TEI mark-up of spoken language data: the BASE experience

Sarah M. Creer & Paul Thompson

School of Linguistics & Applied Language Studies, The University of Reading

Abstract. Transcription and mark-up of spoken language data should ideally present as accurate, full and impartial a representation of the original speech event as possible, but processing of the data record is subject to a number of compromises between the pull of competing forces, such as the demand for *user readability* along with *computer readability*, and the requirement (for purposes of interchangeability) for *conformity to existing standards* vs. the *accurate description of the particularities of the data.* This paper presents problems that we have encountered during the process of creating a corpus of orthographically transcribed spoken language data for the British Academic Spoken English corpus. Limitations in the recommendations of the TEI Guidelines are also discussed.

1. Introduction

The British Academic Spoken English (BASE) corpus is a collection of recordings and marked-up transcripts of academic lectures and seminars that is being developed at the Universities of Warwick and Reading. It is designed to be a British counterpart to the Michigan Corpus of Academic Spoken English (MICASE), a corpus that is to some degree representative of spoken events in academic settings in the USA.

Unlike MICASE it does not include speech events other than lectures and seminars, and the majority of the recordings are on digital video rather than audio tape. The corpus currently consists of 160 lectures and 39 seminar recordings, equally spread over four broad academic domains, of which the majority have been transcribed, and 65 have been marked up (as of 30/6/04).

The main motivation for developing this corpus is to create a resource for research into spoken academic discourse, and it is targeted chiefly at EAP researchers and practitioners. The corpus should provide a wealth of evidence of naturally occurring language in specific contexts, and the intention is to make it highly accessible, by creating a web interface for interrogation of the corpus, and by tailoring the interface to the needs of potential end-users.

Reading Working Papers in Linguistics 8. 149-174.

In the early stages of the project, the recordings of the lectures were transcribed by a variety of paid student and volunteer transcribers, for reasons of expedience, and this resulted in a degree of variability in the quality and accuracy of the transcriptions. Once funding was secured, however, we have worked to achieve consistency in the rendering of spoken language in an orthographic representation, and we have also aimed to make the corpus compatible (and thus interchangeable) with other corpora, in particular the MICASE corpus. The MICASE corpus uses the Text Encoding Initiative guidelines to encode its spoken language data.

The TEI describes itself as "an international and interdisciplinary standard that helps libraries, museums, publishers, and individual scholars represent all kinds of literary and linguistic texts for online research and teaching, using an encoding scheme that is maximally expressive and minimally obsolescent"[1]. They were originally devised for use with written language data but extended to include a tag set for the transcription of speech. The TEI guidelines allow the contextual and paralinguistic information important to the event to be represented alongside the spoken lexical information of the spoken language. The information is represented through the use of XML tags: those that maintain a semantic structure and contain spoken data between them and empty tags acting as temporal markers containing information about particular events within the overall speech event. The tags themselves, known as elements, have attributes which allow additional information to be contained within the tag. For example, the <u> tag consists of an opening and a closing tag containing an utterance between the two, and the opening tag may contain attributes such as an identifier that indicates who the speaker is (in Example 1, 'who' is an attribute of the <u> element, and '001' or '002' is the value of that attribute).

Example 1

```
<u who="001">you've got two extremes of a totally
unrealistic thing being able to be realistic
okay</u>
<u who="002">yeah</u>
                                        RL058pt3 01.46-01.50
```

Having this type of structure allows sections to be selected or suppressed for separate analysis depending on various attributes. Example 1 shows a portion of a lecture with two speakers. This tag structure then provides an

[1] www.tei-c.org (June 30th 2004)

ability to retrieve any utterance with a specified attribute, such as a particular speaker.

Empty tags[2] such as <kinesic/> contain no spoken text, but mark the occurrence of another significant contextual event, as shown in Example 2. This will allow searching and location of these types of events in the text. We follow the TEI guidelines, current edition TEI P4[3], employing the set of TEI elements for mark-up of spoken language data shown in Table 1.

Example 2

```
what precautions are necessary so if materials are sensitive <kinesic desc=”writes on board” iterated=”y” dur=”7”/> to let’s say air and it’s really by that we mean oxygen of course
                                                    RL005pt1 2.10-2.24
```

In Example 2, the <kinesic> tag shows where the kinesic event begins, and the various attributes give information about the event: that the speaker is writing on the board, that the event is iterated and that the length of the iteration is 7 seconds.

Table 1 List of elements used in mark-up of the BASE corpus (attributes not shown)

Tag	Description
<u>	An utterance is a discrete sequence of speech produced by one participant, or group of participants, in a speech event. The tag contains transcription of lexical items.
<pause/>	Indicates a perceived pause, either between or within utterances, of at least 0.2 seconds duration.
<vocal/>	A non-lexical vocal event such as laughter, coughing.
<kinesic/>	A non-vocal communicative event such as putting hand up, handing out paper, etc.
<event/>	An occurrence, not necessarily communicative, usually non-verbal, noted because it affects comprehension of the surrounding discourse. For example fire alarm, playing of audio tape, etc.
<shift/>	A marked change in voice quality for any one speaker.
<writing>	A passage of written text revealed to participants in the course of a spoken text.

[2] An empty tag is identifiable by the forward slash before the closing angle bracket.
[3] http://www.tei-c.org/P4X/

<distinct>	Used for words or phrases in languages other than present-day British English. This includes earlier forms of English but does not include proper names.
<sic>	Used when a speaker makes a mistake without self-correcting, and the error might otherwise appear to be a transcribing error.
<trunc>	Used when a word is truncated.
<gap/>	Used to indicate omissions in the text. Also used when names referred to in the recording are withheld at the request of the participant(s).
<unclear>	Used when transcriber is uncertain of exact word(s).

Transcription and mark-up of spoken language data should ideally present as accurate and impartial a representation of the speech event as possible, but this ideal can never be achieved, as Cook (1995) convincingly argues. The process of data capture and transcription is subject to a number of powerful tensions, repeatedly forcing the transcriber away from the ideal and towards compromise, and this paper describes some of the problems that we have had to deal with during the process of creating our corpus.

2. Practical issues

2.1 Time and money

The greatest practical concerns for anyone involved in the development of a corpus of spoken language data are the intertwined issues of time and money. Processing spoken language data is an extremely time consuming and expensive process and this places constraints and limitations on what can be achieved. On the BASE project we have calculated that one hour's worth of recording takes, on average, at least 10 hours to transcribe, 3-4 hours to check, and then more than 8-15 hours to mark up. Decisions have to be made over the level of detail that can be incorporated in the mark-up, while at the same time allowing for a large number of recordings to be transcribed and annotated. This is one of the primary tensions underlying decisions over the mark-up of the data: **breadth** vs. **depth**. The following sections describe a range of other tensions affecting and constraining the corpus developers' decision-making processes.

2.2 Corpus design - meeting end-users' requirements

Design is continually adjusted and reviewed throughout the process of building a corpus and it is important to base these decisions on the perceived usefulness of that particular design feature for the end-users. Ideally corpora would be designed to provide useful information for all areas of speech research, but the constraints of time and money militate against this. The BASE project is funded by a Resource Enhancement grant from the Arts and Humanities Research Board of England, and deadlines for completion of the project must be met, as well as targets for the numbers of transcripts to be completed.

Equally powerful, however, is the notion of the end-user community: the corpus is designed to constitute a set of resources for language teachers and researchers into English for Academic Purposes, and it is these end-users who must be kept in mind throughout the design process. It is for this reason that the chosen form of representation in the corpus is orthographic transcription. This decision is not without its problems, as will be shown in the following sections, but it is fundamental to the project that the corpus should be non-threatening, in its presentation, for the language teacher.

2.3 Conformity

The spoken language data representation has to be readable not only for humans (**user readability**) but also for computers (**computer readability**). At a basic level, this simply requires a high level of systematicity in mark-up, but at a broader level it argues the case for standardisation, and the importance of interchangeability between corpora. The BASE project has chosen to make its corpus a sister to MICASE, and it is planned that end-users should observe a regularity in annotation and in structuring between the two corpora. On a broader level, the BASE project aims at conformity to the TEI Guidelines. However, as discussed below, there are problems in the guidelines. A further dynamic tension at play here, then, is that between **compliance** with existing guidelines and the need for **customisation** of mark-up to account for the features of the actual data, not all of which have been satisfactorily treated for within the guidelines.

2.4 Data-gathering - quality of recordings

Records of naturally occurring spoken language events are inevitably to some degree partial; at the same time, the corpus developer needs to find ways to produce accounts of these events that, while partial, are of approximately equal degrees of partiality. This principle can be seriously challenged at the first stage: that of data collection in natural settings.

From our experience, the quality of digitised recordings of natural speech events is clearly superior to that of an audiocassette and therefore the transcription itself is more accurate. The BASE recordings were made on minidisc and on digital video in different classrooms and lecture theatres throughout the two universities. In an ideal world, one could get recordings of equal quality in any of these different settings. In reality, however, we encountered problems with the equipment not picking up all of the speech data due to technical difficulties. In some lecture theatres, for example, bulbs in the overhead projectors interfered with the wireless lapel microphones and every time lecturers moved close to the projector to change a transparency, they became inaudible. This kind of data loss can also lead to an inconsistency in the transcripts – gaps in a transcript can be due to a number of factors, such as recording problems, the poor levels of articulation of a speaker, possibly also the need to edit a section for reasons of anonymity (see 2.5 below), and decisions have to be taken over whether, and how, these gaps are to be indicated in the transcript.

Another potential source of inconsistency is the mode of the recordings themselves. The BASE project recordings are mostly on video but some are purely audio. The audio-only recordings do not contain the same level of detail of context and paralinguistic activity that are captured by the video recordings, and this poses a problem for the transcriber who will have to choose between keeping the level of detail restricted to what can be discerned on the audio tracks or having varying levels of detail depending on what is available on video, on the one hand, and audio, on the other. Allied to this dilemma is the practical issue of what software the transcriber should use: a simple audio transcription aid with the use of a VCR for viewing video as a supplementary source of information, or a dedicated computer program interface, such as the freeware program, Soundscriber[4], which permits viewing of the video on the same screen as the text editor and audio controls. The size of the image on the larger monitor may make certain visual features of the recording more salient than they would appear to be when viewed on a small window on a

[4] http://www.lsa.umich.edu/eli/micase/soundscriber.html

computer screen. A screenshot of Soundscriber in use is shown in Figure 1 overleaf.

One feature in which technology does support consistency is in the measurement of pauses. The <pause/> tag in the tag set marks the position and duration of pauses 0.2 seconds or longer, to one decimal place, in the BASE corpus. For accurate measurement of these pauses, Wavesurfer[5] is being used. This freeware program allows both the waveform and spectrogram of the audio recording to be displayed to provide a visual means for location and accurate measurement of the periods of silence that fit the BASE pause definition, without the distraction of the videoed material. The gestures and events recorded by the video which affect the comprehension of the discourse are added on a separate occasion using Soundscriber.

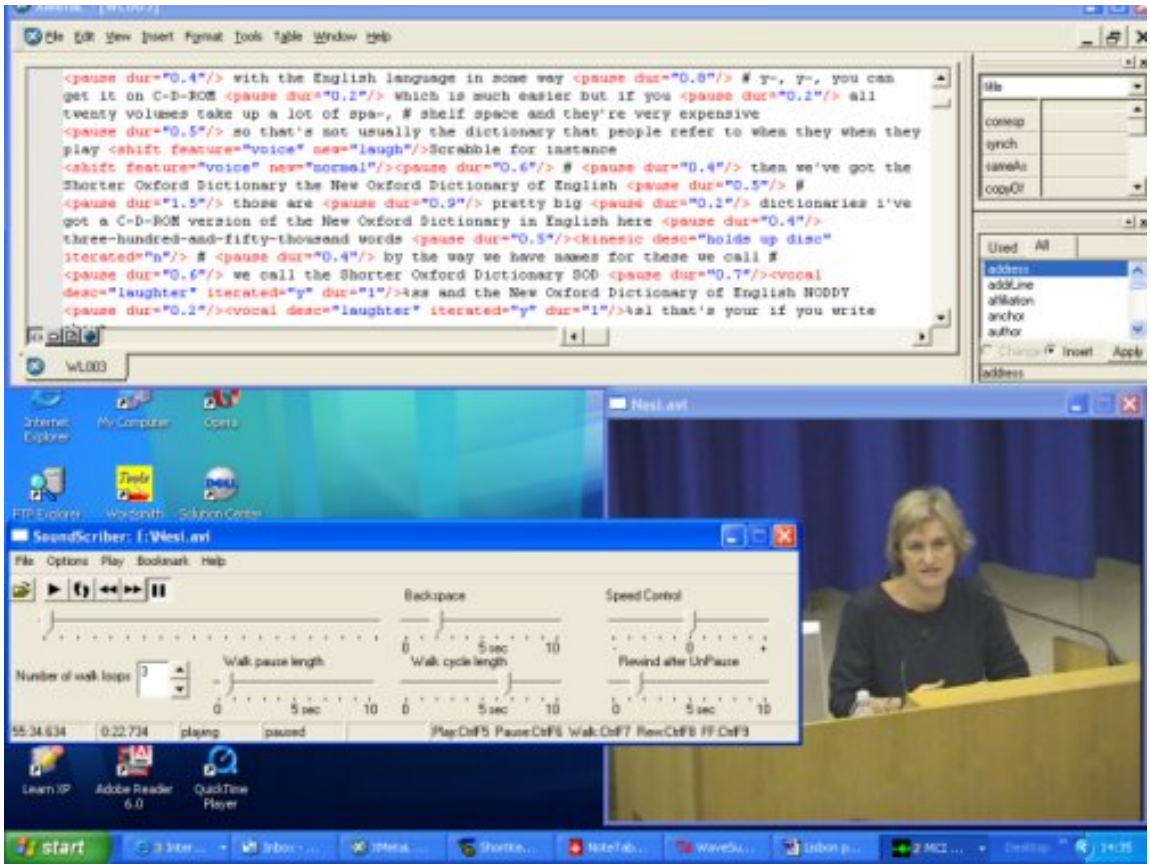

Figure 1 Screenshot of desktop showing the XML editor at top and the Soundscriber controls in bottom left and video screen at bottom right.

[5] http://www.speech.kth.se/wavesurfer/

2.5 Constraints on authenticity

Ideally, we aim to capture naturally occurring speech in an authentic situation, but there is always the question of how authentic the event recorded is when a participant is aware of being observed. Example 3 illustrates an extreme of the observer's paradox – the lecturer adds a spoof French translation of a car name for the benefit of the international students (not present at the lecture) who may listen to the recording later as part of their EAP course.

Example 3

```
it's just oscillating off it goes just oscillates
and for the foreign language students very much
like a Citroen Diane suspension it's getting
towards that or le Diane de Citroen or whatever it
whatever it is am i supposed to interact with this
or should this be a
```

RL027 pt2 09.07-09.46

Another potential threat to the authenticity of the data is the need to protect the identities of the participants. Participants must be aware of the recordings being made, what the data is to be used for and who it will be made available to. This will involve some editing and suppression of the data. A difficult issue is to what extent this suppression should be made. A degree of editing will reduce the authenticity of the data but if the corpus is to be made freely available it is possible that people might make use of data in ways not previously anticipated. For example, the corpus could be used as a source of reference for information that is taken out of context and was not agreed to by the participants. With this in mind it is an ethical decision whether to let the participants look at the transcriptions before they become part of the corpus. This would ensure that all the speakers in the recordings agree to make the information contained in the data available to others and that any information which they see as being problematic or not acceptable in a wider context can be removed. This may be an important ethical issue but it can also seriously compromise the authenticity of naturally occurring data.

3. Representation and interpretation – degrees of partialness

3.1 Orthographic transcription

Choices on how to represent data are inextricably linked with how the data is going to be used and who is going to use it. This choice determines to some extent what information is going to be captured and what is going to be lost. If a decision is made to transcribe data using IPA symbols, many of the intricacies in the realisation of the words produced will be preserved. However, this will restrict the usage of the data to those end-users who are trained in reading and interpreting IPA symbols and limit those who can transcribe the data to those highly trained in IPA transcription. By making an orthographic transcription, much of the individual pronunciation information is lost but it allows consolidation of instances of the same word independent of the pronunciation. Orthographic transcription also increases readability of the transcripts for the end-users and allows for transcription of data by non-experts in IPA transcription. As explained above, the BASE project is aimed primarily at EAP researchers and practitioners, and so the decision was taken to use orthographic transcription. In the following sections, the focus is therefore on problems that have been encountered in creating a consistent and accurate orthographic transcription, and a balanced and detailed mark-up of the data.

3.2 Spelling

A primary difficulty in providing an orthographic transcription of the data is that of spelling. It seems reasonable to assume that there should be standards and conventions available for the representation of language, and transcription guidelines often make reference to a particular spell-checking system used, or cite dictionaries from which spelling rules have been taken. Speech, however, is a transient entity which could provide problems in standardisation and interchangeability of corpora in the short and long term. Some of these problems are:

- Variation in spellings between dictionaries.
- Variation in spellings of words that are all deemed acceptable in English, e.g. *categorise/categorize*, and per cent, percent and per-cent.
- Technical terms and neologisms not yet codified, which can run the danger of idiosyncrasy.

- The representation of made up words, e.g. *skaboodle*. The lack of a one-to-one phoneme-to-grapheme relationship in English provides inconsistency across corpora and ambiguity in pronunciation.
- Language change, e.g. the standard written past participle spelling of earn is earned but from analogy with *learn* and *learnt*, *earnt* is being used more as a standard term. If popular usage dictates the direction that language is moving, as the usage becomes more widespread it could become a standard term, posing long-term problems for the interchangeability of corpora.

Decisions have to be made early on in the transcription process to ensure consistency, which, in turn, permits comprehensive and precise searches, and accurate word frequency lists to be drawn from the corpus. For the BASE project a list of standard conventions will be provided for reference alongside the transcriptions.

On whose authority?
In the BASE corpus, there is a lecture on Pericles which is spelled Pericles in the Oxford Reference Online database, but the notes given by the lecturer used the spelling Perikles. By definition the lecturer is an expert in his field. This raises the issue of whether to follow the spelling of the expert or to use that of the standard language.

3.3 Definitions of a 'word' and orthographic representation of non-words

It is a generally recognised fact that spoken and written language differ in various respects, and these differences create a problem for the representation of spoken language following written standards. Spoken language is a continuous process, a complex interaction of articulators, whereas written language is a discrete representation of segments at the word level. What is written in dictionaries does not define what and how all things are said. If the continual stream of speech was segmented into words using, for example, the definition that a word is the minimum unit in writing to which meaning can be assigned, items such as truncated words and noises do not fit into this categorisation. Questions about whether these individual noises are meaningful enough to warrant word status presents another set of choices to be made.

Example 4

```
so i want the arm to come in as quick as possible
to to that point <shift feature="pitch"
new="high"/> zing <shift feature="pitch"
```

```
new="normal"/> like that do the <vocal
desc="buzzing noise" iterated="y" dur="1"/> do the
welding and move away again now if it's overdamped
the case we saw before it will not come in zing it
will come in <shift feature="tempo" new="ll"/>
zing <shift feature="tempo" new="normal"/> and
eventually get there if it's as the case we're
going to move on to it goes past it overshoots
then of course it would go zing
```

RL027 pt2 01.24-01.50

In Example 4, taken from a lecture on mechanical engineering, the question is whether “zing” is to be represented as a word, or marked with a <vocal/> tag which would make it a non-lexical item. The <shift/> tag ‘feature’ attribute allows a range of options to be specified where a change in the voice quality of a speaker is marked, such as pitch range, loudness and tempo. The ‘new’ attribute specifies how the feature has changed. In this example the first <shift/> tag marks the point where the pitch range of the speaker’s voice changes and the second indicates where it is no longer marked by using ‘normal’ as the ‘new’ value. The third use of the <shift/> tag demonstrates where the tempo of the speaker’s speech changes, where ‘ll’, as the value for ‘new’, represents a very slow tempo, with the final <shift/> tag marking the return to the normal speech tempo. The <vocal/> tag allows a description of the phenomena, noting whether it is an iterated event and how long the phenomena lasts in whole seconds.

Example 4 shows two distinct types of noise made: one is a buzzing noise, which has been tagged here as <vocal/>, and the other is a noise which is here represented as ‘zing’. One influence on the decision to represent it thus is that it consists of sounds of English which can be put together to make a lexical item, and a second factor is that it can be found in a dictionary. In the first three instances of the sound in this extract it performs a role similar to that played by the buzzing noise in describing the sounds made by parts of a welding machine, and it could be argued that the two types of noise should be marked up as being similar, rather than one being <vocal/> and the other lexicalised with a <shift/> tag employed to show the marked prosody of the first three instances. The fourth instance, however, suggests that the sound is performing a more conventional ‘word’ role in the utterance. An alternative is to try to represent the buzzing noise as a word, in the same way as ‘zing’, although there is no clear means (‘buzz’ is not an adequate representation of the sound made in this instance) of doing so.

Another example from the corpus is that of the lecturer who frequently used the phrase “all right” after statements throughout the lecture. The intonation of this remained the same through the lecture but the realisation turned into “mm” rather than using the actual words. This is an example of how a simple transcript of the lecture would hinder the understanding of the speech event. The written transcript can only adequately provide lexical information about the event rather than the paralinguistic and contextual information that the participants are using. This argues the case for linking the transcript to a prosodic transcription or to the audio files, which would help to disambiguate the meaning.

3.4 Visual representation of data above the word level

The visual representation of a transcript has been shown to have an influence on the interpretation and perceptions of the event by the reader. The layout and punctuation provides interpretations of how the spoken language was produced. Edwards (1993) discusses the role of the transcript in spoken language research, and postulates a set of maxims for readability of transcripts which the transcriber should keep in mind to limit misinterpretation:

- Proximity of related events.
- Visual separability of unlike events.
- Time-space iconicity.
- Logical priority.
- Mnemonic marking.
- Efficiency and compactness.

Where written language is delineated by punctuation, the equivalent for spoken language is the complex interaction of prosodic marking. To represent one level of the prosody, and give clues to the phrasing of the data, all pauses have been marked in the BASE corpus and precisely measured. This makes the transcript more difficult to read for the end-user but also provides information about the realisation of the speech. This is an example of the kind of compromise that has to be made in the processing of spoken language data, between the **need for readability**, and the importance of **contextual and paralinguistic information**.

3.5 Interpretation: the transcriber’s dilemma

Any transferral of data from a recording of raw speech into written form will be an interpretation. To represent what is heard means that some

information is inevitably lost, and this will restrict the options for further interpretation of the data by others.

A difficult issue for the transcriber is that they are responsible for interpreting what is relevant to the event. The transcript is necessarily selective and a matter of interpretation. Having visual accompaniments to the transcript can aid the user but it does not overcome the problem that the observer is not a participant in the event. What is deemed relevant by the transcriber may not have been given the same relevance by the participants in the event itself. Often the transcriber does not have all the information, such as just having audio recordings where non-vocal events can be heard but decisions on their relevance have to be made without access to all the information necessary. Audio recordings do not necessarily capture all non-vocal information that is relevant and a degree of editing has therefore been done without the transcriber being aware of that decision.

Decisions about whether to represent a fully or partially articulated word as its phonetically realised form or as its underlying form provide further problems of interpretation. If a word is to be represented orthographically, then the representation depends on whether the word appears in a codified form, for example ‘till’ is a separate codified shortened version of ‘until’ but there is no separate dictionary entry for ‘`kay’ as the partially articulated version of ‘okay’. A further example is the *learnt*/*learned* difference. The past and past participle of *learn* can be either *learnt* or *learned*, where *learnt* is preferred in British English, particularly when the word functions as a participial adjective, according to Fowler's Modern English Usage (1999). In a single recording it is possible to find a speaker switching between a pronunciation of the word which ends in a /t/ sound, and one which ends in /d/. The transcriber then has to interpret whether this distinction exists in the mind of the speaker (they actually thought ‘learnt’ in the first case, and ‘learned’ in the second), and whether or not this should be made clear through the orthography. To make the transcript comprehensively searchable however, the orthography has to be standardised. Representing what is said is a matter of interpretation and the challenge is to constrain the degrees of variability of this interpretation as far as possible.

Making decisions about how to represent the spoken form as written language is particularly apparent in the problem of homophones. The transcriber has to interpret the context within which the word is spoken to assign it the most probable meaning and written form. In cases such as “they’re”, “their” and “there” and “see” and “sea”, reliance on the context can usually disambiguate these forms but the transcriber cannot know that that particular form chosen to represent the occurrence is that form which

the speaker intended to utter. There are instances, such as Example 5, where the context does not provide a weighting towards one or the other of the possible representations due to the non-fluent production of speech.

Example 5A
this is a problem with property a companies always have to have somewhere to live

Example 5B
this is a problem with property a company's always have to have somewhere to live

RL031 pt2 07.20

Making a decision about which of these two forms in Example 5 ('company's' and 'companies') is the one that the speaker intended is making assumptions about cognition and pre-planning of utterances. The aim of making a non-theory dependent corpus then has to be compromised to make a representation of what has been uttered.

The transcriber must transcribe what is heard and not correct the speech to make it grammatical in standard written English. The observer cannot know whether what has been uttered fulfils the intention of the speaker. Correcting what would be seen as grammatical errors in standard written English also implies that the form spoken by the speaker is not acceptable and is a form of judgement on dialectal or idiolectal features of the individual's speech.

4. Difficulties using TEI guidelines

Spoken communication differs from written communication in that it is not only a lexical and syntactic event, but also consists of the context in which the event occurs and additional paralinguistic information (Cook 1995). To become a participant in a spoken communication event, presence at the point in time at which the communication is delivered and membership of the audience at which the communication is aimed are required. The written record of the words spoken is insufficient to capture all that is required to interpret a spoken event. The TEI guidelines on mark-up of spoken language data provide means for making the representation of the event more comprehensive. Problems in the practicalities of adapting tags for spoken language events from linear written language and trying to

represent these events to provide an effective interpretation are discussed in this section.

4.1 Representation

Trying to represent the speech event accurately, particularly the temporal aspects of speech, while conforming to the TEI guidelines, provides a difficulty for the encoder. Representation of an event influences the interpretation by the user as they can only interpret the event from the information provided and how it is presented.

One of the biggest tensions in providing an accurate representation of spoken language data using the TEI guidelines is between customisation and conformity (cf. section 2.3 above). Where there are too many potential ways to encode events (for example, ways to mark time alignment) in the TEI guidelines, there will be ambiguity, confusion and this will limit consistency both within and across corpora.

Furthermore, the representation of spoken language in written form imposes linearity on speech. To a certain extent this is due to the constraints on a human's ability to produce more than one word at a time. However, the TEI guidelines impose linearity on the whole speech event and the events contained within it. Speech is a temporal phenomenon and on levels above and below the representation of the word, other non-linear events, such as prosody, are being constrained by the linear demands of the encoding system. The linear representation also only allows events to occur at certain break points in the transcripts, at word boundaries. These artificial boundaries in a stream of continual speech are enforced by the orthographic transcription system, not necessarily correctly representing the actual temporal events.

This tension between accurately representing the temporal nature of the event and conforming to the TEI standards can be illustrated by <u>, the dividing unit of utterance. It is defined in the TEI Guidelines as "a stretch of speech usually preceded and followed by silence or by a change of speaker". However, speakers do not necessarily finish their utterance when another interrupts. The option for the transcriber is then to either indicate the temporal aspect of the interruption, breaking the utterance artificially or mark the end of the utterance as a whole entity. The overlap can then be indicated either with time stamps or by using style sheets. If the speaker who interrupts fails to gain the floor, his or her utterance then becomes embedded. The difficulty then is to decide where to mark that embedded utterance. The choice is either to break the utterance of the interrupted speaker, following Edwards' (1993) principle of proximity,

which implies in the transcript that there was a break, or to violate this principle allowing the loss of temporal information. This becomes more complicated if other speakers try to interrupt as it then can become unclear, when represented linearly, which speaker is being interrupted. Using <u> in the representation of multi-speaker events (such as small group discussions in interactive lectures, or a heated exchange of opinion during a seminar) becomes highly problematic. Without time stamps or a link to the recorded event itself, the user can only interpret from the information presented and the manner in which it is presented.

Example 6 below shows a portion from a lecture in the corpus which demonstrates the difficulty in representing utterances which are embedded or which overlap with other utterances and how this representation and layout affects the way that the utterances are interpreted by the user. Example 6A shows a conversation analysis type representation of the section where square brackets represent overlapping speech. Example 6B shows a representation of the same portion of text with TEI encoding, where the 'trans' attribute assigns the value that describes the transition between utterances. The difference between these two examples illustrates the difficulty in the decision made to break the utterance after "okay" even though the "yeah" reply is embedded in the first utterance temporally. The layout and tagging in 6B suggests that there is a break in the first utterance at that point where there was not one produced to maintain the semantic flow and temporal nature of the interaction. Example 6C shows a representation following the TEI guidelines but inserting a new tag <timestamp/> which notes the time at which the utterances begin. Without the timestamps and just presented with the transcript, the interaction between speakers is not represented accurately enough for the user to interpret it as it was produced. The timestamps give the transcript the temporal information that has been missing to allow the user to recreate the event to some extent rather than relying on what could be a misleading layout. The overlap is marked, for illustration, with an <overlap/> tag which contains further information in the attributes allowing a reference to another <overlap/> tag to indicate the alignment between the two. For illustrative purposes, in this example, this has been shown by an attribute 'ref' which matches its value to the section of speech it is overlapping. The overlap could be further illustrated through the use of style sheets to render the overlapped sections, for example, in different colours. With the timestamps and overlapping marked, however, the readability of the text is compromised.

Example 6A

```
001: you've got two extremes of a totally
unrealistic thing being able to be realistic okay
by [this]

002: [yeah]

001: definition and a totally realistic thing
being completely unrealistic  [by this definition
so  ]

002:                          [exactly   exactly
it's the]

001:  it  includes  and  excludes  absolutely
everything [at the same time]

002:      [that's right    ]
```

RL058pt3 01.46-01.57

Example 6B

```
<u who=”001” trans=”pause”>you've got two extremes
of a totally unrealistic thing being able to be
realistic okay</u>

<u who=”002” trans=”overlap”>yeah</u>

<u who=”001” trans=”overlap”>by this definition
and a totally realistic thing being completely
unrealistic by this definition</u>

<u who=”002” trans=”overlap”>exactly exactly it's
the</u>

<u who=”001” trans=”overlap”>so it includes and
excludes absolutely everything at the same
time</u>

<u who=”002” trans=”overlap”>that's right</u>
```

RL058pt3 01.46-01.57

Example 6C

```
<u who=”001”><timestamp=”01.46.1”/>you've got two
extremes of a totally unrealistic thing being able
to be realistic okay</u>

<u who=”002”><timestamp=”01.50.3”/><overlap
ref=”a”>yeah</overlap></u>

<u who=”001”><timestamp=”01.50.0”/>by <overlap
ref=”a”>this</overlap> definition and a totally
realistic thing being completely unrealistic
<overlap ref=”b”>by this definition</overlap></u>

<u who=”002”><timestamp=”01.53.3”/><overlap
ref=”b”>exactly exactly it's the</overlap></u>

<u who=”001”><timestamp=”01.54.0”/><overlap
ref=”b”>so</overlap> it includes and excludes
absolutely <overlap ref=”c”>everything at the same
time</overlap></u>

<u who=”002”><timestamp=”01.56.4”/><overlap
ref=”c”>that's right</overlap></u>
```

RL058pt3 01.46-01.57

A further problem lies in the representation of audio-visual elements and other forms of non-verbal evidence/illustration used in academic lectures. These illustrations can be central to the communication of the information contained in the lecture. A decision must be taken about to what extent they should be represented, and also about how to represent them. The tag provided for such an event is the <writing> tag, which contains “a passage of written text revealed to participants in the course of a spoken text” (TEI Guidelines). The main issue with this tag is how the items being revealed are represented. It is only the written text that is specified to be encoded, again privileging the written form. Items such as diagrams, formulae and illustrations can be just as communicative, particularly in disciplines such as Meteorology and Economics that make frequent use of symbolic representation, and should also be included. Further questions such as how far this can be done within the text and to what extent it is to be reproduced in its physical appearance, such as the font and size, need to be addressed.

The <writing> tag does not deal with how to represent other types of illustrations such as pronunciation examples. Example 7 comes from a lecture in which different accents of English are described demonstrating different pronunciations of the word “through”. In the orthographic transcription this would not be marked but as it affects the understanding of the lecture as an illustration of a point, it would be useful to have a phonetic description of what is pronounced. In the BASE corpus, we have chosen to use the <distinct> element with the ‘type’ attribute, the sublanguage or register of the text contained within the tags, set to a ‘sampa’ value, and then to represent the sound using the Sampa[6] system of phonetic transcription, contained in square brackets.

Example 7A: orthographic transcription

```
so if i start saying if i start changing my vowel
in through to through to through or something like
that which many English people do
```

Example 7B: with suggested pronunciation mark-up

```
so if i start saying if i start changing my vowel
in <distinct type="sampa"> [Tru:]</distinct> to
<distinct type="sampa">[Tr}]</distinct> to
<distinct type="sampa"> [TrY]</distinct> or
something like that which many English people do
```

RL032 pt1 06.58-07.08

4.2 Interpretation

A general problem in the TEI guidelines for the processing of spoken language data is that many of the tags demand a great deal of interpretation by the encoder. For example, the difference between the tags <event/> and <kinesic/> is that events marked as kinesic are communicative. Not only does the encoder have to decide whether the event is relevant and to be included, but also whether it has a communicative function.

The tension between depth and breadth of transcription means that there has to be some interpretation of events that are observed. With audio only recordings the encoder is limited to what discourse affecting incidents are heard or noted down as observations at the time of the recording. With the videoed information, the extent to which other communicative and non-communicative events which potentially affect the comprehension of the discourse by the participants, is made clear. Some selectivity of these

[6] http://www.phon.ucl.ac.uk/home/sampa/home.htm

events has to take place due to time and money constraints. There will already be some degree of editing due to the inability of the recording equipment to capture every event taking place in that situation.

To make the encoding system consistent across the corpus, a set of rules should be established in which the justifications for inclusion or non-inclusion of tags are described. However, the development of such a set of rules in in itself problematic, for two reasons. If the encoder decides to mark a particular instance where the lecturer indicates (physically) a point on a slide they are showing, it could be argued that, to achieve consistency, every time a lecturer indicates a point on a slide it should be noted in the corpus. This would imply that the act of indicating a point on a slide is of equal significance, in communicative terms, in each instance, which does not seem to be the case from the interpretive view of the non-participant encoder. Secondly, the selection of this action as worthy of note is effectively a privileging of the action over others which are not marked, but which are not necessarily any less important in the way they affect the comprehension of the discourse or the production of the speech itself. The necessary alternative is to draw up a dynamic list of rules, derived from experience of marking the data, which articulates the encoder's rationale for judgements about whether or not a particular action is worth tagging. In other words, it is extremely difficult for the encoder to maintain consistency without imposing a great deal of his or her own interpretation of events on the transcript. There is very little that can be specified *a priori* to ensure a level of consistency, just a reliance on a non-participator's interpretation of events. Where there is no *a priori* defined rule base, there will always be room for inaccuracies and inconsistencies.

Inconsistencies in encoding data will also occur across corpora due to the individual needs and interpretation of the encoder. For example the definition of <pause/> is "a pause either between or within utterances". In the BASE project a pause is empirically defined as a period of silence from 0.2 seconds long, narrowing down the TEI definition. This definition of <pause/> means that the DTD would have to be altered as it does not allow for a pause to occur within <distinct>, an element which identifies words and phrases that are distinct in some way from the surrounding language. The <distinct> tag was originally created to describe a feature of written, rather than spoken, language but the need for it can be demonstrated in Example 8, where the 'lang' attribute specifies the language of what is contained with the tag. The extendable nature of the TEI guidelines allow for this customisation which can be noted in the DTD. Problems occur, however, when customisation leads to ambiguity in interpretation.

Example 8

```
Passy himself has given instance of th-, in-,
instances of this <pause dur="0.5"/> # <distinct
lang=”french”> dans un parler tant soit peu langue
on distinguera</distinct> <pause dur="0.4"/>
<distinct lang=”french”>trois petites
roues</distinct> <pause dur="0.2"/> that's # three
little wheels
                                        RL011 pt1 05.54-06.04
```

The TEI definition of <pause/> leaves scope for a range of interpretations, some of which will be primarily impressionistic. Customising the definition of <pause/> in turn affects the definition of <u> marking utterances, where the attribute, ‘trans’, which describes the transition between utterances, has the possible values for beginning the following utterance:

smooth - without unusual pause or rapidity.
latching - with a markedly shorter pause than normal.
overlap - before the previous one has finished.
pause - after a noticeable pause.

Interpretations of the terms “unusual” and “normal” then have to be made. Taking the definition of a pause of the BASE project, the attribute ‘smooth’ becomes redundant. A normal pause would be defined as one that is 0.2 seconds or longer and shorter than normal would be less than that, fitting into the categories of ‘pause’ and ‘latching’ respectively. This compromises the degree of interchangeability of corpora as these definitions depend on a definition of a pause decided by the individual encoder. The dependency on the individual interpretation illustrates the tension between the freedom to customise to the needs of the individual and the ability to conform to conventions and standards to make interchangeable corpora.

The impact of these inconsistencies within and across corpora could be reduced by ensuring that the transcriptions are linked to the audio and video files, where this is possible. Although this would not allow information retrieval for locating events within the corpus wherever there are no tags, it would at least provide as full a picture as possible of the original event as the encoder received it.

4.3 Gaps in the TEI guidelines for spoken language data

There are gaps in the TEI guidelines which, if filled, could make it far easier to create corpora that follow a strict set of standards and conventions. This would create a higher degree of interchangeability between corpora, and reduce the onus on the individual transcriber.

One of the main gaps in the guidelines is that there is no satisfactory way of marking speech that is not the speakers' own, such as in the case of directly read or quoted material (a common feature of academic discourse). The TEI guidelines suggest that "reading" could be the description of an event but this seems to be unsatisfactory as it refers to the action of reading as a separate event to the language and speech being produced. This would cause difficulties such as the encoder needing visual information about when the reader was looking at the text to read it and making interpretive decisions about whether the text was being read or not. It is relevant for users to know that a portion of language is written language communicated through speech. If marked with an empty <event/> tag, the read speech would not be easily suppressed from the rest of the corpus if so wished by the user. For these reasons, this is not how read or quoted speech is marked in the BASE corpus. A decision remains to be taken as to whether to adapt an existing TEI-specified tag, that roughly approximates, or to introduce a new tag for use in our corpus, and move for it to be included in future versions of the TEI guidelines. The working definition used for the tag is "text which can be attributed to an identifiable source when it is being quoted and not referenced where the whole text being quoted is at the non-finite clause and above level". This is not a definition without problems but it captures the larger portions of speech which are definitely read or directly quoted. At the same time, there are a number of complications, for example, the extent to which individual words read or quoted would be marked, and how to deal with idioms, proverbs and sayings, and placement of the tags in a stretch of speech with mistakes in the reading or breaks for embellishment. A distinction would also have to be made between text that is being read and text that uses part of the same language to refer to the contents of the text previously read. Information about the source and whether it is present at the time of the speech event or whether it is quoted from memory is also to be considered.

Another area in which the TEI guidelines are lacking is in providing a means for indicating who utterances (or parts of utterances) are addressed to. In a lecture, this does not usually pose many problems, as lectures are often monologic, but in the case of a seminar, particularly a highly interactive discussion, it is possible that some words are directed at

certain individuals while the remainder of an utterance is addressed to other (or to all) participants in the event, and that knowledge of whom each part of the message is addressed to will affect comprehension of the transcript. The question of how to mark up the data is complicated by the possibility that part of the utterance is directed to one or more addressees while the other parts of the utterance are addressed to other participants. The indication of the addressee cannot be included as an attribute of an <u> element, therefore, and will have to be treated within a separate element tag.

Other lacunae are that the TEI guidelines do not address the problem of how to deal with the representation of speech impediments, nor do they fully represent the whole range of possible occurrences of unintelligible fragments of speech such as truncated words.

4.4 <shift/>

The <shift/> tag exemplifies the problems discussed of both interpretation and representation in the TEI guidelines. The <shift/> tag "marks the point at which some paralinguistic feature of a series of utterances by any one speaker changes". There has to be an initial interpretation by the non-participant encoder to select these segments of speech that demonstrate this change. For example, voice quality is noted as a possible shift in speech but creaky voice is a feature of speech that occurs frequently at the end of phrases. The issue here is whether this would only be marked in places where it is not expected and conveying non-lexical information. This means that the encoding process would then become tied to one particular theory of paralinguistic behaviour selected and interpreted by the encoder.
On top of interpretation, the further question then is how the shift is represented. The TEI guidelines use individual features such as voice quality, pitch range and rhythm and represent changes in the features as relative discrete changes. Speech is a continuous stream of moving articulators which do not necessarily move in the discrete way that this definition requires. The imposition of a linear structure on speech events also only allows the tags to be placed at word boundaries, which makes the decision about where to place the tags difficult for the encoder and the output inaccurate.

The design of the <shift/> tag does not allow more than one feature change to be noted in one tag even though it is not one feature but that combination of changes in the parameters that creates the perceived shift. The way that this is designed means that each feature can end at different points in the speech and also be picked out for separate analysis. However,

it also creates a problem of inconsistency. The individual changes in each feature may not be enough to create a perceived shift but are crucial components in combination for creating the shift. The one feature design implies that wherever there is that amount of a change in that feature, there will be a tag indicating it, which is not the case.

Another option of how the change could be represented is by defining the shift as a description of the function e.g. <shift desc="mimicking a child's voice"/>. The problem here is that this is an interpretation made by a non-participant in the speech event making non-empirical impressionistic judgements. It would however be difficult to describe exactly in acoustic terms using the combination of listed features, the output of the tagged speech, as it would also be for speech produced with various emotions. This description type labelling would also privilege the section because of its function rather than its actual acoustic realisation which may not necessarily exactly align, therefore providing an inaccurate representation.

Neither of these options provides a satisfactory way of encoding paralinguistic information and it raises the question of whether these sections should be marked at all if there is no accurate and consistent way of describing them. The attempt to capture valuable information about the details of the event is undermined by interpretative and inconsistent representation. It would seem that a comprehensive prosodic transcription of the text or a link to the audio file would be the only non-interpretive option here. However this would require an exceptionally fine level of detail in mark-up and alignment of transcript to audio file, to make possible the direct searching and retrieval of sites of particular prosodic interest, and this level of detail is not possible for a corpus of the size of BASE. Questions surrounding how to deal with shifts in paralinguistic features using the <shift/> tag therefore remain unanswered.

5. Conclusion

In this paper we have discussed some of the issues involved in processing spoken language data, and some of the difficulties in using the TEI guidelines for the encoding of the BASE corpus.

Basing a system of transcription and encoding for spoken language data on standards for written language raises a number of problems. In the case of orthographic transcription, a major problem is that, though the system assumes a set of codified standards, the reference texts that exist are not entirely comprehensive (and indeed cannot be). This results in

difficulties for maintaining consistency both within a corpus and across corpora, which in turn restricts interchangeability.

Problems in the TEI Guidelines have been identified from the experience of compiling the BASE corpus. The attempt to provide standards across mark-up and therefore interchangeability of corpora will be successful only if there are tags that can encompass all that the encoder wishes to capture, and if there is standardisation of the usage and interpretation of these tags. Where the encoder has to make interpretative, non-empirically measured judgements on the relevance and nature of phenomena in the speech event, there will always be inconsistencies in their application. It must be recognised and made clear to the user that the tagged events deemed relevant for inclusion are not and cannot be empirically measured for their inclusion or non-inclusion, but are dependent on the interpretation of an individual. This supports the argument for linking the original audio and visual information to the transcript.

The TEI tagset for spoken language encoding was extended from that devised for written language. It is not necessarily the current tags themselves which need to be redeveloped, but the linear representation that is more problematic. The representation of non-linear events with a linear system will no doubt lead to inaccuracies. Using time markers and linking the marked up transcripts to the recordings that have been made could provide a more comprehensive, although still not a non-interpretative representation of the event. A time aligned multichannel audio and visual representation of the event alongside the marked up transcript may be able to capture more accurately more of the contextual and paralinguistic events, thus relieving the transcriber/encoder of much of the burden of interpretation, and, at the same time, better taking into account the temporal nature of spoken language.

The importance of providing a resource that is rich in information for the user while, at the same time, limiting or at least recognising the degree of transcriber interpretation and inconsistency has been stressed. Further, we claim that guidelines for the transcription and tagging of spoken language should not be based on the assumption that spoken language is a subset of written text encoding, but that they should provide an accurate reflection of the temporal, non-linear aspects of spoken language.

References

BASE corpus URL: http://www.rdg.ac.uk/AcaDepts/ll/base_corpus/

Cook, G. (1995) Theoretical issues: transcribing the untranscribable. In Leech, G., Myers G. & Thomas J. (eds.) *Spoken English on computer*. Harlow: Longman. 35-53.

Edwards, J. A. (1993) Principles and contrasting systems of discourse transcription. In Edwards, J. A. & Lampert, M. D. (eds.) *Talking data: transcription and coding in discourse research*. Hillsdale, New Jersey: Lawrence Erlbaum. 3-31.

MICASE Michigan corpus of academic spoken English URL: www.hti.umich.edu/m/micase/

Oxford Reference Online URL: www.oxfordreference.com/views/GLOBAL.html

Pocket Fowler's Modern English Usage. (1999) Allen, R. (ed.) Oxford University Press, Oxford Reference Online. URL: www.oxfordreference.com/views/BOOK_SEARCH.html?book=t30

Text Encoding Initiative URL: www.tei-c.org/P4X/ index.html