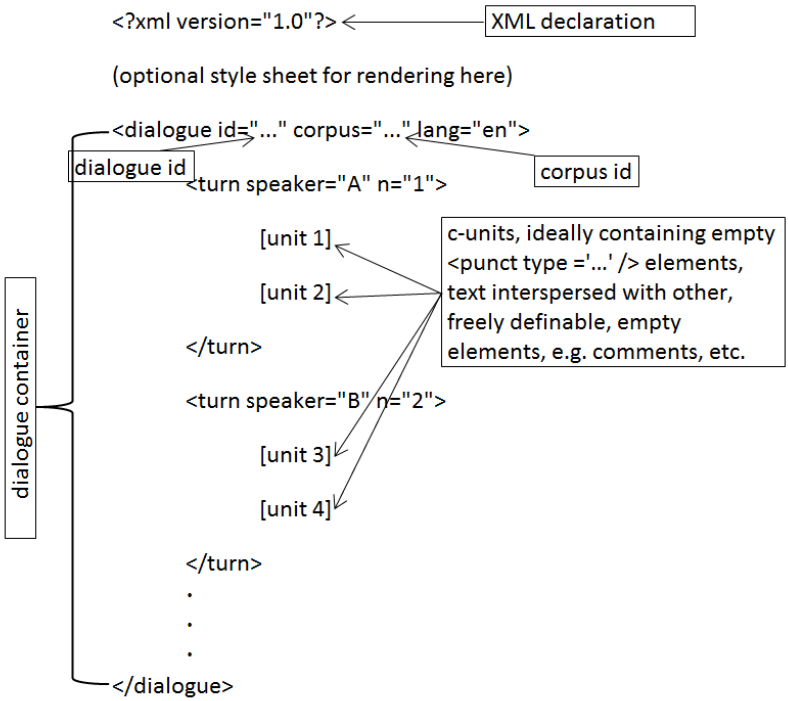

Figure 1: Basic DART XML structure, prior to annotation

The illustration above presents an example of the raw format, prior to adding specific pragmatics-relevant information through the automated annotation process. Rather than incorporating an extensive header containing meta-information, as is e.g. the case in the BNC data, the DART format encodes only minimal information about each dialogue in the form of attributes of the 'dialogue' container tag (<dialogue> … </dialogue>). As with all attributes used in DART, though, these are freely extensible, so that e.g. revision information could easily be added. The only attributes that are essential for the annotation process are the dialogue (id) and corpus identifiers (corpus), while the value of the lang attribute is assumed to be English by default if not specified.

As Figure 1 shows, below the dialogue level in the DART XML hierarchy are the individual speaker turns, each containing a speaker attribute and a unique sequential identifier (n). As far as possible, the speaker attributes should also be unique to enable speaker/group profiling or the investigation of features pertaining to speaker-related, i.e. idiosyncratic, dispersion, something that, sadly, is not the case in the current releases of ICE data, as we shall see in Section 3.1.

Inside each turn in the pre-annotated form, the text is initially split into individual lines corresponding to c-units, i.e. 'units which are independent or self-standing, in that they have no structural connection with what precedes or follows in the conversation' (Biber et al. 1999: 1069). As we shall see below, these units differ from the traditional sentence categories we are generally used to from written-language grammar, and, in the annotated version are contained in sequentially numbered (n) syntactic tags that also contain attributes for speech acts (sp-act), semantico-pragmatic markers (referred to as modes), semantic information (referred to as topics), as well as information about surface polarity. Modes here are essentially the equivalent of IFIDs (Illocutionary Force Indicating Devices) in Searle's terminology (Searle 1969: 30), and constitute interactional signals, i.e. words, phrases, or prosodic information that serve to indicate or trigger the illocutionary force of a unit. They may reflect a variety of interactional features, such as expressions of modality, conditions, reasons, signals of intent or personal preference, predictions, different degrees of conviction, etc. The reason why a syntactic tag is chosen as the container element for the unit is two-fold, a) because, in the first instance, the syntax constrains the speech-act choices available, and b) because it makes the investigation into the connection between form and function possible.

If audio files are available, the turn element could also be complemented by an attribute that contains timestamps in order to be able to find the corresponding passages there, although DART (currently) incorporates no facilities

for processing audio. Example (1) depicts an extract from a DART-annotated dialogue taken from ICE-India that illustrates the final format:

```
(1) <turn n="8" speaker="B">
    <dm n="11" sp-act="init">
    so <punc type="level" />
    </dm>
    <q-wh n="12" sp-act="reqInfo" polarity="positive"
    topic="weather" mode="open">
    how is the weather here sister <punc type="query" />
    </q-wh>
    <q-wh n="13" sp-act="reqInfo" polarity="positive"
    topic="weather" mode="open">
    how do you feel the weather here <punc type="query" /
    >
    </q-wh>
    </turn>
    <turn n="9" speaker="A">
    <decl n="14" sp-act="answer-state"
    polarity="positive" topic="weather">
    i feel very hot <punc type="stop" />
    </decl>
    </turn>
```

The syntactic categories annotated by DART comprise declaratives, interrogatives (wh- & yes-no questions), imperatives, as well as more 'unconventional' ones, such as fragments, DMs, terms of address, and yes/no-responses. Instances of three of these – <dm> (unit 11), <q-wh> (units 12 & 13), and <decl> (unit 14) – can be seen in Example (1). All units also exhibit positive surface polarity, although this is only implicit in the DM as it does not receive a `polarity` attribute. The actual text is always kept separate from the syntactic tags to ensure readability when interacting with the data.

Each c-unit receives one or more speech-act values during the automated annotation phase, unless no suitable inferencing rules are present, in which case the `sp-act` attribute remains empty and needs to be filled in manually during the post-processing phase. Any incorrectly assigned values should also be corrected during that phase.

As already pointed out earlier, DMs have previously already been identified as performing interactional functions in dialogue (cf. Schiffrin 1987 or Aijmer 2002), although the range of functions and realisations has been far more limited than in the proposed annotation scheme. Due to this functionality, DMs are treated as functional units in their own right in the DART scheme, and conse-

quently also receive a speech-act value. In Example 1, `init` signifies that speaker B initiates a new stage of the dialogue at this point, 'prefacing' her two requests for information (`reqInfo`) through the use of the DM. Aijmer (2002: 57) refers to DMs with such an initiating function as "topic changer[s]".

Unit 14 in example (1) illustrates yet another feature of the speech-act annotation in DART, the fact that speech acts can be identified as contributing either to the propositional content (`state`) of the dialogue or as constituting features of the interaction between the participants (`answer`). Thus, not only does the DART speech-act annotation go far beyond that of existing speech-act taxonomies in terms of the number and precision of labels/categories, but also adds the interactional level, enabling the corpus user to search for specific response types. An up-to-date list of all available speech-act labels used/implemented in DART can be found at http://martinweisser.org/DART_scheme.html. This list will also be extended whenever further speech-act inferencing rules and corresponding labels are added.

The `topic` attribute present in units 12–14 is often not directly pragmatics-relevant, but may occasionally indicate specific stages of a dialogue. At other times, such as in the examples here, it represents more of a 'bonus feature' which may potentially be exploited in identifying and extracting exchanges about specific topics that can e.g. be included in textbook passages introducing such topics.

The empty `<punc type="…" />` element was originally meant to help DART identify declarative questions, which would otherwise not necessarily be identifiable automatically, based on formal characteristics, even if one takes the interactional context into account. However, as the categories behind this are essentially meant to capture the prosodic functional character of the whole unit, they can also signal additional functions. Thus, an initiating DM, like the one in unit 11 above, will generally end in an intonation contour signalling non-finality, such as a level tone (hence the value). And although this is clearly an interpretative label, as is the use of punctuation marks in written texts, using such interpretative labels does in fact abstract away from the concrete prosodic realisations, which may be quite different from those that occur in the standard reference accents, especially in the case of such accents as Northern Irish English that occur in SPICE-Ireland. The use of this 'continuation marker', along with the treatment of *so* as an independent unit, here also helps to distinguish it from the use of the same word form as a logical connector. In the same way, the `<punc />` element can also serve to disambiguate between the different realisations of other DMs, such as *well*, although the existing 'punctuation' values, `stop`, `level`, `query`, `incomplete`, and `exclam`, may still need to

be expanded to cover additional attitudinal functions and prosodic contours. Other tone movements that do not coincide with c-unit boundaries can largely be deemed irrelevant for pragmatic analysis unless they are part of the stress-marking mechanism, signalling emphatic or contrastive stress. In such relatively rare cases, though, it would probably be better to insert an empty `<stress />` element, if necessary adding a `syllable` attribute to it that specifies which syllable exactly has received the 'deviant' stress. This of course also presupposes agreement on syllable divisions, which may not be straightforward, especially if the syllable concept employed in the indigenous language differs from the 'standard Western' one and the transcribers are speakers of the variety recorded.

The issue of overlap – not shown in Example (1) – is handled in a similar way to that proposed by Wong et al. (2011: 137) in their suggestions for upgrading the ICE annotation format, except that instead of the two different empty elements recommended there, `<startoverlap />` and `<endoverlap />`, the DART format uses a single `<overlap pos="…" n="…" />` tag. Here the `pos` (position) attribute values are set to `start` or `end`, respectively, and the `n` attribute can be used to indicate which overlapped passages occur in sync, especially if one speaker's turn contains parts that overlap with the preceding and following turns. In this way, it is straightforward to identify overlapping passages, but does not contravene the well-formedness criterion imposed by XML, which prohibits overlapping XML elements. An instance of this overlap marking is depicted in Example (2):

```
(2) <turn n="41" speaker="B">
    Oriya is there completely <punc type="stop" />
    <pause type="short" /> when you go to Chittoor <punc
    type="level" />
    definitely you'll find the <pause type="short" />
    southern influence is there <punc type="stop" />
    <pause type="short" /> see the word <pause
    type="short" /> completely <comment content="one
    word" /> uh <comment content="one or two words in
    Telugu" /> <punc type="level" />
    in <overlap pos="start" n="1" /> in Telugu <punc
    type="level" />
    we use it <overlap pos="end" n="1" /> <punc
    type="stop" />
    </turn>
    <turn n="42" speaker="A">
    <overlap type="pos" n="1" /> Kannada influence
```

```
<overlap pos="end" n="1" /> <punc type="stop" />
</turn>
```

Pauses are also rendered as empty XML elements, but with a `type` element that can be set to either `short`, `long`, or `medium` if no exact length has been measured. An alternative or additional attribute, `length`, may be used if more fine-grained analyses of pauses are required.

3 Formal and content issues

Having covered the most important features of the DART annotation format, we can now move on to discussing the ICE format and the necessary steps for converting the existing ICE data for pragmatics-related annotation.

The update recommendations made by Wong et al. (2011) for the general ICE markup scheme already provide a highly useful starting point for transforming the ICE-data into the kind of resource envisaged for this project. However, some of the forms of XML proposed there are still unnecessarily complex or extensive to be suitable for annotating all pragmatics-relevant features that are currently incorporated into DART. In addition, some of the suggestions made in order to make the ICE data more TEI-conformant may be considered sub-optimal.

For instance, the inclusion of a TEI header generally introduces an unwarranted overhead and may distract from interacting with the file in the post-editing process when most of the information that is generally stored in this header can easily be kept (and looked up) in external files containing such meta data (cf. Leech et al. 2000: 13). Another unfortunate recommendation is the introduction of the TEI `<u>` tag, which is also a feature of the BNC. However, this tag has always been a misnomer because *u* here stands for 'utterance' when a) the label utterance itself is ambiguous (cf. Leech et al. 2000: 56) and therefore best avoided, and b) in fact what is being marked up are generally not single (structural or functional) units, but turns, as they are defined as "contain[ing] a stretch of speech usually preceded and followed by silence or by a change of speaker" (TEI Consortium 2015: 254). In this sense, even short responses that stand on their own – rather than being embedded in the interlocutor's turn – ought to be considered turns in their own right, especially if they follow interrogatives, in which case they are, of course, no longer backchannels, anyway.

In addition, the annotation scheme(s) proposed by the TEI are too 'element-heavy', i.e. introduce especially container elements, i.e. those that contain textual content, far too readily when much of the information contained there is a) not part of the textual material (but looks as if it were), and b) could far more

easily be represented in the form of attributes or – at the very least – empty elements.[1] To illustrate this shortcoming, let us take a look at an excerpt from the TEI guidelines (TEI Consortium 2015: 254):

```
(3) <u who="#ros">yeah well I dont want to</u>
    <incident>
    <desc>toy cat has bell in tail which continues to
    make a tinkling sound</desc>
    </incident>
    <vocal who="#mar">
    <desc>meows</desc>
    </vocal>
    <u who="#ros">because it is so old</u>
```

In example (3), we have two elements that do not describe speech, but rather events that occur within speech, an <incident> and <vocal> event, both of which contain nested <desc> (for description) elements. Both of these presumably occur during the speech event, but using separate elements suggest that they are not a part of the flow and somehow interrupt the speaker's turn, which may or may not actually be the case. According such a status to these elements at the very least implies that they have the same rank as turns (<u> units), which appears odd when it is in fact the speech events that we are interested in, and such events really only become relevant in contexts where they potentially influence the speech, either because a speaker may refer to them, or they may interfere with intelligibility. Furthermore, both descriptions are also represented at the same level in the XML hierarchy as the elements that contain text, which not only makes them look equally important again, but is also probably incorrect because they represent properties of the events, in which case they ought to be represented as attributes, anyway, just like the who (i.e. speaker id) attribute represents a property of the <u> element. To avoid such potential confusion, and to accord these items of information a subordinate status without disturbing the text flow, a better solution would be to embed empty elements within the current speaker's turn, thus rendering it as in the following example:

```
(4) <u who="#ros">
    yeah well I dont want to <incident desc="toy cat has
    bell in tail which continues to make a tinkling sound
    " /> <vocal who="#mar" desc="meows" />
    because it is so old
    </u>
```

Reformatting the information not only distinguishes between the different levels of status/importance in the information, but also makes the rendering more compact, thereby improving overall legibility.

Before we can begin to understand how the pragmatics-relevant annotation of the ICE corpora itself may be carried out, we first need to take a look at two types of issues. The first essentially concerns all the corpora and relates to the ICE markup scheme versions that have been applied to them in the past, to what extent these have been applied consistently, as well as how they may affect any attempts at creating pragmatically annotated versions. The second type of issue predominantly relates to the prior attempt at carrying out the pragmatics-relevant annotation for one of the corpora, the (SP)ICE-Ireland, and concerns issues of both readability/usability and pragmatic content, although some further content-related issues may also affect other corpora.

A third issue that will only be mentioned here is that not all the corpora are currently available in any (plain-text) form that would make them amenable to the type of pragmatic annotation proposed here. For instance, the ICE-AUS is only available for online access through the Australian National Corpus (https://www.ausnc.org.au/), for ICE-USA only the written component exists so far, although parts of the Santa Barbara Corpus are supposed to be integrated as the spoken component, and ICE-GB exists only in binary form and has to be accessed through dedicated software in order to exploit the full potential of the existing annotations.

3.1 Form-related issues

In this section, I will discuss some of the more common form-related issues that may affect the proposed project for pragmatic annotation. To provide an exhaustive discussion of all the relevant issues is impossible here for reasons of space and will thus need to be the subject of a later paper. Before moving on here, though, it is probably important to stress that, at the time the plan for creating the ICE corpora was conceived, and the initial annotation scheme designed, corpus linguists' understanding of the nature of spoken data was still far less advanced than it is these days. Thus, at least some of the issues raised here may well be due to this relative inexperience and the fact that corpus compilers in the early to late 1990ies were still heavily influenced by the approaches to handling written language that they were more familiar with, and that also, once the first corpora had in fact been collected, it would already have been very difficult to change the initial design drastically.

The original SGML (Standard Generalized Markup Language) markup format of the ICE spoken data was defined in Nelson (1991), and later revised in

Nelson (2002). A sample of this markup, taken from ICE-India, can be seen in Example (5), where empty lines have been deleted in order to save space:

```
(5) <I>
    <$A>
    <ICE-IND:S1A-015#1:1:A>
    Uh you know the University
    <$B>
    <ICE-IND:S1A-015#2:1:B>
    <O> A few words </O> as it is
    <$A>
    <ICE-IND:S1A-015#3:1:A>
    Of course we speak on <O> one or two words </O>
    <$B>
    <ICE-IND:S1A-015#4:1:B>
    Uh
    <$A>
    <ICE-IND:S1A-015#5:1:A>
    So specially <,> when I heard about this Indian
    English today
    morning <,,> I felt <,,> uh that uh <,,> he did not
    specifically <,> point
    out uh <,> what exactly is Indian English <,>
    <ICE-IND:S1A-015#6:1:A>
    Because uh <,> in our room <,> there is a
    Maharashtrian <,> and uh
    there is a Punjabi <,> <}> <-> Kannadian </-> <+>
    Kannadiga </+> </}> and <,>
    <w> we've </w> some of Telugu speaking people <,>
```

Some of the ICE SGML features that can be observed in Example (5) are:

(i) the marker for the beginning of a 'subtext' (<I>), as ICE corpus files may contain multiple texts. This is generally complemented by a marker for its end (</I>) later on in the text;

(ii) speaker identifiers, <$A> & <$B> that mark the beginning of speaker turn;

(iii) text unit markers, e.g. <ICE-IND:S1A-015#1:1:A>, which encode a variety of different types of information, such as the number of the subtext, current speaker, etc;

(iv) textual comments, marked between <O> and </O>, which may refer to untranscribed portions of text that are deemed contextually irrelevant or represent '[a]nthropophonics' (Nelson 2002: 6), i.e. vocal noises;

(v) markers for long <,,> and short <,> pauses;

(vi) textual correction extent markers, <}>…</}>;

(vii) textual correction markers, <->…</-> and <+>…</+>;

(viii) orthographic word markers <w>…</w>, which enclose words that contain apostrophes.

At the time when the original decisions about the markup requirements were made, using SGML was certainly a logical choice, and many of the problems involved in using and processing it only became apparent much later. However, SGML, due to various 'shortcuts' that may be taken in the annotation, such as leaving the end of textual divisions (elements) implied (e.g. for speaker turns in text units in Example (5)), etc., has always been relatively difficult to process and also highly error-prone. This makes it problematic to ensure consistency and track annotation errors. Some of the issues caused by this have already been pointed out in Wong et al. 2011 in their discussion on how to update the annotation system to XML, which has since essentially replaced SGML as a standard in linguistic annotation. However Wong et al. primarily focussed on issues in ICE-AUS, but the problems affecting the inconsistencies between the different ICE corpora are far greater, and, in order to be able to illustrate this briefly, I shall present a number of these – based on a small investigation and attempt to create a universal converter – in order to suggest how they may be dealt with in preparation for pragmatic annotation.

The use of subtexts markers was necessitated by the fact that many of the ICE data files are composites of multiple dialogues in order to bring each file up to 2,000 words, a feature apparently adopted from the design of the earliest text corpora, the BROWN and the LOB, in order to make the data comparable. However, not only does this approach contravene Sinclair's 'whole text principle' (Sinclair 2005) by mixing texts, but also, in this day and age, should no longer be necessary, as more sophisticated techniques for norming frequencies exist, even if most of these sadly still tend to be word-based only, rather than adopting more sensible units of text as a basis. Apart from 'mixing' data from different speakers, though, such a practice also creates issues in identifying dispersion, unless the analysis software employed by the corpus user

makes it possible to extract and categorise data from the subtexts, something that most corpus analysis programs available these days do not feature. Rather than having composite documents, it therefore makes sense to keep each dialogue separate because, while it is easy to group dialogues together for analysis later, it is more difficult to extract them from composite documents. In the DART scheme, each dialogue is always a self-contained entity, which is also the approach adopted in the conversion of the ICE data. Interestingly, though, one of the later ICE corpora, ICE-Nigeria, already follows this practice, too. And, apart from the above-mentioned advantages, removing the subtext marker also reduces the number of tags required for the scheme by one.

The current format of text unit markers also appears to be partly due to the necessity of having to identify parts of composite texts. The text unit marker we encountered in Example (5), however, whenever present, introduces a high level of redundancy, as, apart from containing the unit number, it in fact repeats numerous bits of information, such as the name of the corpus, the text identifier, the number of the subtext, and finally the id of current speaker. All of these could easily be recovered if the annotation scheme were changed into a more hierarchical one, as also suggested by Wong et al. (2011: 130). Text unit markers in spoken language in ICE should ideally, and in theory, correspond roughly to the c-units marked up in DART, but in practice, generally correspond more to sentences as we know them from the representation of written language in that they subsume DMs and other short c-units that ought to be treated as having independent function. I will have more to say about this in Section 3.2.

In addition, the division into textual units appears to have been carried out fairly inconsistently in different corpora, e.g. carefully in ICE-IND, but less so in ICE-HK. Especially in the latter, some of the data contain line breaks in arbitrary places, where markup codes sometimes span multiple lines, and are therefore difficult to identify automatically without normalising the text first or using a dedicated SGML parser. In contrast, if the separation into textual units has been conducted well, even if there are arbitrary line breaks present, they may be used to identify and join textual units.

As shown in (v) above the ICE format has two separate, though similar, tags for marking pauses, one long and one short. The DART format subsumes these under one empty `<pause … />` tag, but indicates the length via a `type` or `length` attribute, again reducing the number of different tags by one, while allowing for more fine-grained distinctions. There is, however, another difference between the ways in which pauses are handled. Nelson (2002: 4) suggests: "[i]f a pause occurs between speaker turns, insert the pause at the end of the first speaker turn.". However, as it is essentially always the next speaker who

chooses to not start the turn immediately, it makes more sense to add this at the beginning of the other speaker's turn. In multi-party dialogues, there may of course be multiple speakers who opt not to take over the turn immediately. Nevertheless, even in such a case, the one who does in fact take over the turn has waited for the duration of the pause, so that this feature may be assigned to them. On the other hand, a speaker who relinquishes the turn or hands it over by asking a question is never the one who pauses. In the same way, pauses that occur within a turn should always be marked before the start of a new c-unit, as the speaker has essentially either hesitated before uttering the next unit or in order to wait for another speaker to take the turn. In that way, it also becomes easier to identify hesitation phenomena and create speaker profiles.

For other markup features, such as text correction, etc., similar mechanisms can be employed, where instead of introducing container elements that may potentially add non-existent text to the materials, this content can be stored inside appropriately labelled attributes of the empty element and thus conveniently excluded when performing word counts or carrying out other types of analysis.

The use of SGML as a markup format, coupled with a suitable lack of validation, has also allowed errors or omissions in coding to find their way into the data, e.g. in the sample taken from ICE-HK depicted in Example (6):

```
(6) <{1> <[1> <,> uhm </[1> I was born in <{2> <[2> <O>
    $A laughs </> </[2> uhm I was born in <,> a hospital
    in <.> Kow </.> in Kwun
    Tong <{3> <[3> <,>
```

In Example (6), I have highlighted the omission of the label for the end of the editorial comment marker indicating the anthropophonic produced by speaker ($)A by putting a box around it. Such an omission could easily be caught – and thereby avoided – if XML was used for the annotation, and a simple well-formedness check on the file carried out. This could be achieved very easily by opening the file in a web browser, which would then signal the error and attempt to indicate where it was located, even if frequently such error reporting is notoriously imprecise.

Another feature that may cause similar issues is the annotation of overlap. The ICE guidelines specify the use of what we might refer to as 'overlap extent markers' (<{>…</{>). However, in addition, there are also (sometimes numbered) overlap start and end markers, visible in Example (6), which are mainly used to indicate (and label) multiple overlaps that span two speaker turns. Apart from the potential of the extent marker, which is a special character that is diffi-

cult to type on some keyboards, being typed in the wrong way or confused with the normalisation extent marker we saw earlier in Example (5), this again introduces a certain level of redundancy and is potentially error-prone. In addition, the overlap extent marker spans two elements and contravenes the well-formedness requirements for XML (and, strictly speaking SGML, too), where all start and end markers need to be fully embedded to guarantee the integrity of a hierarchical structure. Handling overlap, which is a very common feature of spoken interaction, is a shared problem that has existed in markup languages for a long time. Yet it has so far not been dealt with in such a simple, uniform manner as previously shown for the DART format.

Apart from the option for allowing well-formedness checking in a simple manner, adopting XML as a markup language also has at least two further advantages, due to its encoding defaulting to the Unicode format UTF-8: Furthermore, there is no longer any need to escape what was referred to as '[u]nusable characters' in the original specifications by so-called 'character reference entities' (e.g. é to replace *é* in words of French origin), and languages, even in different character sets are now freely mixable. In addition, this also allow the annotator to include phonetic characters in a transcript in editorial comments.

In general, most of the existing corpora tend to apply the conventions set out either in Nelson 1991 or 2002, partly depending on when they were created. Over the years, some projects created a number of additional tags, though, that have sometimes found their way into the revised guidelines, such as the inclusion of the <foreign>...</foreign> tag to mark non-English text. At other times, these additions have remained confined to specific corpora, as e.g. in the case of the <ea_>...<ea/> or <ea/> tags to indicate text in an East African language in ICE-EA. Unfortunately, here, we can also see two more inconsistencies, in the slashes that mark the end of units in ICE-EA occur before the closing angle bracket, and that, for spans of text, the opening tag may contain an underscore before the closing angle bracket, despite the fact that this creates additional redundancy because all end tags are already marked by a slash. These features appear to be remnants of the original annotation model that were never fixed in any editions of the corpus.

One of the most recent corpora, ICE-Nigeria, radically departs from the general format in that it provides three versions, one as general plain text, one in a form of XML (extension eaf) with time alignment to audio-recordings, and a PoS-tagged version. None of these, though, contain any of the SGML markup we encountered above. Instead, the manual (Gut 2014: 2) refers to a number of XML tags used in the corpus, which, however, I have been unable to find in the

versions currently available in the form described there, although the values associated with their names do appear in <ANNOTATION_VALUE>…</ANNOTATION_VALUE> tags.

While all the corpora I had access to generally marked speaker turns and the units within them, ICE-NZ uses no turn, but only unit markings. Again, ICE-Nigeria also seems to differ from the rest of the corpora in that the speaker information, which is normally indicated both as part of the turn and the unit marker, only appears in the form of the PARTICIPANT attribute of the <TIER> tag. Even if the turns are marked by start tags consistently, though, their end tags are not provided and have to be inferred from a subsequent speaker change, so this format does not really represent well-formed XML. Therefore, all turns in all corpora will need to be marked by the appropriate closing element. In terms of the speaker IDs, as already indicated above, it would be best to be able to distinguish between different speakers in order to be able to handle idiosyncratic dispersion features better. As part of the re-formatting and re-annotation of the ICE corpora, it would thus be better to create unique identifiers for all speakers, based on whatever documentation is available for the individual sub-projects. Sadly, however, almost no such documentation, apart from for ICE-Nigeria, appears to currently be publicly available.

As the options for turn attributes in the DART scheme are extensible, it would, for instance, also be easy to add a `status` attribute, which, if present, could contain the value `exclude`, for turns deemed to be irrelevant to the current discourse, produced by non-indigenous speakers, or that are completely in an indigenous language. This is a feature that crops up in many dialogues in ICE-HK, where perhaps the interlocutor was a speaker of an inner-circle variety or other nationality and should thus be excluded from any analyses. This way, yet another element of the original transcription scheme could be removed, again simplifying the tag scheme and reducing the number of nested elements.

In terms of case handling, most corpora contain 'sentence-initial' capitals, which makes disambiguation between homograph proper nouns and other word classes more difficult than necessary for analysis purposes, while ICE-NZ has everything downcased, including proper nouns (e.g. *tuesday*), which does not allow users to recognise proper nouns easily. Only ICE-Nigeria seems to strike the right balance here in using lowercase characters for all words apart from proper nouns.

As the examples above have shown, the current format also contains a variety of different symbols that are not necessarily self-evident and need to be interpreted through making recourse to the manual. Although this is not so much of an issue if data is processed by the computer, in terms of interacting with the

text in order to do the annotations in the first place, as well as understanding materials presented in concordances, etc., this reduces the overall readability of the corpus materials. Unfortunately, the situation gets worse with the only one of the ICE corpora that has been enriched with any form of pragmatic annotation, the SPICE-Ireland, where a number of additional codes have been added. These tags reflect the speech-act-related markup, and are predominantly based on Searle's five categories of *representatives* (<rep> … </rep>), *directives* (<dir> … </dir>), *commissives* (<com> … </com>), *expressives* (<exp> … </exp>), and *declaratives* (<dec> … </dec>). Those tag options are augmented by three further ones, *indeterminate conversationally-relevant utterances* (<icu> … </icu>), *social expressions* (<soc> … </soc>), and *not analysable at pragmatic level* (<xpa> … </xpa>), and a tag part for *keyed utterances* (<…K> … </…K>). The latter is appended to one of the other categories to indicate that the speech act already marked up should not be interpreted literally, but may instead need to be interpreted as ironic or humorous usage. In addition, DMs have an asterisk (*) appended to the end of the word or expression, and question and declarative tags an @ symbol. The ends of intonational phrase boundaries are marked by percent signs (%), and stressed syllables capitalised, which, unfortunately, 'clashes' with the marking for sentence-initial words. Furthermore, prosodic "tunes" (Kallen and Kirk 2012: 43) are marked in the form of numerical codes (1–11) preceding the syllable that initiates the tune. Some, but not all of these features can be seen in Example (7), but for more details, see Kallen and Kirk 2012:

(7) <P1B-021$A> <#> <dec> And so to our studio 2pAnel% </dec> <#> <rep> Father Brian 2D'Arcy% <,> uh he 's a native of 2FermAnagh% but spent many years in the 1RepUblic% before he returned to Northern 1Ireland% </rep> <#> <rep> Brian 1D'Arcy% is a regular 8brOAdcaster% and Sunday newspaper 1cOlumnist% </rep> <#> <rep> Doctor Bill 2ROlston% lectures in 2sociOlogy% at the University of 8Ulster% and he has written 1extEnsively% on Northern 1Ireland% </rep> <#> <rep> And joining us from 8DUblin% is Michael 8FArrell% <,> a leading light in the civil 1rIghts movement% of the late 1sIxties% and he is now a practising 1lAwyer% </rep> <#> <dir> Now* Michael Farrell <,> do you think people would be 1surprIsed in the Republic% uh by that opinion poll survey that

```
says like* only fifty-two percent of 1CAtholics%
really want a united 1Ireland% </dir>
```

Apart from not following the recommended format for marking text units, here only indicated through the unit start tag <#>, this enriched type of display – which no doubt contains highly useful additional linguistic information – becomes even less legible and, what is worse, less searchable in standard corpus linguistics software.

To increase readability (and searchability), the empty <punc type="…" /> elements introduced in Section 2 can act as a replacements for those functionally relevant prosodic tunes in SPICE that occur at intonational phrase boundaries and signal the end of functional units. Tunes that do not have such a function may either be deleted or, in case they correlate with tone movements that indicate stress, be replaced by the <stress … /> elements referred to earlier. Using empty elements to achieve this should allow the corpus user to read and search the text far more easily, especially if the analysis program used provides means of displaying empty elements in a way that makes them easy to distinguish from the actual text. Searchability can then still remain a partial issue, as the user may need to 'interpolate' optional elements into search term patterns, as such elements might be freely interspersed with the words of the text. However, DART already provides at least partial support for this, as its n-gram analysis feature first removes empty elements, as well as fillers, but then performs the re-interpolation when concordancing on such n-grams, 'filling in the blanks' as it were. To illustrate this feature, let us look at such a tri-gram from ICE-HK, *in your holidays*, which occurs in the interrogative *will you find any summer job in your uhm holidays in Australia <punc type="query" />*, but could not be identified as a tri-gram without removing the filler *uhm* first. To recover the original version, though, DART produces the hyperlinked search term, including interpolated fillers and empty elements shown in Example (8):

(8)
```
in (?:(?:<.+?/|(?^:\b[ue][hr]?m?\b)|)
)?your (?:(?:<.+?/|(?^:\b[ue][hr]?m?\b)|)
)?holidays.
```

For user-defined concordance searches, the user currently still needs to specify a regular expression to achieve this, if necessary. However, specifying and interpolating patterns that involve empty elements is still much easier than covering all the different options of all additional markup interspersed with the words in SPICE.

3.2 Content-related issues

The preceding section has mainly focused on presenting formal issues in creating the kind of unified and consistent format necessary for carrying out large-scale pragmatic annotation of all ICE corpora. In this section, I now want to focus more on the kind of content and level of detail that ought to be represented in such a format in order to make it maximally usable for comparative research into interaction. In addition, I shall briefly discuss features related to particular varieties that may affect the automatic processing stages of the annotation.

In Example (7), we have already seen one instance of an approach towards the annotation of pragmatics-relevant detail in the sample from SPICE-Ireland. However, as useful as the marking of DMs or tags is, the inclusion of intonational phrase boundaries, at close scrutiny, appears to provide little pragmatics-relevant information unless it happens to coincide with the end of a functional unit, in which case it is generally marked by the end of a speech-act tag anyway. Returning to the issue of DMs and tags, for the former, as it has long been known that they do fulfil an independent pragmatic function (cf. e.g. Schiffrin 1987: 31ff), especially when occurring in (turn-)initial position, it would certainly be better to treat them as independent pragmatic units and also annotating their function via a `speech-act` attribute. Only marking them by a *, while it at least makes it possible to search for them, still does not highlight their functional aspects enough. With tags, in contrast, we encounter the opposite issue, as they are generally marked separately as directives (`<dir…</dir>`), even though they rarely do represent genuine directives, i.e. instructions to do something, in the strict sense. Instead, their main function in most of the dialogue materials I have analysed to date appears to be to indicate a request for confirmation. As such, they are also intricately linked to the propositional content of the unit they follow/are attached to, something that can also be seen in their 'syntactic' potential to turn declaratives into interrogatives.

The decisions to 'lump together' what essentially constitutes very different speech acts under the one label 'directive', though, stems from the fact that Kallen and Kirk, for very practical and sensible reasons, chose to adopt Searle's limited speech-act taxonomy (Searle 1975: 354–358) and only extend it slightly. To annotate all 300 texts, comprising 626,597 words (Kallen and Kirk 2012: 9), manually, using a more extensive taxonomy of speech-act tags would have required far more resources, apart from increasing the potential for inconsistency in the annotations with each additional (sub-)category, as human annotators are often likely to have differing perceptions about which category exactly to apply. Thus, since no software to aid a more extensive pragmatic annotation was available at the time, this decision is fully understandable under the given

circumstances. However, the constraint on annotating large amounts of data with a higher level of depth has now disappeared due to the availability of DART. Thus, a re-annotation of SPICE using DART would now also offer considerable potential in analysing to what extent different types of taxonomy may be useful for different research purposes, such as is the case for using different tagsets for PoS tagging (cf. Weisser 2016a: 109).

Another issue that affects the pragmatic annotation of the ICE data is code-switching, or, to be more precise, both *code-switching* and *code-mixing*, where the former is defined as switching to a language for a complete syntactic unit, while the latter occurs intra-sententially, i.e. predominantly involves the use of single words from an indigenous language, interspersed with English (Llamas et al. 2007: 208). As the first stage in the DART annotation process involves identifying the syntactic category of any c-unit to be annotated, indigenous words (other than proper nouns) occurring inside a unit or whole units in the other language will obviously influence this process. Furthermore, especially for whole non-English units, it may be impossible to identify any of the pragmatics-relevant features envisaged without first modelling the necessary linguistic resources in that particular language or to devise suitable routines for ignoring whole non-English units, so that they can be annotated manually later. This will necessitate the help of native speakers of that particular indigenous language or at least someone very familiar with it. Even if this kind of support were not available, it should at least already be possible to model DMs, terms of address, or other short formulaic sequences, such as greetings or farewells, as these can often be found relatively easily in phrase or text books. A minimal requirement for the annotation process, however, would at least be to compile a lexicon that contains all indigenous word forms, based on simple word lists, that could either be used to ignore these words in the annotation process, or, if language support is available, to add their parts of speech.

A further complication may be introduced by one of the very features the corpora were originally devised to investigate, the use of differing, non-standard, grammatical structures (see e.g. the example of a 'fake' declarative question in Example (10) below). Using the grammar implemented in DART so far, which predominantly models standard English, such structures may trigger annotation errors on the syntactic or other levels. However, what may initially look like a problem could in fact be used to an advantage in that these non-standard features can then be identified while correcting the annotation errors, and subsequently modelled in the grammar to later annotate them automatically across corpora of varieties that may share such characteristics. And, once anno-

tated, these features would then, in turn, also become countable, and could be used for comparative purposes.

4 Case study

In this section, I report on a small case study that I conducted in order to test the feasibility of the proposed project, as well as to identify how useful the resulting annotations could prove to the analysis of ICE data. To this end, I randomly selected a single, rather lengthy dialogue from ICE-India, S1A-015. Unfortunately, this dialogue was not available in audio format along with the transcribed material, which may potentially have led to some errors in my interpretation while re-formatting the data. This, however, is a problem with all of the current releases of ICE sub-corpora, apart from some short sample clips available from the ICE website. For a large-scale re-formatting and pragmatic annotation along the lines of what I am proposing here, it would thus be absolutely necessary to involve the original teams who compiled the individual sub-corpora, gain access to the original recordings, as well as possibly re-transcribe substantial parts of the dialogues, at least those that do exhibit inconsistencies.

Dialogue S1A-015 contains 71 turns; 290 units, 1,868 words in the final DART markup version, almost equally distributed between the two speakers A (146 units; 940 words) & B (144 units; 928 words). In this dialogue, the two interlocutors are discussing a lecture on Indian English that they both attended. To be able to annotate and analyse the file in DART, I first wrote a converter, designed to allow the automated conversion of as many different types of ICE data, and converted the file into XML. Errors in the automated conversion process in the next stage revealed a number of issues related to slight inadequacies in the design of the original coding scheme or the implementation of the scheme on the part of the creators of the corpus, which then necessitated some manual pre-processing prior to annotation.

The first of these issues relates to the original specifications for representing textual normalisations, where the format does not always make it clear whether the corrections are editorial or represent self-corrections by the speakers themselves. As editorial corrections according to the guidelines fulfil the function of aiding the parsing process (cf. Nelson 2002: 9), passages containing repetition, self-corrections, or hesitations are to be marked for deletion. However, since these constitute important features of spoken interaction, 'deleting' them in this way is sometimes counterproductive, and it should generally be the task of a dedicated parser to handle them appropriately. Furthermore, especially in 'non-native' data, but of course also Inner Circle varieties, such 'performance errors'

may in fact be acceptable features of a given variety. Thus, in the interest of increasing readability and consistency of the data, and because the parser built into DART can already handle dysfluency features such as complex repetitions of up to three words in a row (even containing intervening fillers) I chose to remove these markings.

Other features that are due to the rather idiosyncratic interpretation of the conventions on the part of the corpus compilers or the original transcribers are that sometimes two sequential turns were marked for the same speaker, or that overlap was treated in ways not specified in the guidelines. While the former may only potentially lead to parsing errors and cause issues for the conversion process, the latter is far more serious because it actually constitutes a clear misrepresentation of the interaction, as can be seen in Example (9), where the relevant parts are highlighted through boxes:

```
(9) <$A>
    <ICE-IND:S1A-015#11:1:A>
    They make take up for example word like uh work <,>
    <{> <[> uh
    <,> you know unless they spell it <,> sometimes they
    may not be able to read
    it up <,> even when I read
    <$B>
    <ICE-IND:S1A-015#12:1:B>
    <[> Uh </[> </{>
```

In Example (9), the backchannel uh is marked as part of speaker ($)A's turn, but then again as a separate turn and textual unit by speaker ($)B. Now, apart from the fact that this form of rendering creates the illusion that there are more turns and textual units than actually exist, it also duplicates textual content, thereby causing redundancies in lexical/frequency analyses, as well as potentially causing errors in parsing. As far as I was able to ascertain, this feature is not just an exception in the dialogue I analysed, but also occurred in a number of other dialogues, so that we can assume that it may affect at least some parts of the corpus that were transcribed by the same person.

Once these errors were eradicated, and the lines manually split or joined into c-units, I carried out an automated annotation in DART, followed by a post-processing stage to fix the annotation results.

The annotation process in DART integrates a number of different levels of processing, essentially combining morpho-syntactic tagging, shallow parsing, and an inferencing process based on the occurrence of semantico-pragmatic and/or semantic markers, as well as surface polarity. Thus, it is very difficult to

determine a general level of accuracy, such as for a morpho-syntactic tagger, where one essentially tries to determine whether all words in a text have been annotated correctly and can then simply report a percentage. However, even in the case of taggers, there may be disagreement between raters as to whether a correct tag has been assigned in a particular context, and often taggers either assign defaults (e.g. noun or verb) if they are unable to identify the correct word class or a tag that signals multiple options. DART, however, as pointed out earlier, does not assign any default speech acts in case it does not have an appropriate inferencing rule, and it is left to the human annotator to fill the annotation gaps, along with revising existing annotations in order to either correct or fine-tune them. Furthermore, while it may already be difficult for human raters to decide on the assignment of a correct PoS tag or parse tree, the ambiguity inherent in speech acts is even more difficult to eliminate, especially the more indirect a speech act is. Hence, DART was designed around the notion of assigning high-level speech acts in many cases where it is not unambiguously possible to determine an exact speech-act. Thus, in the absence of precise clues, it may assign speech acts such as `state`, `acknowledge`, or `suggest`,[2] and then the user can decide on whether more specific options, such as `stateCondition`/`stateConstraint`, `agree`, or `offer`, may be appropriate, taking further contextual clues into account. It therefore facilitates the pragmatic annotation process by automating the basic annotation processes that would not only potentially take many days to carry out 'manually', as well as being highly error-prone, due to human inter-rater and individual inconsistency and other factors (e.g. increasing tiredness and lack of concentration), but at the same time provides sensible high-level or precise suggestions for applicable speech acts. Nevertheless, despite having said that it is difficult to judge DART's accuracy, before continuing this case study, I want to provide a brief evaluation of its performance on the particular dialogue under discussion.

The initial annotation process, prior to post-processing, identified the following syntactic elements listed in Table 1, where they are contrasted with the post-processed version:

Table 1: Comparison initial and post-processed annotations (syntax)

syntactic category	# initial annotation	# post-processed
decl(arative)s	98	136
frag(ment)s	122	80
DMs	52	55
imp(erative)s	7	2
yes(-responses)	11	11
no(-responses)	2	2
q-wh	4	5
total	**296**	**291**

The results in Table 1 indicate that the built-in parser already has a very high degree of precision when it comes to identifying and/or splitting off lexico-grammatical patterns related to DMs, yes/no-responses or questions. The mismatch between declaratives and fragments is potentially due to missing syntax rules or non-standard grammar, while the apparent discrepancy that can be observed in the (mis-)identification of imperatives is often due to elliptical structures that contain initial (elliptical) pro-drop, so that the structures only appear to be imperatives. The higher number of units identified by DART for the initial annotation may be due to the fact that the program automatically splits off shorter syntactic units, such as DMs, and yes/no-responses, but sometimes does so mistakenly because their form may be indistinguishable from the beginning of declaratives, etc.

Out of the 296 units identified in the automatic annotation, only 38 (12.84%) did not receive any speech-act attributes, due to the absence of suitable inference rules, and thus most had to be supplied during the post-processing phase. However, this number would probably still have to be reduced, as the initial annotation identified five more units than were actually present, so that some syntactic elements had to be re-combined. All of the missing speech-act attributes occurred in either declaratives (10) or fragments (28), where the potential for employing a greater variety of speech acts related to expressing, stating, referring, or predicting exists, while the speech-act potential for the other syntactic categories is narrower or more clearly defined through their surface structure. For a more in-depth discussion of the speech-act categories and comparison with other annotation formats, see Weisser (2015).

In the following, I will now continue to discuss the results of the analyses of the annotated dialogue in terms of the communicative strategies exhibited by the speakers, along with pointing out some further issues that I identified during the post-processing stage. The discussion will be divided into two parts, first looking at the predominantly 'information-bearing' syntactic categories of declaratives and fragments, together with their associated speech acts, followed by the more interactional categories. Since the number of units produced by both speakers is very close, no frequency norming has been carried out on the data.

Table 2: Frequencies of speech-act realisations for declaratives and fragments

syntax	**speech act**	**frequency A**	**frequency B**
decl	(elab-)state	28	34
decl	report	6	6
decl	refer	6	2
decl	stateReason	5	3
decl	stateReason-expressPossibility	1	0
decl	stateCondition	3	2
decl	stateConstraint	3	2
decl	stateIntent	1	0
decl	expressOpinion	4	5
decl	expressPossibility	4	4
decl	echo-expressPossibility	0	1
decl	expressImPossibility	3	2
decl	suggest	0	2
decl	expressAwareness	2	2
decl	expressNonAwareness	0	1
decl	agree	2	0
decl	reqInfo	1	0
frag	(elab-)refer	26	22
frag	state	5	7
frag	stateOpt	1	1
frag	stateReason	1	1
frag	report	0	1
frag	predict	0	1

syntax	speech act	frequency A	frequency B
frag	expressOpinion	1	1
frag	expressPossibility	0	1
frag	echo	0	1
frag	echo-refer	2	0
frag	reqConfirm	1	0

As Table 2 shows, the declarative units predominantly `state` or `report` facts. The distinction between these two adopted here is that the former expresses a proposition in the present tense (e.g. *they make take up for example word like uh work*), while the latter is essentially its past-tense counterpart, referring to events or conditions that happened/pertained in the past (e.g. *even uh <pause type="long" /> R K Narayan was writing*). Referring expressions in declarative form provide mainly temporal or conditional contextualisation, as in e.g. *sometimes when we use the word like <pause type="long" /> uh master*. The remaining stating declaratives, apart from `stateIntent`, tend to provide more precise descriptions in terms of explicitly providing underlying reasons, conditions or constraints. To provide just a few examples, the unit expressing the dual speech act `stateReason-expressPossibility` is *there may be some slight changes because <pause type="long" /> their mother tongue is uh <punc type="incomplete" />*, where the modal expression expresses the possibility and the subordinate clause the reason, `stateConstraint` is, for instance expressed through *we are not authorised to bring <pause type="short" /> changes in grammar and all those things*, and the dual act `echo-expressPossibility` in *<comment content="pro-drop" /> accent may be different*, where, apart from the probably obvious modal expression, we also find an instance of an `echo`ing speech act realised through the fact that the speaker repeats part of what the prior speaker has said. This interactional feature is often used by speakers in order to either confirm emphatically/agree with what the interlocutor has said or may be used as a strategic device for verifying that information has been received correctly. In our example, we in fact have a case of the former because the previous speaker says that accent may be different, which is repeated by the interlocutor, but not verbatim, as the subject expressed through the demonstrative that is omitted, creating an instance of pro-drop, i.e. pronoun/subject omission.

Regarding speaker strategies, it appears that speaker A favours such explicit means of stating facts over more general propositions, which explains why the number of 'simple' statements uttered by him is slightly lower, although the

overall number of statements is the same for both speakers. This apparent tendency towards greater explicitness can also be seen in A's use of a `stateIntent` speech act, which signals volition, in this case the wish to emphasise his own intention in saying *what i <pause type="short" /> wanted to impress upon you is uh <pause type="long" /> the influence of English on vernacular languages <pause type="short" /> definitely is there*. Expressions of opinions or (epistemic) modality are relatively rarer, and also more or less equally distributed between the two speakers. These suggest a relatively low degree of hedging or personal stance, underlining the academic, fact-oriented, nature of the dialogue. In contrast, the interactional (or interpersonal) nature of the dialogue is reflected in signals of `(non-)awareness` ('knowing') or the two instances of `agree`ment by speaker A.

The request for information in declarative form, perhaps contrary to expectation, does not represent a declarative question, but instead probably a case of the type of question inversion that has frequently been remarked upon as a feature of Indian English (cf. Sedlatschek 2009: 289 ff.). To help us understand this better, it is necessary to look at a slightly larger context, shown in Example (10).

```
(10)<q-wh n="145" sp-act="reqInfo" polarity="positive"
    mode="open-query">
    <pause type="short" /> what exactly is the meaning of
    English <punc type="query" />
    </q-wh>
    <dm n="146" sp-act="phatic">
    <pause type="long" /> <comment content="pro-drop" />
    mean <punc type="level" />
    </dm>
    <decl n="147" sp-act="reqInfo" polarity="positive"
    mode="poss2-alternative-query">
    how really it differs from this standardized British
    English or uh <pause type="short" /> whatever you may
    call it <punc type="query" />
    </decl>
```

In unit 145, the speaker poses a (potentially rhetorical) question, followed by a DM – marked as `phatic`, although (I) mean – where we here encounter another instance of pro-drop – is frequently seen as initiating a self-correction – followed by what could be considered a re-formulation exhibiting syntactic inversion and the absence of periphrastic *does*, the equivalent of *how does it really differ…* in Inner Circle varieties.

Syntactic fragments (`frag`) are incomplete or non-well-formed c-units. One of their main functions in speaker-initiated, more 'expository'/information-providing sequences of units is to indicate deictic reference in the shape of single NPs or PPs. These phrases generally represent leading or clefted adverbials that standard grammar would treat as part of main clauses, but which are split off in the DART scheme and treated as independent pieces of information, partly because they are usually separated from the next unit prosodically via a pause, and partly because they provide focussing or circumstantial information, as in the following example:

```
(11)<frag n="41" sp-act="refer" polarity="positive"
    mode="partial-frag">
    <pause type="short" /> such people <punc type="level"
    />
    </frag>
    <decl n="42" sp-act="refer" polarity="positive"
    topic="time" mode="frag">
    when they are speaking <punc type="level" />
    </decl>
    <decl n="43" sp-act="state" polarity="positive"
    mode="poss3-alternative">
    <pause type="short" /> there may be some difference
    in accent or <pause type="short" /> uh their use in
    the words <punc type="stop" />
    </decl>
```

In example (11), the fragment *such people* puts the subject of the following declarative unit into focus, while the latter, in turn, provides circumstantial information about the declarative unit that follows it.

In responses to questions, frags tend to constitute elliptical answers, although this is not the case in our data, apart from perhaps one instance of a potential elaboration to an answering response. I deliberately say potential here because what the DART annotation picked up as an answering response was the single indigenous word *nei*, whose meaning I have been unable to ascertain. Judging by its context and position at the beginning of the turn following the inverted question from Example (10), *nei* here may either be an acknowledging DM, or yes- or no-response that has been marked as a fragment in the absence of any further evidence as to its exact function. As it follows a request for information, though, DART's inferencing system 'concluded' that it should be an answering speech act and thus marked the following c-unit as an elaboration.

Another option for fragments is that they represent statements, reports, expressions of opinions or epistemic modality, similar to the ones discussed for declaratives. In this case, the only difference to declaratives is that they are not completely grammatically well-formed or complete, but nevertheless constitute more than single phrases, usually also incorporating verbs, such as e.g. in *such a differences may be there*, where there is incorrect agreement between the indefinite determiner and the noun. However, such fragments, which include the one instance of a prediction (*spoken English accent will be different*) are relatively rare in the data, due to the fairly high proficiency of the two speakers.

The remaining three categories at the bottom of Table 2 again represent more interactive features, where `echo`es in general represent repetition of the previous speakers' content, as shown earlier, often as a kind of backchannel mechanism that may either fulfil an acknowledging/agreeing function or allow the speaker to invite the interlocutor to clarify whether the message has been understood properly. The only difference between `echo` and `echo-refer` here is that the latter again contains a deictic reference in the form of NPs or PPs. The one instance of a `reqConfirm` produced by speaker A is similar to a tag question inviting a request for confirmation, except that it consists of the single word *correct* that at the same time constitutes a whole turn.

The remaining syntactic categories all embody speech acts related to the interpersonal and interactional level, rather than expressing ideational/propositional content.

Table 3: Frequencies of speech-act realisations for the remaining syntax categories

syntax	speech act	frequency A	frequency B
dm	init	13	9
dm	hesitate	11	6
dm	phatic	1	6
dm	exclaim	0	1
dm	agree	1	0
dm	expressConviction	2	3
dm	contrast	1	0
imp	direct	1	1

syntax	speech act	frequency A	frequency B
yes	acknowledge	4	6
yes	confirm-acknowledge	0	1
no	negate	2	0
q-wh	reqInfo	4	1

Amongst the DMs, those that indicate (initalise) a new topic clearly dominate the list for both speakers, although speaker A again seems to be somewhat more precise in marking transitions between topics. At first glance, this positive feature seems to be 'offset' by the higher incidence of hesitation on his part, but of course we need to remember that such a still relatively low number of hesitation markers in a rather lengthy dialogue in no way represents the mark of an unskilled speaker when dealing with unscripted, impromptu, conversation. Speaker B, in contrast, appears to be slightly more concerned with 'conversational niceties', as he uses more phatic devices, such as *see* or *(I) mean*, or the exclamative *ah*, although this could also be a mis-transcription of the hesitation marker, which is represented as *uh* throughout the data. The function of the few DMs expressing agreement or conviction are more attitudinal in nature, as is the use of the single contrastive marker *yet*, expressing consensus with the interlocutor or emphasising the speaker's own beliefs. Again, however, such features of personal commitment are largely absent from the dialogue, which remains predominantly fact-oriented.

The purpose of the two imperatives (*say public schools* and *then take another*), one produced by each speaker, is unlike that of stereotypical imperatives, i.e. commands, but in fact similar to that of the initiating DMs discussed above, each time introducing new discourse referents into the topic and alerting the other speaker to them. They thus also both fulfil a focussing function similar to that of the deictic references discussed earlier.

Of the yes-responses provided by both speakers, only the one marked as confirm-acknowledge is in fact a response to a type of question, i.e. a request for confirmation. All other instances are simply backchanneling signals that constitute separate turns, rather than being embedded in the other interlocutor's turn. Likewise, the two no-units produced by speaker A only react to B's saying *may be i don't know anything*, and are in fact an example of an emphatic repetition within the same turn.

The five wh-questions that occur in the dialogue are predominantly rhetorical in nature, as almost all of them raise questions for discussion, rather than

really calling for answers. The only real exception to this is *where did you get those <pause type="short" /> uh wood laws entry*, which, even given more context, does not seem to make any sense and may represent a transcription error or third-party talk that simply was not marked as such because removing it from the dialogue does not interfere with the coherence.

To summarise: the speech-act behaviour identified through the analysis indicates that the two speakers seem to be more or less fully aware of the genre conventions associated with academic discussions and are able to apply these successfully, even if the grammar used may contain some 'performance errors' related to specific features of Indian English.

Amongst the ones I observed, partly due to their affecting the analysis, are what we might term 'determiner-drop' (cf. Sedlatschek 2009: 204 ff.), as we saw in the example of *accent may be different* earlier. The question here is whether this structure should be seen and treated as a declarative in Indian English (or other varieties that exhibit the same feature) or as the fragment it would constitute in Inner Circle varieties? In addition, it would be interesting to see to what extent the same feature might conceivably occur in Inner Circle varieties as a kind of 'elliptical construction' where the initial determiner is reduced so strongly that it (seemingly) disappears.

In some instances, especially involving forms of the verb *mean*, there also seemed to be pro-drop, e.g. in *[I] mean* or *[it] means that is another kind of trend which going on*, which I could not find attested in the literature.

I have already commented on the effect of the indigenous word *nei* on the interpretation and automatic annotation of discourse processes above. This is a clear example of how such indigenous words or expressions may affect the annotation process and thus necessitate the involvement of other researchers more familiar with the varieties to be annotated, and, sadly, contradicts the assumption I expressed earlier that common expressions like discourse markers could probably be found in phrasebooks with relative ease.

These, and some of the smaller issues I identified during the post-processing phase of the dialogue also demonstrate that it should be possible to catch such features automatically if the relatively simple grammar currently implemented in DART would be expanded to include more precise analyses of concord features, etc.

5 Conclusion

In this article, I have tried to discuss some of the most important steps that would be required in carrying out a large-scale project to enrich, as much as pos-

sible, all available ICE corpora through adding levels of pragmatics-relevant annotations, including those of syntax, semantico-pragmatics, semantics, polarity, and last, but not least, pragmatics, in the form of speech acts. I began by describing the proposed DART annotation format. This format, which could be used to either upgrade the existing corpora or create parallel, pragmatically annotated versions, not only features the relevant annotations on the levels listed above, but also has clear additional advantages in terms of processability, legibility, and analysability of the data, as well as simplifying the comparison of the corpora in ways that were previously not possible.

In the next section, I pointed out and discussed a number of issues, both on the levels of form and content, that have affected the prior annotation conventions and their application in various ICE corpora, as well as how these may cause general problems concerning processability and readability of the data. At the same time, I have also attempted to point out what kind of effect they may have on carrying out automated annotations in DART. By necessity, this discussion had to be limited to the most salient and relevant features, as a more detailed account would probably take up the space of a whole article, as well as require more in-depth analyses of all existing corpora in terms of how the initial and revised conventions have been applied, and to what extent this has been done consistently.

The case study described in the final section involved the analysis of speaker strategies in a single dialogue from ICE-India. Here, I hope to have demonstrated how the proposed annotation can be successfully applied, not only in order to create speaker profiles that reflect the pragmatic strategies and competence of the individual speakers, but also to demonstrate how the annotation process can bring to light how specific features of different varieties of English may affect the annotation, thereby also shedding more light on their exact nature, as well as potentially offering later options for modelling such features. This, in turn, could then lead to improved corpus-linguistic methodologies for annotating, researching, charting and comparing all varieties.

Notes

1. So-called ‘empty’ elements do not have start and end tags, but consist of a single set of angle brackets with a name, potential attributes, and a forward slash before the closing bracket, e.g. <pause length="1s" />. They provide information that is not part of the document hierarchy.
2. For full details of the taxonomy and explanations for all labels see http://martinweisser.org/DART_scheme.html.

References

Aijmer, Karin. 2002. *English discourse particles: Evidence from a corpus*. Amsterdam/Philadelphia: John Benjamins.

Aijmer, Karin and Christoph Rühlemann (eds.). 2015. *Corpus pragmatics: A handbook*. Cambridge: Cambridge University Press.

Australian National Corpus. n.d. Accessible at https://www.ausnc.org.au/.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Fischer, Kerstin (ed.). 2006. *Approaches to discourse particles*. Amsterdam: Elsevier.

Gibbon, Dafydd, Inge Mertinse and Roger Moore (eds.). 2000 *Handbook of multimodal and spoken language systems*. Dordrecht: Kluwer Academic Publishers.

Gut, Ulrike. 2014. ICE Nigeria. Available from http://sourceforge.net/projects/ice-nigeria/.

Kallen, Jeffrey and John Kirk. 2012. *SPICE-Ireland: A user's guide*. Queen's University Belfast, Trinity College Dublin, and Cló Ollscoil na Banríona.

Kirk, John. 2013. Beyond the structural levels of language: An introduction to the SPICE-Ireland corpus and its uses. In J. Cruickshank and R. McColl Millar (eds.). *After the storm: Papers from the Forum for Research on the Languages of Scotland and Ulster Triennial Meeting, Aberdeen 2012*, 207–232. Aberdeen: Forum for Research on the Languages of Scotland and Ireland.

Leech, Geoffrey, Martin Weisser, Andrew Wilson and Martine Grice. 2000. Survey and guidelines for the representation and annotation of dialogue. In D. Gibbon, I. Mertins and R. Moore (eds.). *Handbook of multimodal and spoken language systems*, 1–101. Dordrecht: Kluwer Academic Publishers.

Llamas, Carmen, Louise Mullany and Peter Stockwell (eds.). 2007. *The Routledge companion to sociolinguistics*. London/New York: Routledge.

Nelson, Gerald. 1991. *International Corpus of English. Markup manual for spoken texts*. University College, London: Survey of English Usage.

Nelson, Gerald. 2002. *International Corpus of English. Markup manual for written texts*. University College, London: Survey of English Usage Accessed 24 September 2015, at: http://ice-corpora.net/ice/manuals.htm.

Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Searle, John. 1975. A taxonomy of illocutionary acts. In K. Gunderson (ed.). *Language, mind and knowledge* (Minnesota Studies in the Philosophy of Science III), 344–369. Minneapolis: University of Minnesota Press.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and change*. Amsterdam/Philadelphia: John Benjamins.

Sinclair, John. 2005. Corpus and text – basic principles. In M. Wynne (ed.). *Developing linguistic corpora: A guide to Good Practice*. Oxford: Oxbow Books, 1–16. Available online from http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm.

Stenström, Anna-Brita. 1994. *An introduction to spoken interaction*. London: Longman.

TEI Consortium. 2015. *TEI P5: Guidelines for Electronic Text Encoding and Interchang*e. Version 2.8.0.

Tsui, Amy. 1994. *English conversation*. Oxford: Oxford University Press.

Weisser, Martin. 2015. Speech act annotation. In K. Aijmer and C. Rühlemann (eds.). *Corpus pragmatics: A handbook*, 84–113. Cambridge: Cambridge University Press.

Weisser, Martin. 2016a. *Practical corpus linguistics: An introduction to corpus-based language analysis*. Malden, MA & Oxford: Wiley Blackwell.

Weisser, Martin. 2016b. DART – the Dialogue Annotation and Research Tool. *Corpus Linguistics & Linguistic Theory, 12(2)*. DOI 10.1515/cllt-2014-0051.

Weisser, Martin. 2016c. Profiling agents & callers: A dual comparison across speaker roles and British vs. American English. In L. Pickering, E. Friginala and S. Staples (eds.). *Talking at work: Corpus-based explorations of workplace discours*e. London: Palgrave Macmillan.

Wichmann, Anne. 2004. The intonation of please-requests: A corpus-based study. J*ournal of Pragmatics* 36: 1521–49.

Wong, Deanna, Steve Cassidy and Pam Peters. 2011. Updating the ICE annotation system. *Corpora* 6 (2): 115–144.

World Wide Web Consortium. n.d. Extensible Markup Language. http://www.w3.org/ XML/.