1. Overview

Throughout the development of corpus linguistics there has been a noticeable focus on analysing written language, and with written corpora now exceeding the one-billion-word mark, the possibilities for generating new insights into the way in which language is structured and used are both exciting and unprecedented. Spoken corpora, on the other hand, tend to be much smaller in size and thus often unable to offer the same level of recurrence of individual items and phrases when compared to their written counterparts. In addition, the analysis of spoken discourse as recorded in spoken corpora requires specific attention to elements beyond the text, such as intonation, gesture and discourse structure, which cannot easily be explored with the use of the kinds of frequency-based techniques used in the analysis of written corpora.

Nevertheless, spoken corpora provide a unique resource for the exploration of naturally occurring discourse, and the growing interest in the development of spoken corpora is testament to the value they provide to a diverse number of research communities. Following on from the early developments of relatively small spoken corpora in the 1960s, such as the London–Lund Corpus for example, the past two decades have seen major advances in the collection and development of spoken corpora, particularly in the English language, but not exclusively. Some examples of spoken corpora are the Cambridge and Nottingham Corpus of Discourse in English (CANCODE; McCarthy 1998), a five-million-word corpus collected mainly in Britain; the Limerick Corpus of Irish English (LCIE; Farr et al. 2004); the Hong Kong Corpus of Spoken English (HKCSE; Cheng and Warren 1999, 2000, 2002); the Michigan Corpus of Academic Spoken English (MICASE; Simpson et al. 2000) and the spoken component of the British National Corpus (BNC; see Lee, this volume). In addition, there is a growing interest in the development of spoken corpora of international varieties of English (e.g. the ICE corpora) and other languages, as well as of learner language (e.g. De Cock et al. 1998; Bolton et al. 2003). These corpora provide researchers with rich samples of spoken language-in-use which form the basis of new and emerging descriptions of naturally occurring discourse.
In recognition of the fact that spoken discourse is multi-modal in nature, a number of spoken corpora that are now being developed align audio and visual data streams with the transcript of a conversation. Examples of these include the Nottingham Multi-Modal Corpus (NMMC: a 250,000-word corpus of videoed single-speaker and dyadic discourse; see Knight et al. 2006; Carter and Adolphs 2008); the Augmented Multi-Party Interaction Corpus (AMI: a 100-hour meeting room video corpus; see Ashby et al. 2005) and the SCOTS Corpus (a corpus of Scottish texts and speech, comprised of audio files with aligned transcriptions; see Douglas 2003; Anderson and Beavan 2005; Anderson et al. 2007; see Thompson, this volume).

Research outputs based on the analysis of spoken corpora are wide-ranging and include, for example, descriptions of lexis and grammar (e.g. Biber et al. 1999; Carter and McCarthy 2006), discourse particles (Aijmer 2002), courtroom talk (Cotterill 2004), media discourse (O’Keeffe 2006) and healthcare communication (Adolphs et al. 2004; see also Biber, this volume, for further references). This research covers phenomena at utterance level as well as at the level of discourse. A number of studies start with the exploration of concordance outputs and frequency information as a point of entry into the data and carry out subsequent analyses at the level of discourse (e.g. McCarthy 1998), while others start with a discourse analytical approach followed by subsequent analyses of concordance data.

Before a spoken corpus can be subjected to this kind of analysis, the data has to be collected, transcribed and categorised in a way that allows the researcher to address specific research questions. This chapter deals with the basic steps that need to be taken when assembling a spoken corpus for research purposes. We will discuss the different considerations behind corpus design, data collection and associated issues of permission and ethics, transcription and representation of spoken discourse.

2. Corpus design

Issues of corpus design need to be addressed prior to any discussion of the content of the corpus, and the methods used to organise the data. Often, design and construction principles are locally determined (Conrad 2002: 77); however, Sinclair’s principles articulated in relation to corpus design can be seen as general guidelines for both spoken and written corpora (for similar prescriptions see Reppen and Simpson 2002: 93; Stuart 2005: 185; Wynne 2005). Sinclair (2005) sets out the following guidelines:

1. The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.
2. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.
3. Only those components of corpora which have been designed to be independently contrastive should be contrasted.
4. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
5. Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
6 Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.

7 The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.

8 The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.

9 Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.

10 A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

Some of these guidelines can be difficult to uphold because of the nature of language itself; a ‘population without limits, and a corpus is necessarily finite at any one point’ (Sinclair 2008: 30). The requirement for ensuring representativeness, balance and homogeneity in the design process is thus necessarily idealistic. They are also specific and relative to individual research aims, and thus have to be judged in relation to the different questions that are asked of the data.

With regard to the guidelines above, there are a number of issues that pertain specifically to the construction of spoken corpora. These are best described in relation to the fundamental stages of construction (see also Psathas and Anderson 1990; Leech et al. 1995; Lapadat and Lindsay 1999; Thompson 2005; Knight et al. 2006):

- Recording
- Transcribing, coding and mark-up
- Management and analysis

These have to be considered at the design stage and are best seen as interacting parts of a research system, with each stage influencing the next. The stage of recording data is determined by the type of analysis that is planned, which in turn determines the granularity and detail of transcription, coding and mark-up. It is therefore important to plan the development of a corpus carefully and to consider all practical and ethical issues that may arise (see Psathas and Anderson 1990 and Thompson 2005 for further discussion on the importance of planning). While the planning phase is an important stage of the design process, the approach that emerges during the process of construction is to be reconsidered and modified throughout.

It is important to note that the process of planning necessarily leads to an informed selection of discourse events that are being recorded, and thus it is impossible to create a ‘complete picture’ of discourse in corpora (Thompson 2005; also see Ochs 1979; Kendon 1982: 478–9; Cameron 2001: 71). This is true regardless of whether the corpus is of a specialist or of a more ‘general’ nature.

Given that the process of selection is unavoidable, it is the responsibility of the corpus constructor(s) to limit the potential detrimental effect that the selection may have on the representativeness, homogeneity and balance of the corpus. The use of a checklist or log to chart progress can be helpful, and acts as an invaluable point of reference for discussing anomalies or ‘gaps’ that may occur in the data, as well as accounting for interesting patterns that may be found in subsequent analyses.

We will now discuss the different stages in the construction of spoken corpora.
**Recording**

The recording stage is the data collection phase. At this stage it is important that all recordings are both suitable and rich enough in the information required for in-depth linguistic enquiry as well as being of a sufficiently high quality to be used and re-used in a corpus database (Knight and Adolphs 2006). It is therefore advisable to strive to collect data which is as accurate and exhaustive as it can be, capturing as much information from the content and context of the discursive environment as possible (Strassel and Cole 2006: 3). This involves documenting information about the participants, the location and the overall context in which the event takes place, as well as about the type of recording equipment that is being used, and the technical and physical specifications that are being applied to the recording itself. This is because the loss or omission of data cannot be easily rectified at a later date, since real-life communication cannot be authentically rehearsed and replicated.

It is ‘impossible to present a set of invariant rules about data collection because choices have to be made in light of the investigators’ goals’ (Cameron 2001: 29, also see O’Connell and Kowal 1999). The decisions made concerning the recording phase are thus specific to the research aims. Despite this, it is vital to ensure that ‘the text collection process for building a corpus … [is] principled’ (Reppen and Simpson 2002: 93) in order to achieve those aims.

As regards the equipment used for recording data, there are now a number of high-quality voice recorders available for recording spoken interaction. Most are now digital and recordings can be easily transferred to a PC or other device. Video recordings of spoken interactions are increasingly becoming an important alternative to pure sound recordings, as the resulting data offers further scope for analysis. Video recording equipment is thus starting to offer a useful alternative to sound recorders, and the availability of very small and unintrusive video recording equipment means that this procedure can now be used in a variety of different contexts (see Crabtree et al. 2006; Morrison et al. 2007; Tennent et al. 2008).

As part of the planning process for the recording stage it is essential that decisions are made determining the design of the recording process (i.e. what kind of data needs to be recorded and how much), as well as the physical conditions under which the recordings take place (i.e. the when and where the recording is to take place and the equipment that is being used).

Since the construction of spoken corpora is very expensive, the issue of cost effectiveness is an important factor when it comes to all stages of construction. Thompson (2005) highlights that there is a need to decide between ‘breadth’ and ‘depth’ of what is to be recorded. The cost–benefit consideration assesses the relative advantages between capturing large amounts of data (in terms of time, number of encounters or discourse contexts), the amount of detail added during the transcription and annotation phase, and the nature of analyses that might be generated using the recordings. In terms of the number of hours of recording needed to achieve a particular word count, in previous corpus development projects, such as the CANCODE project for example, one hour of recorded casual conversation accounted for approximately 10,000 words of transcribed data. This is only a very broad estimate as the number of words per hour recording depends on a range of different factors, including among others the discourse context and the rate of speech of the participants.

This discussion leads to the wider question of ‘How much data is enough?’, a complex and challenging question which requires a number of different perspectives to be considered.
At the heart of this question is the variable under investigation and a key ‘factor that affects how many different encounters you may have to record is how frequently the variable you are interested in occurs in talk’ (Cameron 2001: 28). Thus, the amount of data we need to record to analyse words or phrases that occur very frequently is less than we would need to study less frequent items. The study of minimal response tokens in discourse, such as *yeah, mmhm*, etc., which are very frequent in certain types of interaction, therefore requires less data to be collected than the study of more lexicalised patterns which function as response tokens and which may be less frequent, such as *that’s great* or *brilliant*.

### 3. Metadata

Apart from the process of recording the actual interaction between participants engaging in conversation, it is also important to collect and document further information about the event itself. Metadata, or ‘data about data’, is the conventional method used to do this. Burnard states that ‘without metadata the investigator has nothing but disconnected words of unknowable provenance or authenticity’ (2005: 31). Thus, metadata is critical to a corpus to help achieve the standards for *representativeness, balance and homogeneity* etc, outlined above (Sinclair 2005).

Burnard (2005) uses ‘metadata’ as an umbrella term which includes editorial, analytic, descriptive and administrative categories:

- Editorial metadata – providing information about the relationship between corpus components and their original source.
- Analytic metadata – providing information about the way in which corpus components have been interpreted and analysed.
- Descriptive metadata – providing classificatory information derived from internal or external properties of the corpus components.
- Administrative metadata – providing documentary information about the corpus itself, such as its title, its availability, its revision status, etc.

Metadata can be extremely useful when the corpus is shared and re-used by the community, and it also assists in the preservation of electronic texts. Metadata can be kept in a separate database or included as a ‘header’ at the start of each document (usually encoded through mark-up language). A separate database with this information makes it easier to compare different types of documents and has the distinct advantage that it can be further extended by other users of the same data. The documentation of the design rationale, as well as the various editorial processes that an individual text has been subjected to during the collection and archiving stages, allows other researchers to assess its suitability for their own research purposes, while at the same time enabling a critical evaluation of other studies that have drawn on this particular text to be carried out.

### Permission and ethics

Typical practice in addressing ethics on a professional or institutional level suggests that corpus developers should ensure that formal written consent is received from all participants involved, *a priori* to carrying out the recording. Conventionally, this consent stipulates how recordings are to take place and how data is presented, and outlines the
research purposes for which it is used (Leech et al. 1995; Thompson 2005). While a participant’s consent to record may be relatively easy to obtain, and commonly involves a signature on a consent form, it is important to ensure that this consent holds true at every stage of the corpus compilation process.

It is also important to consider consent to distribute recorded material. This is distinct from the consent to record and should include reference to the way in which recorded data is accessed and by whom, as well as the channel by which it may be distributed. Participants should be informed that once the data is distributed, it can be difficult or impossible to deal with requests for retraction of consent at a later stage. It is therefore important that systems used for distribution and access of data can deal with different levels of consent, and replay the data, accordingly, to different end users.

A further issue when dealing with ethics of recording relates to the notion of anonymity and the process that is applied to the data to ensure that anonymity can be maintained. Traditional approaches to corpus development emphasise the importance of striving for anonymity when developing records of discourse situations, as a means of protecting the identities of those involved in the interaction. To achieve this, the names of participants and third parties are often modified or completely omitted, along with any other details which may make the identity of referents obvious (see Du Bois et al. 1992, 1993). The quest for anonymity can also extend to the use of specific words or phrases used, as well as topics of discussion or particular opinions deemed as ‘sensitive’ or ‘in any way compromising to the subject’ (Wray et al. 1998: 10–11).

Issues of anonymity are more easily addressed when constructing text-based, monomodal corpora. If the data used is already in the public domain and freely available, no alterations to the texts included are usually required. Otherwise, permission needs to be obtained from the relevant authors or publishers of texts (copyright holders), and specific guidelines concerning anonymity can subsequently be discussed and addressed with the authors, and alterations to the data made as necessary. Similar procedures are involved when constructing spoken corpora which are based solely on transcripts of the recorded events. Modifications in relation to names, places and other identifiers can here be made at the transcription stage, or as a next step following initial transcription of data.

Anonymity is more problematic when it comes to audio or video records of conversations in corpora. Audio data is ‘raw’ as it captures vocalisations of a person, existing as an ‘audio fingerprint’, which is specific to an individual. This makes it relatively easy to identify participants when audio files are replayed. Alteration of vocal output for the purpose of anonymisation can make it difficult for the recording to be used for further phonetic and prosodic analysis and is thus generally not advisable. A similar problem arises with the use of video data. Although it is possible to shadow, blur or pixellate video data in order to conceal the identity of speakers (see Newton et al. 2005 for a method for pixellating video), these measures can be difficult to apply in practice (especially with large datasets). In addition, such measures obscure facial features of the individual, blurring distinctions between gestures and language forms. As a result, datasets may become unusable for certain lines of linguistic enquiry.

Considering the difficulties involved in the anonymisation of audio-visual data, it is important to discuss these issues fully with the participants prior to the recording, and to ensure that participants understand the nature of the recording and the format of distribution and access. And while those who agree for their day-to-day activities to be recorded for research purposes may not be concerned about anonymity, the issue of protecting the identity of third parties remains an ethical challenge with such data, as
does the issue of re-using and sharing contextually sensitive data recorded as part of multi-modal corpora.

This situation then raises a challenge that is central to the development of any corpus, namely, how can multi-site, multi-user, multi-source, multi-media datasets protect the rights of study participants as an integral part of the way in which they are constructed and used? Reconciling the desire for the traceability and probity of corpus data with the need for confidentiality and data protection requires serious consideration and should be addressed at the outset of any corpus development project.

4. How do I transcribe spoken data?

One of the biggest challenges in corpus linguistic research is probably the representation of spoken data. There is no doubt that the collection of spoken language is far more laborious than the collection of written samples, but the richness of this type of data can make the extra effort worthwhile. Unscripted, naturally occurring conversations can be particularly interesting for the study of spoken grammar and lexis, and for the analysis of the construction of meaning in interaction (McCarthy 1998; Carter 2004; Halliday 2004). However, the representation of spoken data is a major issue in this context as the recorded conversations have to undergo a transition from the spoken mode to the written before they can be included in a corpus. In transcribing spoken discourse we have to make various choices as to the amount of detail we wish to include in the written record. Since there are so many layers of detail that carry meaning in spoken interaction, this task can easily become a black hole (McCarthy 1998: 13) with a potentially infinite amount of contextual information to record (Cook 1990). The reason for this is that spoken interaction is essentially multi-modal in nature, featuring a careful interplay between textual, prosodic, gestural and environmental elements in the construction of meaning.

In terms of individual research projects it is therefore important to decide exactly on the purpose of the study, to determine what type of transcription is needed. It is advisable to identify the spoken features of interest at the outset, and to tailor the focus of the transcription accordingly. For example, a study of discourse structure might require the transcription to include overlaps but not detailed prosodic information. Transcription is thus ‘a theoretical process reflecting theoretical goals and definitions’ (Ochs 1979: 44, also see Edwards 1993 and Thompson 2005). It is best viewed as being ‘both interpretative and constructive’ (Lapadat and Lindsay 1999: 77; also see O’Connell and Kowal 1999: 104 and Cameron 2001).

At the same time there is a need to follow certain guidelines in the transcription in order to make them re-usable by the research community. This, in turn, would allow both the size and quality of corpus data available for linguistic research to be enhanced, without individuals or teams of researchers expending large amounts of time and resources in starting from scratch each time a spoken corpus is required.

There are now a number of different types of transcription conventions available, including those adopted by the Network of European Reference Corpora (NERC), which was used for the spoken component of the COBUILD project (Sinclair 1987). This transcription system contains four layers, ranging from basic orthographic representation to very detailed transcription, including information about prosody. Another set of guidelines for transcribing spoken data has been recommended by the Text Encoding Initiative (TEI) and has been applied, for example, to the British National
Corpus (BNC – see Sperberg-McQueen and Burnard 1994). These guidelines include the representation of structural, contextual, prosodic, temporal and kinesic elements of spoken interactions and provide a useful resource for the transcription of different levels of detail required to meet particular research goals.

The level of detail of transcription reflects the basic needs of the type of research that they are intended to inform. One of the corpora that has been transcribed to a particularly advanced level of detail is the London–Lund corpus (Svartvik 1990). Alongside the standard encoding of textual structure, speaker turns and overlaps, this corpus also includes prosodic information and has remained a valuable resource for a wide range of researchers over the years.

Wray et al. (1998) provide a rubric for some of the more universally used notations for orthographic transcription, which mark, for example, who is speaking and where interruptions, overlaps, backchannels and laughter occur in the discourse, as well as some basic distinct pronunciation variations. Basic textual transcription can be extended through the use of phonemic and phonetic transcription. Phonemic transcription is used to represent pronunciation. Phonetic transcription, on the other hand, uses the International Phonetic Alphabet (IPA – refer to Laver 1994; Wray et al. 1998; Canepari 2005), and indicates how specific successive sounds are used in a specific stretch of discourse.

The amount of information that can be included within a phonetic transcript is substantial, and as a result, again, it pays to be selective and to concentrate on marking only those features relevant to a specific research question.

**Layout of transcript**

Once decisions have been taken as to the features which are to be transcribed, and the level of granularity and detail of the information to be included, the next step is to decide on an appropriate layout of the transcription. There are many different possibilities for laying out a transcript but it is important to acknowledge that ‘there will always be something of a tension between validity and ease of reading’ (Graddol et al. 1994: 185).

The most commonly used format is still a linear representation of turns with varying degrees of detail in terms of overlapping speech, prosody and extra-linguistic information. The example in Figure 4.1 is taken from a sub-component of the Nottingham

<table>
<thead>
<tr>
<th>&lt;$2$&gt; &lt;$=&gt;$ Oh well I &lt;$=&gt; I’m just reading things at the moment and just$+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;$1$&gt; Right.</td>
</tr>
<tr>
<td>&lt;$2$&gt; +kind of vague$+$</td>
</tr>
<tr>
<td>&lt;$1$&gt; So what given that the amount of stuff on metaphor is huge?</td>
</tr>
<tr>
<td>&lt;$2$&gt; Yeah well I’ve been looking through some of the stuff on scientific metaphors and$+$</td>
</tr>
<tr>
<td>&lt;$1$&gt; Uh-huh.</td>
</tr>
<tr>
<td>&lt;$2$&gt; +er particularly how they’re used for educational purposes in$+$</td>
</tr>
<tr>
<td>&lt;$1$&gt; Right.</td>
</tr>
<tr>
<td>&lt;$2$&gt; +explaining concepts erm$+$</td>
</tr>
<tr>
<td>&lt;$1$&gt; Yeah.</td>
</tr>
</tbody>
</table>

*Figure 4.1* An example of transcribed speech, taken from the NMMC.
Multi-Modal Corpus (NMMC). While this corpus is fully aligned with audio and video streams, the basic transcriptions are formatted as shown.

In Figure 4.1, speakers are denoted by < $1 > and < $2 > tags, false starts are framed by < $ = > and < \$ = >, while interruptions are indicated by the presence of the + tag.

The transcript seen in Figure 4.1 presents the data linearly, ordering the conversation in a temporal way, rather like a conventional drama script. It is relevant to note that, using a linear format of transcription, it is particularly difficult to show speaker overlap, and for this reason some prefer to use different columns and thus separate transcripts according to who is speaking (see Thompson 2005). Therefore, since speech is rarely ‘orderly’ in the sense that one speaker speaks at a time, linear transcription may be seen as a misrepresentation of discourse structure (Graddol et al. 1994: 182). This criticism is particularly resonant if, for example, four or five speakers are present in a conversation and where there is a high level of simultaneous speech.

The use of column transcripts (Figure 4.2) allows for a better representation of overlapping speech, presenting contributions from each speaker on the same line rather than with one positioned after the other (for further discussion on column transcripts see Graddol et al. 1994: 183; Thompson 2005).

A final, alternative method of representing speech in a transcript can be seen in Figure 4.3 (based on Dahlmann and Adolphs 2007; see also Dahlmann and Adolphs 2009).

Here the speech is presented as a musical score, with the talk of each speaker arranged on an individual line (or track) on the score. Speech is arranged according to the time at which it occurred. Overlapping contributions are indicated (as with Figure 4.2) as text which is positioned at the same point along the score, the time line, across each individual speaker track. The contributions of multiple speakers can be represented using this method of transcription. This method of transcription is based on a similar principle to that used in transcription and coding software, such as Anvil (Kipp 2001) and DRS (French et al. 2006).

Ongoing advancements in the representation and alignment of different data streams have started to provide possibilities for studying spoken discourse in an integrated framework including textual, prosodic and video data. The alignment of the different elements and the software needed to analyse such a multi-modal resource are still in the early stages of development, and at the present time it is probably beyond the scope of

<table>
<thead>
<tr>
<th>&lt;$1$&gt;</th>
<th>&lt;$2$&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right.</td>
<td>&lt;$=&gt;$ Oh well I &lt;$=&gt;$ I’m just reading things at the moment and just+</td>
</tr>
<tr>
<td>So what given that the amount of stuff on metaphor is huge?</td>
<td>+kind of vague+</td>
</tr>
<tr>
<td>Yeah well I’ve been looking through some of the stuff on scientific metaphors and+</td>
<td></td>
</tr>
<tr>
<td>Uh-huh.</td>
<td>+er particularly how they’re used for educational purposes in+</td>
</tr>
<tr>
<td>Right.</td>
<td>+explaining concepts erm+</td>
</tr>
<tr>
<td>Yeah.</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.2** A column-based transcript.
the majority of individual corpus projects to develop a searchable resource that includes the kind of dynamic representation that would address the need for a less linear layout of transcription.

**Coding spoken data**

The coding stage refers to ‘the assignment of events to stipulated symbolic categories’ (Bird and Liberman 2001: 26). This is the stage where qualitative records of events start to become quantifiable, as specific items that are relevant to the variables under consideration are marked up for future analyses (Scholfield 1995: 46). The coding stage is essentially a development of the transcription stage, providing further detail to the basic systems of annotation and mark-up applied through the use of transcription notation. The coding stage thus operates at a higher level of abstraction compared to the transcription stage, and may include, among others, annotation of grammatical, semantic, pragmatic or discoursal features or categories.

Coding is a key part of the process of annotating language resources. This process is often undertaken with the use of coding software. The majority of corpora include some type of annotation as they allow corpora to be navigated in an automated way.

Early standards for the mark-up of corpora, known as the SGML (Standard Generalised Mark-up Language, which has been succeeded by XML), were developed in the 1980s. SGML was traditionally used for marking up features such as line breaks and paragraph boundaries, typeface and page layout, providing standards for structuring both transcription and annotation. SGML is used in the 100-million-word BNC corpus. With modern advances in technology, and associated advances in the sophistication of corpora and corpora tools, movements towards a redefinition of SGML have been prompted.
Most notably, in the late 1990s, early advances were made for what was termed the CES, *Corpus Encoding Standard*. The CES promised to provide ‘encoding conventions for linguistic corpora designed to be optimally suited for use in language engineering and to serve as a widely accepted set of encoding standards for corpus-based work’ (Ide 1998: 1). The CES planned to provide a set of coding conventions to cater for all corpora of any size and/or form (i.e. spoken or written corpora), comprising of an XML-based mark-up language (known as XCES). The standardised nature of these conventions aims to allow coded data and related analyses to be re-used and transferred across different corpora. However, the majority of corpora today still use modified versions of the SGML, or adopt their own conventions (mainly based on XML, as used in the BNC).

5. Analysing spoken corpora

There are a number of chapters in this volume which deal with the analysis of spoken corpora, so this issue will not be covered here in any detail. Other chapters in this volume discuss the implications of the unique nature of spoken discourse in terms of its implications for any type of spoken corpus analysis (see, for example, chapters by Evison, Vo and Carter, Rühlemann, this volume). The discourse level frameworks that may be of use for the analysis of spoken corpora are not necessarily compatible with the kind of concordance-based and frequency-driven analysis that is often used in large-scale lexicography studies. One of the key differences between spoken and written corpora is that most spoken discourse is collaborative in nature and as such it is more fluid and marked by emerging and changing orientations of the participants (McCarthy 1998).

Yet it is important to identify external categories for grouping transcripts in a corpus, especially when we are concerned with analysing levels of formality and other functions which need to be judged against the wider context of the encounter. This is often more straightforward when dealing with written texts as many of the genres that tend to form the basis of large written corpora are readily recognised as belonging to a particular category, such as fiction versus non-fiction, letters versus e-mails, etc. The group membership of such texts is more clearly demarcated than is the case with the majority of spoken discourse. The development of suitable frameworks for analysing spoken corpus data is thus particularly complex and further research is needed to explore and evaluate ways in which the analysis of concordance data and discourse phenomena can be fully integrated.

The development of techniques and tools to record, store and analyse naturally occurring interaction in spoken corpora has revolutionised the way in which we describe language and human interaction. Spoken corpora serve as an invaluable resource for the research of a large range of diverse communities and disciplines, including computer scientists, social scientists and researchers in the arts and humanities, policy makers and publishers. In order to be able to share resources across these diverse communities, it is important that spoken corpora are developed in a way that enables re-usability. This can be achieved through the use of guidelines and frameworks for recording, representing and replaying spoken discourse. In this chapter we have outlined some of the issues that surround these three stages of spoken corpus development and analysis. As advances in technology allow us to develop new kinds of spoken corpora, which include audio-visual data-streams, as well as a much richer description of contextual variables, it will become increasingly important to agree on conventions for recording and representing...
this kind of data, and the associated metadata. Similarly, advances in voice-to-text software may ease the burden of transcription, but will also rely heavily on the ability to follow clearly articulated conventions for coding and transcribing communicative events. Adherence to agreed conventions of this kind, especially when developing new kinds of multi-modal and contextually enhanced spoken corpora, will significantly extend the scope of spoken corpus linguistics in the future.

Further reading


References


Knight, D. and Adolphs, S. (2006) Text, Talk and Corpus Analysis [academic online module, restricted access], University of Nottingham, UK.


