17 The Transcription of Discourse

JANE A. EDWARDS

0 Introduction

Recordings are essential tools in discourse research, but are not sufficient by themselves for the systematic examination of interaction. It is simply impossible to hold in mind the transient, highly multidimensional, and often overlapping events of an interaction as they unfold in real time.

For this reason, transcripts are invaluable. They provide a distillation of the fleeting events of an interaction, frozen in time, freed from extraneous detail, and expressed in categories of interest to the researcher.

As useful as they are, transcripts are not unbiased representations of the data. Far from being exhaustive and objective, they are inherently selective and interpretive. The researcher chooses what types of information to preserve, which descriptive categories to use, and how to display the information in the written and spatial medium of a transcript. Each of these choices can affect the researcher's perceptions of the structure of the interaction (Ochs 1979), making some types of regularities easier to detect in the data and others more difficult.

For example, arranging utterances by different speakers in separate columns (**column-based format**) gives the impression of asymmetry between the speakers, with the leftmost speaker appearing to be the most dominant. In contrast, arranging them one above the other in a single column (**vertical format**) gives the impression of interdependence and equal dominance. Vertical format is useful for conversations between adults of equal status, but would be misleading for interactions between adults and very young children, which tend to be child-centered and therefore child-dominated. For those interactions, Ochs (1979) recommended using column-based format, with the child's column leftmost.

The best choice of conventions in a given instance depends on the nature of the interaction, the theoretical framework, and the research question. In fact, Mishler (1991) presents several examples from published literature in which the same interaction was transcribed differently for contrasting purposes – in some cases, even by the same researcher at different times.

Transcription is an open-ended process. A transcript changes as the researcher's insights become progressively refined (Ehlich 1993; Ehlich and Switalla 1976; Gumperz and Berenz 1993). To ensure that significant but subtle factors are not left out, it is important to listen to recordings repeatedly throughout the course of a study and to update the transcript to reflect developing insights.

This chapter focuses on the interplay of theory and methods. It begins with general principles of design which are relevant regardless of research question. Next it surveys alternative conventions and their underlying assumptions. Then discussion turns to practical issues of applying transcription conventions to actual data in a consistent and efficient manner. Finally, it reviews some historical precursors to transcription, and summarizes developing standards and future trends.

1 General Principles

1.1 Encoding processes

Transcripts contain basically three types of encoding, termed here **transcription**, **coding**, and **markup**.

Transcription is the process of capturing the flow of discourse events in a written and spatial medium. This includes primarily: *who* said *what*, *to whom*, *in what manner*, and *under what circumstances*. It involves the kinds of information found in the script of a play, only with more systematic and detailed specification.

Many categories found useful in discourse research are interpretive in nature, rather than being tied strictly to objective physical measurements. Interpretive categories are necessary because the goal of discourse research is to capture aspects of interaction as they are perceived by human participants, and these are not yet specifiable by means of physical parameters. For example, perceived pause length depends not only on physically measurable time, but also on speech rate, location of the pause (e.g. within a clause, between clauses, between speaker turns), and other factors. There are many distinctions of interest to discourse researchers which have less obvious relationships to physically measurable properties. This is not a problem, so long as they can be applied reliably by human observers, on the basis of clearly specified criteria.

At a certain level of abstraction and complexity, transcribing shades into coding (also called "annotation" or "tagging"), which is even more interpretive and more closely tied to particular theoretical frameworks. Some examples of coding include: syntactic categories (such as nouns, verbs, adjectives, etc.), semantic distinctions (e.g. motion verbs, manner verbs), or pragmatic acts (e.g. directive, prohibition, claim). Coding establishes equivalence classes which expedite analysis and computer search by enabling otherwise dissimilar items to be efficiently brought together for closer examination.

Mark-up concerns format-relevant specifications rather than content. It is intended to be interpreted by a typesetter or computer software for such purposes as proper segmentation of the text and cataloging of its parts, in the service of formatting, retrieval, tabulation, or related processes. It also plays a central role in data exchange and emergent encoding standards, discussed in the closing section of this chapter.

1.2 Other representations

An important technological advance in recent years has been the ability to link transcripts to recordings. Bloom (1993) linked transcripts to videotaped recordings by means of SMTPE time codes, for purposes of navigating through the recordings more easily.

Some projects link transcripts to digitized audiorecordings (e.g. the MARSEC project, described in Knowles 1995; the HCRC project, described in Thompson et al. 1995, and the ToBI project, described in Roach and Arnfield 1995; the "Transcriber" interface, described in Barras et al. 1998). The listener can relisten to any utterance (or turn), with a simple click of a mouse.

Some projects link transcripts to digitized videorecordings (e.g. the SignStream project, described in Neidle and MacLaughlin 1998), enabling systematic encoding and analysis of visual language data (e.g. sign language and data). Duranti (1997) mentions the value of photographs, maps, and diagrams for providing supplementary information about an interaction. These other representations do not usually affect the form or content of transcripts, but are simply alternative perspectives on the same data.

The focus of this chapter is the representation of spoken language in a written/spatial medium. There are three general design principles which are pertinent regardless of research question. These are principles of category design, computational tractability, and readability.

1.3 Principles of category design

Transcription and coding systems are divided into subdomains (e.g. pause length, intonation contour, syntactic category). The categories used in describing a particular subdomain (e.g. "short" or "long" pause) function as alternatives to one another. That is, they constitute a "contrast set." To be descriptively useful, the categories within each contrast set must satisfy three general principles:

1 They must be *systematically discriminable*. That is, for each portion of the interaction it must be clear whether or not a given category applies. Category membership can be based on either defining characteristics or similarity to prototypical exemplars.
2 They must be *exhaustive*. That is, for each relevant aspect or event in the data, there must be a category which fits (even if, in hopefully rare cases, it is only "miscellaneous").
3 They must be *usefully contrastive*. That is, they must be focused on distinctions of importance to the research question. For example, a "short" pause in information flow in monologues might be 0.2 seconds, whereas a "short" pause in research on turn-taking might be 0.5 seconds.

The categories within a contrast set usually cannot be interpreted without knowledge of the number and type of other categories in that set. Firth (1957: 227) expressed this property as follows: "The 'meaning' of the grammatical category noun in a grammatical system of, say, three word classes, *noun*, *verb*, and *particle*, is different

from the meaning of the category *noun* in a system of five classes in which *adjective* and *pronoun* are formally distinguished from the *noun*, *verb*, and *particle*."

This is true also when interpreting symbols in a transcript. Punctuation marks are convenient and hence ubiquitous in transcripts, but may not serve the same purposes in all projects. They may be used to delimit different types of units (e.g. intonational, syntactic, pragmatic) or to signify different categories of a particular type. For example, a comma might denote a "level" utterance-final contour in one system and "nonrising" utterance-final contour in another. The only guarantee of comparability is a check of how the conventions were specified by the original sources. (For instances of noncomparability in archive data, see Edwards 1989, 1992b.)

1.4 Principles of computational tractability

For purposes of computer manipulation (e.g. search, data exchange, or flexible reformatting), the single most important design principle is that *similar instances be encoded in predictably similar ways*.

Systematic encoding is important for uniform computer retrieval. Whereas a person can easily recognize that *cuz* and *'cause* are variant encodings of the same word, the computer will treat them as totally different words, unless special provisions are made establishing their equivalence. If a researcher searches the data for only one variant, the results might be unrepresentative and misleading. There are several ways of minimizing this risk: equivalence tables external to the transcript, normalizing tags inserted in the text, or generating exhaustive lists of word forms in the corpus, checking for variants, and including them explicitly in search commands. (Principles involved in computerized archives are discussed in greater detail in Edwards 1992a, 1993a, 1995.)

Systematic encoding is also important for enabling the same data to be flexibly reformatted for different research purposes. This is an increasingly important capability as data become more widely shared across research groups with different goals. (This is discussed in the final section, with reference to emerging standards for encoding and data exchange.)

1.5 Principles of visual display

For many researchers, it is essential to be able to read easily through transcripts a line at a time to get a feel for the data, and to generate intuitive hypotheses for closer testing. Line-by-line reading is often also needed for adding annotations of various types. These activities require readers to hold a multitude of detail in mind while acting on it in some way – processes which can be greatly helped by having the lines be easily readable by humans. Even if the data are to be processed by computer, readability is helpful for minimizing error in data entry and for error checking.

In approaching a transcript, readers necessarily bring with them strategies developed in the course of extensive experience with other types of written materials (e.g. books, newspapers, train schedules, advertisements, and personal letters). It makes sense for transcript designers to draw upon what readers already know and expect from written

media, both because readers are good at extracting information in these ways, and because strategies based on reading habits and perceptual predispositions may be difficult to suspend even if it is desired to do so.

Written materials often make systematic use of two cues in particular: space and visual prominence. For example, chapter titles are expected to be printed in a large font, possibly centered or ruled off, and placed above the body of the text at some vertical distance, rather than, say, being embedded in the body of a text and in the same font size and type.

In looking across transcripts of various types, one notices some recurring strategies using these two cues for highlighting information and indicating relationships of interest. Six of them are summarized here. Some of these overlap with properties discussed by Du Bois (1991) and Johansson (1991). These are discussed with examples in Edwards (1992b, 1993b).

1 *Proximity of related events:* Events or types of information which are more closely related to each other are placed spatially nearer to each other than those which are less closely related. For example, prosodic information, such as prominent syllable stress, is often indicated by a mark (e.g. an apostrophe or an asterisk) placed immediately before the relevant syllable (cf. Svartvik and Quirk 1980; Gumperz and Berenz 1993).
2 *Visual separability of unlike events:* Events or types of information which are qualitatively different from each other (e.g. spoken words and researcher comments, codes, and categories) tend to be encoded in distinctly different ways. For example, codes may be enclosed in parentheses, or expressed as nonalphabetic characters (rather than alphabetic) or upper case letters (in contrast to lower case). This enables the reader to know what kind of information is about to be read before actually reading it, and thereby speeds reading and minimizes false attributions (e.g. perceiving a word as having been part of the speech stream, when it was really part of a metacomment or code).
3 *Time-space iconicity:* Temporally prior events are encountered earlier on the page (top to bottom or left to right) than temporally later events. This can include utterances, gestures, door slams, laughs, coughs, and so forth.
4 *Logical priority:* Logically prerequisite information for interpreting utterances tends to be encountered earlier on the page than the utterance(s) for which it is relevant. Information concerning the circumstances of data gathering and the relationships among the speakers tends to be given at the top of the transcript, whereas changes in circumstances or activities during the course of the interaction tend to precede the utterances they contextualize or potentially influence.
5 *Mnemonic marking:* Coded categories are encoded either in directly interpretable abbreviations or in symbolically iconic ways in order to expedite recovery of their meaning during rapid reading. An example of this is the use of a slash (/) for rising intonation and a backslash (\) for falling tone, rather than vice versa or instead of an arbitrary numerical code (e.g. "7"), as in the following example:

(1) **London–Lund Corpus, text 1.3 (Svartvik and Quirk 1980):**

```
1 3  7212280 1 1 A  11  and at ^h\/ome#.                    /
1 3  7212290 1 1 A  11  she's not a ^b\it the way she is at c/ollege#  /
```

Direct readability is also helped by using conventions already known from other written contexts. Du Bois (1991) notes that a number of transcription conventions derive from literary conventions found in novels and plays. Some examples are the use of three dots (. . .) for pauses, or a dash (–) for interrupted thoughts or utterances.

6 *Efficiency and compactness:* Coded distinctions should be marked with as few symbols as possible (e.g. nonredundantly, using short abbreviations), so long as meaning is easily recoverable (i.e. encoded mnemonically). This serves to minimize nonessential and distracting clutter in the transcript. For example, the use of a slash (/) for rising tone is more compact and efficiently read than would be the use of the full word, "rising." The encoding of spoken words and prosodic information on the same line instead of on separate lines is also a type of compactness.

All transcripts contain at least some of these devices. They vary in the specific types of information they foreground, background, and interrelate.

We turn now to a brief survey of some of their differences and how they relate to underlying theories of interaction.

2 Contrasting Methods and Assumptions

There are primarily two types of decisions which affect researcher perceptions in transcripts: format-based decisions and content-based decisions.

2.1 Format-based decisions

Format-based decisions are those involving layout and symbol choice. If the data have been systematically encoded, it is possible to convert between these choices by means of computer programs.

2.1.1 Layout

The main layout considerations are: arrangement of speaker turns (i.e. vertical, column-based, and partiture formats) and placement of codes and researcher comments relative to the discourse events they clarify (multilinear, column-based, and interspersed formats).

2.1.1.1 Arrangement of speaker turns

The three main choices are vertical format, column-based format, and "partiture" or musical score notation.

As mentioned in the opening section, **vertical** format implies symmetry and equal dominance of speakers, whereas **column-based** format gives the impression (due to left-to-right reading bias) that the speaker whose utterances are leftmost is more dominant in the interaction. Vertical and column-based format are similar in highlighting the incremental aspect of interaction – that is, the fact that discourse is built

up gradually out of smaller units, contributed one at a time. For both vertical and column-based formats, time is preserved vertically from the top to the bottom of the transcript, and to a more limited degree from left to right. In both of these formats, overlapping stretches of speech are signaled by marks placed before and after the parts which overlap. In vertical format, indentation and brackets may also be used:

(2) **From Du Bois et al. (1993: 49):**
Jeff: That's all it does.
It doesn't [even] reach a conclusion.
Sarah: [mhm]

Similar conventions are found in Jefferson (1984) and Psathas (1990).

In **partiture** notation (e.g. Ehlich 1993; Erickson and Shultz 1982), turns by different speakers are put on different lines in a manner similar to instruments on a musical score. This format highlights the collaborative aspects of interaction – that is, that discourse is a unified "accomplishment" achieved jointly by multiple participants. Partiture preserves both time and simultaneity in a directly interpretable manner (which is why it is useful for musical notation), and eases the study of temporal coordination, overlaps, and conjoined action. Its disadvantage is that it may require specialized software (such as HIAT2 – Ehlich 1993) to ensure the alignment is preserved whenever changes are made. Also, the boundaries of turns are less prominent than in vertical format.

2.1.1.2 Placement of codes and comments

The three main possibilities are multitier, column-based, and interspersed formats.

Multitier (or interlinear or multilayer) format: The most widespread format involves placing codes or annotations on separate lines beneath the datum they clarify. It was used in the Berkeley Crosslinguistic Language Acquisition project (Slobin 1967), which was one of the earliest coded computer corpora:

(3) 2;0a 002A ALL CLOSE UP Q. {notes back of bus is open}.
2;0a 002B = -NO -V PC -YN QT
2;0a 002C == CPSP {close-up} PERF {all}
2;0a 002D (Q+POT PERF (C (P SP))) #

In this example, the top tier contains the child's utterance and contextual comments, and subsequent tiers contain syntactic and semantic analyses of the utterance. In the ChiLDES archive of child language data (MacWhinney 1995), the top tier contains the child utterance and subsequent tiers contain phonetic, prosodic, gestural-proxemic, or other types of information. In multilingual studies, the top line is used for the utterance, the second line for an interlinear morpheme by morpheme gloss, and the third line for a free translation (discussed in Duranti 1997: 158; see also Du Bois et al. 1993). In the ToBI (Tones and Break Indices) project, concerning prosody, the orthographic rendering of the utterance is followed by three tiers: a tone tier (for specifying the tonal properties of the fundamental frequency contour), a break index tier (for specifying the degree of disjuncture between adjacent words), and a miscellaneous tier (for additional notations). Multitier format is also used in the MATE

(Multilevel Annotation Tools Engineering) project, a large European project concerned with establishing standards for encoding for speech and language technologies using corpus data.

Multitier format enables users to access each type of information independently from the others in an efficient manner. However, this format also has some drawbacks. Unless there is a strict, sequential, one-to-one correspondence between main-tier elements and elements on the other tiers, additional provisions are needed (such as numerical indices) to indicate the word(s) to which each code or comment pertains (that is, its "scope"). Otherwise it is not possible to convert data automatically from this format into other formats, and the data are less efficient to use where it is necessary to correlate individual codes from different tiers (see Edwards 1993b for further discussion).

Also, it is generally less useful in discourse research than other methods because it requires the reader to combine information from multiple tiers while reading through the transcript, and because it spreads the information out to an extent which can make it difficult to get a sense of the overall flow of the interaction.

Column-based format: Rather than arranging the clarifying information vertically in separate lines beneath the discourse event, codes may be placed in separate columns, as in the following example from Knowles (1995: 210) (another example is the Control Exchange Code described in Lampert and Ervin-Tripp 1993):

(4)

phon_id	orthog	dpron	cpron	prosody
525400	the	Di	D@	the
525410	gratitude	‘gr&tItjud	‘gr&tItjud	,gratitude
525420	that	D&t	D@t~p	that
525430	millions	‘mIll@@nz	‘mill@nz	~millions
525440	feel	fil	‘fil	*feel
525450	towards	t@’wOdz	t@’wOdz	to’wards
525460	him	hIm	Im	him

If the codes are mostly short, column-based format can be scanned more easily than multitier format because it is spatially more compact.

Column-based coding is also preferable when annotating interactions which require a vertical arrangement of speaker turns, such as interactions with very young children. For this reason, Bloom's (1993) transcript contained columns for child utterances, coding of child utterances, adult utterances, coding of adult utterances, and coding of child play and child affect. The columns in her transcript are of a comfortable width for reading, and it is relatively easy to ignore the coding columns to gain a sense of the flow of events, or to focus on the coding directly.

Interspersed format: Where codes are short and easily distinguished from words, they may be placed next to the item they refer to, on the same line (i.e. "interspersed"), as in this example from the London–Oslo–Bergen corpus:

(5) A10 95 ^ the_ATI king’s_NN$ counsellors_NNS couched_VBD their_PP$
A10 95 communiqué_NN in_IN vague_JJ terms_NNS ._.

Brackets may be used to indicate scope for codes if they refer to more than one word. Gumperz and Berenz (1993) indicate such things as increasing or decreasing tempo or loudness across multiple words in a turn in this manner.

Information is encoded in both the horizontal and vertical planes in the following example, from the Penn Treebank, in which the vertical dimension indicates larger syntactic units:

(6) **(from Marcus et al. 1993):**

```
((S
  (NP Battle-tested industrial managers
      here)
   always
  (VP buck
      up
     (NP nervous newcomers)
     (PP with
        (NP the tale
           (PP of
              (NP (NP the
                      (ADJP first
                            (PP of
                                (NP their countrymen)))
                      (S (NP *)
                         to
                        (VP visit
                            (NP Mexico))))
                  ,
                  (NP (NP a boatload
                          (PP of
                             (NP (NP warriors)
                                    (VP-1 blown
                                       ashore
                                       (ADVP (NP 375 years)
                                             ago)))))
                  (VP-1 *pseudo-attach*)))))))
.)
```

2.1.2 Symbol choice

The choice of symbols is mainly dictated by the principles of readability already discussed above. Examples (5) and (6) are readable despite their dense amount of information. This is due to the visual separability of upper and lower case letters, and to a consistent ordering of codes relative to the words they describe. That is, in (5), the codes follow the words; in (6) they precede the words.

With systematic encoding and appropriate software, it is possible for short codes, such as those in example (5), to serve as references for entire data structures, as is possible using the methods of the Text Encoding Initiative (TEI) (described more

fully in McEnery and Wilson 1997: 28). Alternatively, tags can be left out of the text entirely, by numbering the words in a text sequentially and linking data structures to their identification numbers (as in Du Bois and Schuetze-Coburn 1993).

2.1.3 Converting between formats

With consistent encoding and appropriate software, it is possible to translate easily between alternative formats, and to protect the user from clutter by displaying only those parts of the information which are needed for specific research purposes.

This kind of flexibility of representation was the main motivation behind the TEI, a large international project designed to establish guidelines for text exchange among the various fields using textual data, discussed in greater detail in the final section of this chapter.

2.2 Content-based decisions

Unlike "format-based" biases, "content-based" biases cannot be adjusted by computer program. To change these, it is often necessary to change both the number and type of categories used to encode the interaction. It is the content-based aspects which most distinguish different systems and which are primarily of interest with respect to the issue of the impact of theory on methods.

Content-based decisions are of mainly two types: the sorts of information which are encoded, and the descriptive categories used.

Though transcripts differ across many dimensions, some of the domains in which transcripts differ most often (and which are often the most theory-relevant) are the following:

- words
- units of analysis
- pauses
- prosody
- rhythm and coordination
- turn-taking
- nonverbal aspects and events.

2.2.1 Words

In encoding words, there are mainly two types of decision to be made. The first is whether standard orthography is sufficient, or whether to preserve nuances of pronunciation (e.g. regional accents or idiolects). If details of pronunciation are to be preserved, the second choice is whether to use phonemic or phonetic transcription (which is rigorous but requires some special training) or modified orthography (which requires less training but is also less precise).

Because English spelling has many inconsistencies, modified orthography is often ambiguous. It is also likely to be less accessible for nonnative speakers of English than for native speakers. In addition, it suggests departure from the educated standard

(Duranti 1997) and may cause participants to appear less educated or intelligent (Gumperz and Berenz 1993; Preston 1985). Where modified orthography is used for words which are pronounced in the standard way (e.g. "wuz" in place of "was"), it serves to imply a manner of speaking without actually adding precision regarding pronunciation. Duranti (1997) observes that modified orthography may serve in some cases to remind researchers of specific aspects of a recording which they know intimately, rather than encoding events in a manner precisely interpretable by those who have not heard the recording. (For further discussion, see Duranti 1997; Edwards 1992c.)

2.2.2 Units of analysis

Next the researcher must decide how to subdivide the text into units for purposes of analysis. Should the unit of analysis be an idea unit, a unit containing a predicate, a speaker turn, a unit bounded by pauses or uttered under a completed intonational contour, or some combination of these? Should text be subdivided into paragraphs or episodes? These are just a few of the possibilities.

This choice will determine which dimensions of structure are highlighted for purposes of analysis (e.g. prosody, syntax, information packaging), as well as the relevant scope of descriptive codes. (For further discussion, see Edwards 1993b; Lampert and Ervin-Tripp 1993.)

This choice affects the location of line breaks. In some transcription systems, line breaks occur before each intonation or ideational unit (as in Du Bois et al. 1993). Where analysis is focused on turn-taking, line breaks may be less common, perhaps occurring only between turns, or for long utterances (to keep them on the screen or page).

The unit of analysis also has implications for the temporal organization of the transcript. In the ChiLDES archive, utterances are the primary units of analysis. Gestures are treated as clarifying information, tied to specific utterances. They are placed on subordinate tiers beneath the utterances they are believed to clarify. If the gesture occurs before the utterance, this is indicated by adding the tag "<bef>" to the gestural-proxemic tier. Time is preserved spatially only for utterances in that format. Where a gesture or event is deemed relevant to more than one utterance, it is duplicated for each utterance (without notation distinguishing this case from the case in which the gesture itself is repeated in the interaction). This introduces ambiguity, and hinders automatic conversion from this format to others.

An alternative approach is to place verbal and nonverbal communication events in the transcript in order of occurrence. This approach is more theory-neutral because researchers are not required to guess the scope of relevance of nonverbal events (as is required in the former approach). In addition, having utterances and nonverbal acts in chronological order provides a more immediate sense of the flow of an interaction. This second approach is the more common in discourse research (e.g. Bloom 1973; Ehlich 1993; Jefferson 1984; Psathas 1990; Gumperz and Berenz 1993).

2.2.3 Pauses

Some researchers measure pauses to the nearest tenth of a second (Jefferson 1984). However, a pause may seem longer if embedded in rapid speech than if embedded in

slower speech. For this reason, some researchers quantify pauses as the number of beats of silence, based on the speaker's preceding speech rate. If all participants have the same speech rate, these approaches will be equivalent.

The perceived length of a given pause is also affected by its location in the discourse. It may seem longer if it is within an utterance than between turns by different speakers. Pauses which are longer or shorter than expected for a given location may be highly significant to interactants, indicating communicative strain or high rapport depending on communicative norms of that social group (as discussed by Erickson and Shultz 1982; Tannen 1984). For this reason, some researchers include normative judgments in their estimates of pause length. (To avoid circularity, communicative significance is established independently of the pause.)

Some systems explicitly mark all detectable pauses (e.g. Chafe 1993; Du Bois et al. 1993), while others mark only pauses that depart strongly from expectation (e.g. Ehlich 1993; Gumperz and Berenz 1993).

Even if the pause is measured in tenths of a second, its classification as "short" or "medium" depends on the research purpose. Researchers concerned with turn-taking smoothness may consider a "short" pause to be 0.5 seconds, while those interested in information packaging may consider it to be 0.2 seconds.

Another issue is the positioning of the pause relative to units of analysis. For monologs, it is sufficient to adopt a consistent convention, such as putting the pause at the beginning of each intonation unit (e.g. Chafe 1987, 1993). For dialogs, decisions are needed regarding who is responsible for an interturn pause. If a pause is perceived as "shared" by interactants, it makes sense to place it midway between the two turns (e.g. on a separate line in vertical format). If the first speaker asks a question and the second speaker says nothing, the pause may signal reticence. In that case, there is some logic to viewing it as belonging to the second speaker and transcribing it as if it is an empty turn (Tannen 1981).

All of these categories are potentially useful in some contexts. It is important simply to be aware of their interpretive nature (Tannen 1981) and to make allowances for their biases in the results.

2.2.4 Prosody

Prosodic features are properties that "generally extend over stretches of utterances longer than just one sound" (Cruttenden 1997: 1). These include such things as perceived duration, prominence, and intonation. These are perceptual/linguistic rather than acoustic phenomena. Although they are related to objectively measurable properties, the correspondence is far from perfect.

Listeners make many adjustments which acoustic measuring machines do not. There are far more frequency variations in the speech signal than are noticed by the listener (see, for example, Couper-Kuhlen 1986: 7). An utterance may be sprinkled with sudden high frequencies at high vowels (e.g. /i/) and silent spots at devoiced stop consonants (e.g. /p/) (Cruttenden 1997), but somehow the listener looks past these perturbations and perceives what seem to be reasonably smooth frequency contours.

Seemingly simple categories such as "rising intonation" actually cover a wide variety of acoustic contours. Contours may stretch over utterances of different lengths, or

have differing numbers of pitch peaks or different speeds of pitch change, and still be judged as belonging to the same contour category. These adjustments rely on norms:

> As Crystal (1975) has pointed out, we apparently do use norms or standards in auditory perception. For one, we can form a notion of "natural speaking level" and are able to determine (regardless of individual voice range) whether someone is speaking near the top or the bottom of his/her voice. (Couper-Kuhlen 1986: 9)

Since discourse researchers wish to describe interactions in categories which are as similar as possible to perceptions by participants, it is necessary to use interpretive categories. A variety of interpretive categories has been found useful. We examine them with reference to three aspects of prosodic encoding: prominence, duration, and intonation.

Prominence: A common feature of English is that some syllables are perceived as more prominent than others. The location of a prominence is determined in part lexically. In *ELephants* the first syllable is the most prominent; in *esCAPED*, the last. When these words occur in the same utterance, one of them will typically receive more prominence than the other, depending on such things as information focus or surprisingness of content (cf. Bolinger 1986; Tench 1996). For example, in response to "What happened today?" the reply might be "The *elephants* escaped," with the greater prominence on *elephants*, whereas in response to "Did you feed the elephants today?" the response might be "The elephants *escaped*."

All transcription systems mark unusual prominence (e.g. contrastive stress and "boosters"). Some systems mark many other prominences as well, such as primary stress (') and secondary stress (^) in the following example:

(7) **From Du Bois, et al. 1993: 58:**
G: . . . (2.2) 'a=nd of course,
a 'lot of herb ^tea,
when I'd 'rather be drinking ^whiskey.

These prominences are also marked in the London–Lund Corpus (Svartvik and Quirk 1980).

Perceived prominence presents considerable challenges to automatic detection by computer. It may arise from a marked change in pitch, or increased intensity, lengthening, or a combination of these and other factors. The same speaker may use different combinations of these cues in different types of discourse, or even within a single stretch of discourse (Brown et al. 1980).

Duration (lengthening and shortening): The length of syllables is determined to some degree lexically, as a function of which syllable is stressed in the word. For example, the second syllable is longer in *subJECT* than in *SUBject*. In addition, speech rate tends to speed up at the beginnings of phrases (**anacrusis**) and to slow down at the ends (**phrase-final lengthening**). Those discourse researchers who mark syllable lengthening or shortening tend to mark it only where it deviates from norms or is interactively significant for other reasons.

Intonation: There is no definitive "phoneme" in prosodic research, that is, no units which correlate with meaning in such a way that the principle of distinctive contrast

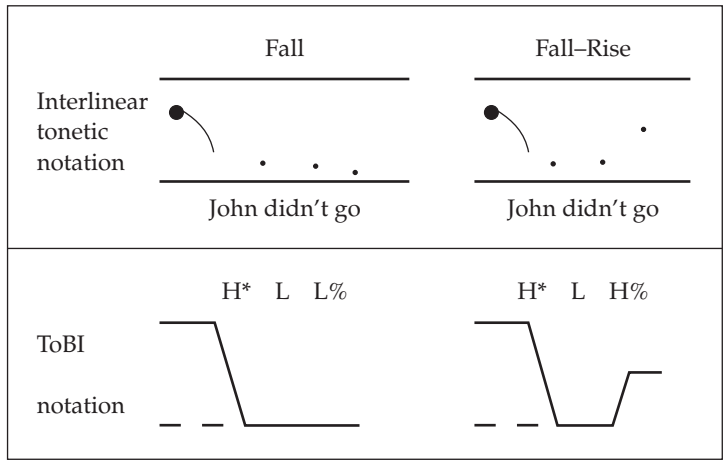

Figure 17.1 Interlinear tonetic and ToBI notations

can apply. However, we know that intonation is used systematically for communication – quite apart from the contributions of words and syntax. The analyst's task is to determine, as Crystal (1969) expressed it, what nonsegmental contrasts are meaningful, within a set of postulated systems.

Researchers differ in the size of unit they believe to be meaningful. Some have attributed meanings to entire contours (e.g. Bolinger 1986). Others have subdivided the contours and sought meaningful generalizations regarding those subparts (e.g. that a falling unit-final contour suggests conviction or closure). Tench (1996) compares a number of those proposals. Some researchers attribute meaning to the degree of rise (or fall) relative to the speaker's comfortable range (cf. "boosters", described in Altenberg 1990).

Systems also differ in their treatment of "declination," that is, the tendency for fundamental frequency to drift downward from the beginning to the end of an intonation unit. Acoustically oriented researchers may represent intonation contours as being superimposed upon a declining baseline (e.g. 't Hart et al. 1990). Others draw the contours as being superimposed on a level baseline – an approach which is less acoustic and more interpretive.

A widespread and intuitively accessable notation is that known as "interlinear tonetic notation," which is illustrated in the top pane of figure 17.1. In that format, "the top and bottom lines represent the top and bottom of the speaker's pitch range and each dot corresponds to a syllable, the larger dots indicating stressed and/or accented syllables" (Cruttenden 1997: xv).

An especially important difference between researchers is that between "levels" and "movements." Some researchers encode the relative height of individual syllables – e.g. Pike (1945); Ladd (1996); and the ToBI (Tones and Break Indices) prosodic conventions (Beckman and Ayers 1997). Others focus on pitch movements, or "nuclear tones" – e.g. the "British School" (Cruttenden 1997; Crystal 1980; Svartvik and Quirk 1980).

Notational systems of "movements" have included: (1) "nuclear tones," which extend from the main prominence in a stretch of speech to the end of that unit (e.g.

high-rising, low-falling, fall–rise); (2) pitch change on the final syllable of the unit (e.g. rising, falling, continuing); (3) larger patterns such as the rise–fall–rise pattern observed in the so-called "contradiction contour" (see Cruttenden 1997).

The focus on levels versus movements inspired considerable differences in their notation conventions. Figure 17.1 compares their notations with reference to two very common contours in American and British English. Within the British school, represented by Cruttenden (1997: 61), they are called "falling" and "fall–rise" nuclear tones, and are expressed in "interlinear tonetic notation." In ToBI notation, represented here by Pierrehumbert and Hirschberg (1990: 281), these receive less immediately transparent labels: H* L L% and H* L H%.

There is a variety of other differences between British school and ToBI notations (see Cruttenden 1997). For tables of partial correspondences between the two systems, see Ladd (1996: 83) and Roach (1994). Additional methods of prosodic transcription are surveyed by Ehlich and Switalla (1976), Gibbon (1976), and Leech et al. (1998).

An important branch of prosody research involves comparing observed acoustic properties (i.e. measured aspects of waveforms) with the auditory perceptions of listeners (i.e. psychological and linguistic categories). This work serves to clarify the acoustic substrates of listener perceptions of prosody (e.g. 't Hart et al. 1990). Text-to-speech conversion (e.g. Dutoit 1997; Svartvik 1990) is another area which promises to advance our knowledge of prosody. These approaches seek the most reliable correspondences between acoustic and auditory descriptions, often making explicit also contributions of other factors (i.e. syntax, semantic, pragmatic, and other information) in order to generate acceptably natural prosodic contours.

Because the prosodic distinctions are difficult to convey in writing alone, documentation should include recordings in addition to manuals. This was appreciated by Armstrong and Ward (1942), who made their examples available for purchase on three grammophone records (1942: vii). Audiorecordings are available for the following recent systems: Cruttenden (1997) (via audiocassette, Cambridge University Press ISBN: 0 521 62902 0), Du Bois et al. (1993) (via tutorial on computer diskette, available from the Linguistics Department, University of California at Santa Barbara), and Beckman and Ayers (1997) (via audiocassette and ftp-transferable digital records, available from Mary Beckman).

2.2.5 Rhythm and coordination

One key property not discussed so far is that of rhythm and coordination. Regardless of the degree to which nonverbal actions and utterances contribute independently to the content of an interaction, they are often unfolding jointly in time. Some researchers have attempted to systematize these observations and look into their temporal organization and coordination with one another. Erickson and Shultz (1982) and Ehlich (1993) did this by incorporating nonverbal actions as well as utterances into their "partiture" (or musical score) format, and by looking for synchrony within and across speakers as a reflection of communicative smoothness or difficulty in interview situations.

Scollon (1982) suggests that people contribute to an interaction in keeping with its rhythm at the time. This is an exciting area, not yet routinely indicated in transcripts. This may be due in part to a premature dismissal of stress timing (though see

Couper-Kuhlen 1993 and Dutoit 1997 for reasons why this was unjustified). It may also be due to a lack of tools facilitating this type of encoding (such as exist now for prosodic encoding). A practical and systematic approach to encoding it is found in Couper-Kuhlen (1993).

2.2.6 Turn-taking

Categories concerned with turn transition include unusually short pauses between one speaker and the next (**latching**), interruption by the second speaker, and simultaneous talk (**overlap**). These conventions are among those devised by Gail Jefferson for **conversation analysis** (Jefferson 1984; Psathas 1990) and used widely in any area of discourse research concerned with coordination of turns across speakers (e.g. Gumperz and Berenz 1993).

Transcription systems differ as to whether they mark only the beginnings of overlapping sections or also the ends (see Leech et al. 1998). They differ too in whether they mark overlap only by adjacent marks or also by indentation.

2.2.7 Nonverbal aspects and events

Nonverbal actions constitute a communicative resource which is partially independent of spoken language. As such, it raises many of the issues already discussed with reference to spoken language, such as how detailed the descriptive system can be without overburdening a viewer, and what is the best format for displaying the information.

Partiture format is the one most often used for capturing nonverbal events (e.g. Ehlich 1993; Neidle and MacLaughlin 1998). One approach (by Farnell 1995, cited in Duranti 1997: 149) involves the use of Laban notation, borrowed from choreography. The SignStream Project (discussed in the next section) provides a computer interface for handling complex data of this type as well as linking the transcript to the videorecording.

For capturing gaze patterns during conversations, Jefferson (1984) proposed inserting a line above or below an utterance to indicate the stretch of words during which mutual gaze occurred. A couple of additional conventions are used to indicate transition into and out of mutual gaze, to identify who is involved, and other details. This system has been found useful in several studies (e.g. Goodwin and Goodwin 1992).

3 Practicalities

Transcription is a notoriously time-consuming process. The exact amount of time required depends on the type of speech involved (e.g. amount of interruption, self-repair, or overlapping speech) and the amount of detail (e.g. word flow, prosodics, turn-taking). Estimates for word-level transcription with minimal added information are in the range of 10 or 11 minutes of transcribing for every 1 minute of speech (e.g. Crowdy 1995; Gibbon et al. 1997: 171). To produce a transcript containing the types of information encoded in most discourse transcripts (e.g. overlaps, pauses, stress or

prominence, notable changes in rate or loudness, etc.), the time estimates increase to 20:1 (Ervin-Tripp 2000). Coding can take considerably longer, depending on the number and complexity of the categories involved.

A great saving in time would be possible if off-the-shelf speech recognition software (such as that produced by Dragon Systems, Lernhout and Hauspie, or IBM) could be used to produce word-level discourse transcripts. When attempted, this has so far yielded only modest results (e.g. Coniam 1998). Current off-the-shelf speech recognizers need a training phase, are designed to recognize only one person's speech per session, and are hindered by noisy conditions. Perhaps it may be possible someday. Automatic speech recognition is progressing not only in word recognition (for an overview, see Bourlard and Morgan 1994), but also in automatic detection of speaker shift, topic shift, utterance boundaries, and stressed syllables. In the meantime, specialized software interfaces can greatly facilitate human efforts on both transcription and coding.

Transcriber (Barras et al. 1998) is a highly developed software tool, which gives the human transcriber virtually unlimited control over the playback of a digital recording, and provides a convenient computer interface for data entry. Transcriber is distributed as free software under GNU General Public License, at www.etca.fr/CTA/gip/Projects/Transcriber/. Its basic format is what was above called vertical format, that is, speaker turns are arranged in a single column down the page. It does not yet allow fine-grained encoding of overlapping speech, but it is in others ways surprisingly flexible, and seems likely to continue to develop.

For partiture or multitiered format including nonverbal information, relevant software includes HIAT-DOS (Ehlich 1993; http://www.daf.uni-muenchen. de/HIAT/HIAT.HTM) and SignStream (Neidle and MacLaughlin 1998; www.bu.edu/asllrp/SignStream/), the latter linking the transcript to a digitized videorecording.

Software supporting coding or annotation includes: the Corpus/Annotation Toolbox (Garside and Rayson 1997; McEnery and Rayson 1997), which is associated with Lancaster University's UCREL research center (www.comp.lancs.ac.uk/ucrel/); the SAM (Speech Assessment Method) software tools (Gibbon et al. 1997: appendix E), which is part of the EAGLES project (Expert Advisory Group on Language Engineering); and the MATE Coding Workbench, which is part of the Multilevel Annotation Tools Engineering project (described at mate.nis.sdu.dk/). Additional software is listed on the resources web page of the Linguistic Data Consortium (morph.ldc.upenn.edu/annotation/).

Apart from the practicalities of adding distinctions to a transcript, several guidelines may help ensure that the resulting encoding (transcription or coding) is as accurate and consistent as possible.

Lampert and Ervin-Tripp (1993) note that learning to use a coding system is similar to learning a second language: in both cases, the learner must learn how to extend what is already known in useful ways. To expedite learning, they recommend that coders be trained in groups using clear and concrete documentation, and that intercoder agreement be evaluated regularly, to guard against differential uses and "category drift." Intercoder agreement can be assessed with Cohen's Kappa. It is preferable to percent agreement because it corrects for chance agreement. If Kappa is low for a particular section of a code, it may indicate the need for additional coder training or better documentation, or it may signal the need for revising the categories themselves.

Additional suggestions for transcription are the following:

- Do not encode more than what is known to be needed in a project. Additional distinctions can always be added later, as needed.
- Limit the load on working memory by restricting the number of decisions made on any one pass through the data.
- Carefully document all distinctions in the form of a transcription manual or coding manual. Where possible the manual should contain not only a conceptual definition, but also both good examples of a category and boundary cases (Lampert and Ervin-Tripp 1993). For prosodic distinctions, audiorecordings are useful in documenting distinctions, and are being made available increasingly by their authors. Manuals serve not only in encoding but also when using data encoded by others.
- Documentation should also include information regarding the participants and their relationships, the social context and setting, the task (if any), and other factors. For discussion of typologies of discourse types, see Leech et al. (1998).

4 Transcription: Past and Future

Much as phoneticians believed, prior to Abercrombie, that phonetics began in 1830 (Fromkin and Ladefoged 1981), one sometimes finds claims in discourse literature such as the following: "transcription – the entextualization of speech – has only recently been taken seriously. [E]arlier research practices assumed . . . that what was said could be represented adequately in the form of paraphrases or summaries" (Mischler 1991). In fact, attempts to transcribe discourse can be traced to antiquity.

4.1 Ancient origins

The earliest attempts to capture spoken language in writing date back at least 22 centuries, to the golden age of oratory in ancient Greece. According to Parkes (1993), during the fourth century BCE, when the ideal of eloquence was central, writing was viewed as a record of the *spoken language*:

> Texts were mostly read aloud. A reader on his own might murmur the sounds of the words to himself, but the ideal was a kind of expressive declamation with well modulated pronunciation, in which the text was carefully phrased (*distincta*) by means of appropriate pauses. (Parkes 1993: 9)

Punctuation marks came to be added to early texts by teachers and educated readers in order to clarify structure and facilitate oral performance. There are parallels in the types of properties which were marked even if the specific distinctions were not the same. Some texts from the first and second centuries CE show a tripartite division of pauses: "minor medial pause" (a low dot), "major medial pause" (a dot at medium height), and "final pause" (a high dot). By the sixth century CE, punctuation marks had been employed to signal various aspects of interest in modern transcripts:

unit boundaries, pauses, rising and falling intonation on syllables, and some aspects of rhythm (Parkes 1993; Steele 1779). Some modern work even uses ancient distinctions (e.g. *ictus* and *remiss*, in Abercrombie's and in Halliday's work on rhythm, cited in Butler 1985: 139).

There are parallels in the typographical devices as well, dictated no doubt by perceptual and cognitive factors such as what is most easily noticed amidst words on a page, and most consistently produced with the methods and materials at hand. These fall into four categories.

Punctuation: The ad hoc marks devised by ancient readers (as today) tended to be simple geometric forms (wedges, curls, dots, dashes, hooks, etc.), placed on the line or elevated above it (similar to our distinction between comma and apostrophe), alone or in combination (similar to our semicolon and colon). In Parkes (1993), one encounters discussions concerning what, whether, and how much to punctuate. For example, Cicero argued that the reader should be guided by the constraint of rhythm, "not by the need for breath nor by copyists' marks," and used only few marks in his texts. In 1220, Bene of Florence "ridiculed the notion that punctuation should attempt to represent intonation or delivery: 'for if we wish to vary the written marks according to our manner of delivery it would look like an antiphonary' " (Parkes 1993: 45).

Metacomments: Prior to the development of quotation marks and question marks, scribes indicated these by means of linguistic markers, such as "dicit" for indirect speech, or "scriptum est" for quotations (Parkes 1993: 11). This is similar to the modern-day use of abbreviations such as "ac" meaning "accelerando," in Gumperz and Berenz's (1993) conventions, to indicate some property of a stretch of speech.

Visual prominence: In inscriptions from as early as 133 BCE the practice is already seen of placing the first letters of a new paragraph to the left of the line, enlarging it (*litterae notabiliores*) (Parkes 1993: 10). This is not unlike contemporary practices of printing stressed words in captial letters (e.g. Tedlock 1983; Bolinger 1986) or boldface (e.g. Goodwin and Goodwin 1992).

Spatial arrangement: Scribes in the second century BCE were already leaving blank space between major sections of a text (e.g. chapters or paragraphs (*per capitula*) (Parkes 1993: 10). In the fourth century CE, St. Jerome introduced a spatial innovation into the formatting of Bibles, for the purpose of clarifying the structure of ideas and avoiding misunderstandings of religious doctrine. This format, *per cola et commata*, involves starting each sense unit on a new line. "Where the unit of sense is too long to be accommodated in a single line, the remainder is inset on the next line and the insetting is continued until the end of the unit" (Parkes 1993: 16). In his prologue to his translation of Ezekiel, Jerome writes: "that which is written *per cola et commata* conveys more obvious sense to the readers" (Parkes 1993: 15). St. Jerome applied it only to Isaiah and Ezekiel; scribes later extended it to all the other books of the Bible. The Codex Amiatinus (from 716 CE), which was one of the earliest complete Bibles to survive, used these spacing conventions (from Parkes 1993: 179):

(8) SED IN LEGE DOMINI UOLUNTAS EIUS
ET IN LEGE EIUS MEDITABITUR
DIE AC NOCTE

The English translation is:

(9) But his will is in the law of the Lord
and in His law shall he meditate
day and night

Many centuries later, this format is found in virtually any transcript which is organized in terms of idea units, such as the following (from Schegloff 1982: 82):

(10) *B*: Uh now I could've walked, the three or
four blocks,
to that cab stand,

or Chafe's (1987) work on information flow, or Tedlock's (1983) work on Native American narratives. This use of space is especially pronounced in the Penn Treebank, where there are several more levels of indentation to signal additional subordination (see example 6).

After the sixth century CE, it became fashionable to read silently (Parkes 1993), and written conventions diverged progressively from spoken ones, took on properties of their own, and became less relevant to transcription.

4.2 The novel

During the eighteenth century, the novel arose as a new literary form, which attempted to capture conversations in a realistic way. Some conventions which arose in this medium and became adopted in some transcription conventions are the use of three dots (. . .) for pauses, or a dash (–) for interrupted thoughts or utterances (Du Bois 1991), interpretive metacomments (e.g. "he said cheerfully"), and modified spelling to capture variant phonology or dialects.

4.3 Scientific notations

As early as the 1500s, scholars attempted to encode spoken language for scientific study. Some of their practices have parallels in modern transcripts.

In 1775, Joshua Steele proposed a notation intended to capture the melody and rhythm of English "in writing" (1775: 15), in a manner similar to musical notation. With no access to modern devices for recording or measuring speech, Steele repeated a sentence over and over, while finding its notes on his bass viol, and expressed rhythms with reference to human gait or the beats of a pendulum.

The use of quasi-musical notation is found in modern approaches, though stylized in various ways. In interlinear tonetic transcription (e.g. Cruttenden 1997), a two-line staff is used to represent the top and bottom of the speaker's natural pitch range; Brown et al. (1980: 64) use a three-line staff. In Bolinger's work (e.g. Bolinger 1986), the words themselves flow up and down with no staffs and no lines.

Phoneticians had been devising shorthand systems for capturing speech sounds since the sixteenth century (MacMahon 1981), and Fromkin and Ladefoged (1981) speak of "the seventeenth-century search for a universal phonetic alphabet that could be distinguished from the separate alphabets required for particular languages" (1981: 3). In the 1800s, many different shorthand systems were being developed in England and continental Europe. Among the developers were Alexander Melville Bell (1867) and Henry Sweet (1892), whose work contributed to the development of the International Phonetic Alphabet (IPA).

In the mid-1800s one finds sporadic mention of markings for stress, voice quality including ingressive versus egressive whisper, and voice (Bell 1867: 48). Stress markings are found in all intonation-oriented transcripts; ingressive/egressive is part of Gail Jefferson's conventions (Jefferson 1984), which is the standard used in conversation analysis. Bell (1867) distinguished fall–rise and several other intonation types, which are widespread in contemporary approaches.

Despite these parallels with modern work, researchers attempting to systematically study any of these factors before the twentieth century faced serious technological limitations. In the 1930s, Boas relied on native speakers' ability to speak slowly and clearly as he wrote down their narratives (Duranti 1997: 122). This method worked less well for conversations, where his informants were less careful in their speech. Henry Sweet's shorthand methods reportedly enabled a recording rate of 150 words per minute (MacMahon 1981: 268), but here too there were no double checks on accuracy. Joshua Steele's methods, though resourceful, enabled at best a very crude approximation of melody and rhythm compared to what is possible with modern signal processing technology. Even as late as the 1920s, prosodic analyses were still often impressionistic, and conflicting claims were difficult to assess (Crystal 1969).

4.4 Technological advances

In the 1950s, recording technology became available for research use, making it possible to replay an interaction indefinitely many times, and to analyze timing and pronunciation to virtually unlimited degrees of refinement. Signal processing technology has made it possible to measure the physical properties of the speech signal, reflecting such things as fundamental frequency, energy, or other parameters. And computer interfaces make it possible to enter data effortlessly, to search quickly through even the largest databases, and to view transcripts and acoustic wave forms simultaneously, enabling an increasing interpenetration of visual and verbal information.

Technology does not provide answers on its own, however. For example, the measured aspects of a speech signal (such as wave frequency and amplitude) do not correspond perfectly with the perceived aspects (such as fundamental frequency and loudness). The ability to connect transcripts to acoustic measurements is an important step toward understanding those relationships, but it does not arise automatically with the technology of linking the two. The technology itself must be harnessed in various ways to be of benefit to research. The transcript draws attention to distinctions which people find meaningful and can make consistently. By providing categories which are relevant to human interactants, the transcript helps bridge the gap between technology and discourse understanding.

4.5 Convergence of interests

There is now an increasing convergence of interests between discourse research and computer science approaches regarding natural language corpora. This promises to benefit all concerned.

Language researchers are expanding their methods increasingly to benefit from computer technology. Computer scientists engaged in speech recognition, natural language understanding, and text-to-speech synthesis are increasingly applying statistical methods to large corpora of natural language. Both groups need corpora and transcripts in their work. Given how expensive it is to gather good-quality corpora and to prepare good-quality transcripts, it would make sense for them to share resources to the degree that their divergent purposes allow it.

Traditionally, there have been important differences in what gets encoded in transcripts prepared for computational research. For example, computational corpora have tended not to contain sentence stress or pause estimates, let alone speech act annotations. This is partly because of the practicalities of huge corpora being necessary, and stress coding not being possible by computer algorithm. If discourse researchers start to share the same corpora, however, these additional types of information may gradually be added, which in turn may facilitate computational goals, to the benefit of both approaches.

One indication of the increasing alliance between these groups is an actual pooling of corpora. The CSAE (Corpus of Spoken American English), produced by linguists at UC Santa Barbara (Chafe et al. 1991), was recently made available by the LDC (Linguistics Data Consortium), an organization previously distributing corpora guided by computational goals (e.g. people reading digits, or sentences designed to contain all phonemes and be semantically unpredictable).

4.6 Encoding standards

Another indication of the increasing alliance is the existence of several recent projects concerned with encoding standards relevant to discourse research, with collaboration from both language researchers and computational linguists: the TEI, the EAGLES, MATE, and the LDC. While time prevents elaborate discussion of these proposals, it is notable that all of them appear to have the same general goal, which is to provide coverage of needed areas of encoding without specifying too strongly what should be encoded. They all respect the theory-relevance of transcription.

McEnery and Wilson (1997: 27) write: "Current moves are aiming towards more formalized international standards for the encoding of any type of information that one would conceivably want to encode in machine-readable texts. The flagship of this current trend towards standards is the Text Encoding Initiative (TEI)." Begun in 1989 with funding from the three main organizations for humanities computing, the TEI was charged with establishing guidelines for encoding texts of various types, to facilitate data exchange. The markup language it used was SGML (Standard

Generalized Markup Language) – a language which was already widely used for text exchange in the publishing industry. Like HTML (the markup language of the worldwide web), SGML and XML (the eXtensible Markup Language) use paired start and end tags (e.g. <s> and </s>) to indicate the boundaries of format-relevant segments of text, which are then interpreted by an appropriate interface (or browser) so that the desired visual display is accomplished without the user needing to see the tags themselves. SGML and XML derive their power in addition from other structural aspects, which are discussed in detail elsewhere (Burnard 1995; Johansson 1995; Mylonas and Allen 1999).

The subcommittee on Spoken Language Encoding for the TEI (Johansson et al. 1992) began its work with a large-scale survey of transcription methods, in order to identify as comprehensively as possible all major dimensions encoded in transcripts. Then the subcommittee proposed how each dimension should be encoded in TEI-compliant format. Rather than dictating what users should encode in a transcript, the TEI approach was to catalog distinctions used by others and to establish TEI-compliant ways of encoding them if the researcher wishes to include them in a transcript. TEI standards were designed to facilitate data exchange across projects, to enable the same transcript to be flexibly formatted in different ways for different research purposes, and to support technological upgrading from text-only to text aligned with digital records.

TEI conventions have been adopted in the 100,000,000-word British National Corpus (Crowdy 1995), and in the 1,000,000-word International Corpus of English (McEnery and Wilson 1997: 29). The TEI is now a consortium, and is shifting from SGML to XML. For more information see Burnard (1995), Johansson (1995), Mylonas and Allen (1999), and the website, http://www.tei-c.org/.

Within the EAGLES project, three developments are relevant here. The first is the impressive survey of LE (language engineering) transcription and annotation methods prepared by Leech et al. (1998). The second is the *Handbook of Standards and Resources for Spoken Language Systems* (Gibbon et al. 1997), which offers precise, extensive, and useful technical guidelines on such things as recording equipment and computer-compatible IPA, based on European projects. The third is the XCES (XML Corpus Encoding Standard) developed for EAGLES by Nancy Ide and Patrice Bonhomme at Vassar College (www.cs.vassar.edu/XCES).

Leech et al.'s (1998) survey concerns encoding conventions which can be of use in large corpora for applied purposes involving training and the use of computer algorithms. Leech et al. feel it is premature to favor a particular standard at this point and instead favor the general approach taken in the TEI: "For future projects, we recommend that as much use as possible should be made of standardized encoding schemes such as those of the TEI, extending them or departing from them only where necessary for specific purposes" (1998: 14).

In a domain which is as theory-relevant as transcription, TEI (or something like it) is really the only workable basis for standardization. A standard of this type seeks to provide a mechanism (via markup conventions) for systematic encoding of data, such that the data can be flexibly reformatted later in various ways as dictated by the purpose at hand, but leaves it to the discretion of individual researchers to decide what exactly to encode and what categories to use.

5 General Discussion and Conclusion

The present chapter has provided an overview of factors which are relevant whenever transcripts are used. The transcript is an invaluable asset in discourse analyses, but it is never theory-neutral. Awareness of alternatives and their biases is an important part of their use. It is hoped that this chapter contributes to effective use of transcripts and to the continued development of discourse methodology more generally.

REFERENCES

Altenberg, B. 1990. Some functions of the booster. In *The London–Lund Corpus of Spoken English: Description and Research*, ed. Jan Svartvik. 193–209. Lund: Lund University Press.

Armstrong, Lilias E., and Ward, Ida C. 1942. *Handbook of English Intonation*. Cambridge: W. Heffer and Sons.

Barras, Claude, Geoffrois, Edouard, Wu, Zhibiao, and Liberman, Mark. 1998. Transcriber: a free tool for segmenting, labeling and transcribing speech. *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, 1373–6. May 1998. (Available at www.etca.fr/CTA/gip/publis.html.)

Beckman, Mary E., and Ayers, Gayle Elam. 1997. *Guidelines for ToBI Labelling, Version 3*. Columbus, OH: Ohio State University Research Foundation.

Bell, Alexander Melville. 1867. *Visible Speech: The Science of Universal Alphabetics, or Self-interpreting Physiological Letters, for the Writing of All Languages in one Alphabet*. London: Simpkin, Marshall.

Bloom, Lois. 1973. *One Word at a Time: The Use of Single Word Utterances Before Syntax*. The Hague, Mouton.

Bloom, Lois. 1993. Transcription and coding for child language research: the parts are more than the whole. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 149–66. Hillsdale, NJ: Lawrence Erlbaum.

Bolinger, Dwight. 1986. *Intonation and its Parts: Melody in Spoken English*. Stanford: Stanford University Press.

Bourlard, Herve, and Morgan, Nelson. 1994. *Connectionist Speech Recognition: A Hybrid Approach*. Boston: Kluwer Academic.

Brown, Gillian, Currie, Karen L., and Kenworthy, Joanne. 1980. *Questions of Intonation*. London: Croom Helm.

Burnard, Lou. 1995. The Text Encoding Initiative: an overview. In *Spoken English on Computer: Transcript, Mark-up and Application*, eds Geoffrey Leech, Greg Myers, and Jenny Thomas. 69–81. New York: Longman.

Butler, Christopher. 1985. *Systemic Linguistics: Theory and Applications*. London: Batsford Academic and Educational.

Chafe, Wallace L. 1987. Cognitive constraints on information flow. In *Coherence and Grounding in Discourse*, ed. Russell S. Tomlin. 21–51. Philadelphia: John Benjamins.

Chafe, Wallace L. 1993. Prosodic and functional units of language. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 33–43. Hillsdale, NJ: Lawrence Erlbaum.

Chafe, Wallace L., Du Bois, John W., and Thompson, Sandra A. 1991. Towards a new corpus of spoken American English. In *English Corpus Linguistics*, eds Karin Aijmer and Bengt Altenberg. 64–82. New York: Longman.

Coniam, David. 1998. Speech recognition: accuracy in the speech-to-text process. *TEXT Technology*, 8.1–13.

Couper-Kuhlen, Elizabeth. 1986. *An Introduction to English Prosody*. London: Edward Arnold.

Couper-Kuhlen, Elizabeth. 1993. *English Speech Rhythm: Form and Function in Everyday Verbal Interaction*. Philadelphia: John Benjamins.

Crowdy, Steve. 1995. The BNC Corpus. In *Spoken English on Computer: Transcript, Mark-up and Application*, eds Geoffrey Leech, Greg Myers, and Jenny Thomas. 224–34. New York: Longman.

Cruttenden, Alan. 1997. *Intonation*. 2nd edn. Cambridge: Cambridge University Press.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. New York: Cambridge University Press.

Crystal, D. 1975. *The English Tone of Voice: Essays in Intonation, Prosody and Paralanguage*. London: Edward Arnold.

Crystal, D. 1980. The analysis of nuclear tones. In *The Melody of Language*, eds Linda R. Waugh and C. H. van Schooneveld. 55–70. Baltimore: University Park Press.

Du Bois, John W. 1991. Transcription design principles for spoken language research. *Pragmatics*, 1.71–106.

Du Bois, John W., and Schuetze-Coburn, Stephan. 1993. Representing hierarchy: constitute hierarchy for discourse databases. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 221–60. Hillsdale, NJ: Lawrence Erlbaum.

Du Bois, John W., Schuetze-Coburn, Stephan, Cumming, Susanna, and Paolino, Danae. 1993. Outline of discourse transcription. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 45–89. Hillsdale, NJ: Lawrence Erlbaum.

Duranti, Alessandro. 1997. *Linguistic Anthropology*. New York: Cambridge University Press.

Dutoit, Thierry. 1997. *An Introduction to Text-to-speech Synthesis*. Boston: Kluwer Academic.

Edwards, Jane A. 1989. *Transcription and the New Functionalism: A Counterproposal to CHILDES' CHAT Conventions*. Technical Report, No. 60. Berkeley: UC Berkeley, Cognitive Science Program.

Edwards, Jane A. 1992a. Computer methods in child language research: Four principles for the use of archived data. *Journal of Child Language*, 19.435–58.

Edwards, Jane A. 1992b. Design principles for the transcription of spoken discourse. In *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82, Stockholm, August 4–8, 1991*, ed. Jan Svartvik. 129–47. New York: Mouton de Gruyter.

Edwards, Jane A. 1992c. Transcription in discourse. In *Oxford International Encyclopedia of Linguistics, vol. 1*, ed. William Bright. 367–71. Oxford: Oxford University Press.

Edwards, Jane A. 1993a. Perfecting research techniques in an imperfect world. *Journal of Child Language*, 20.209–16.

Edwards, Jane A. 1993b. Principles and contrasting systems of discourse transcription. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 3–31. Hillsdale, NJ: Lawrence Erlbaum.

Edwards, Jane A. 1995. Principles and alternative systems in the transcription, coding and markup of spoken discourse. In *Spoken English on Computer: Transcript, Mark-up and Application*, eds Geoffrey Leech, Greg Myers, and Jenny Thomas. 19–34. New York: Longman.

Ehlich, Konrad. 1993. HIAT: a transcription system for discourse data. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 123–48. Hillsdale, NJ: Lawrence Erlbaum.

Ehlich, Konrad, and Switalla, B. 1976. Transkriptionssysteme – Eine exemplarische Uebersicht. *Studium Linguistik*, 2.78–105.

Erickson, Frederick, and Shultz, Jeffrey. 1982. *The Counselor as Gatekeeper: Social Interaction in Interviews*. New York: Academic Press.

Ervin-Tripp, Susan. 2000. Methods for studying language production. In *Studying Conversation: How to Get Natural Peer Interaction*, eds Lise Menn and Nan Bernstein Ratner. 271–90. Mahwah: Lawrence Erlbaum.

Farnell, Brenda. 1995. *Do You See What I Mean?: Plains Indian Sign Talk and the Embodiment of Action*. Austin: University of Texas Press.

Firth, J. R. 1957. *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press.

Fromkin, Victoria A., and Ladefoged, P. 1981. Early views of distinctive features. In *Towards a History of Phonetics*, eds R. E. Asher and Eugenie J. A. Henderson. 3–8. Edinburgh: Edinburgh University Press.

Garside, Roger, and Rayson, P. 1997. Higher-level annotation tools. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, eds Roger Garside, Geoffrey Leech, and A. M. McEnery. 179–93. London: Longman.

Gibbon, Dafydd. 1976. *Perspectives of Intonation Analysis*. Bern: Herbert Lang.

Gibbon, Dafydd, Moore, Roger, and Winski, Richard. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. New York: Mouton de Gruyter.

Goodwin, C., and Goodwin, M. H. 1992. Assessments and the construction of context. In *Rethinking Context: Language as an Interactive Phenomenon*, eds Allessandro Duranti and Charles Goodwin. 147–89. Cambridge: Cambridge University Press.

Gumperz, John J., and Berenz, Norine B. 1993. Transcribing conversational exchanges. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 91–121. Hillsdale, NJ: Lawrence Erlbaum.

't Hart, Johan, Collier, Rene, and Cohen, Antonie. 1990. *A Perceptual Study of Intonation: An Experimental–Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Jefferson, Gail. 1984. Transcript notation. In *Structures of Social Action: Studies in Conversation Analysis*, eds J. Maxwell Atkinson and John Heritage. ix–xvi. Cambridge: Cambridge University Press.

Johansson, Stig. 1991. *Some Thoughts on the Encoding of Spoken Texts in Machine-readable Form*. Oslo: University of Oslo, Department of English.

Johansson, Stig. 1995. The approach of the Text Encoding Initiative to the encoding of spoken discourse. In *Spoken English on Computer: Transcript, Mark-up and Application*, eds Geoffrey Leech, Greg Myers, and Jenny Thomas. 82–98. New York: Longman.

Johansson, Stig, Burnard, Lou, Edwards, Jane, and Rosta, Andrew. 1992. *Chapter P2.34: Text Encoding Initiative.*

Spoken Text Working Group, Final Report. TEI, Department of English, University of Oslo.

Knowles, Gerry. 1995. Converting a corpus into a relational database: SEC becomes MARSEC. In *Spoken English on Computer: Transcript, Mark-up and Application*, eds Geoffrey Leech, Greg Myers, and Jenny Thomas. 208–19. New York: Longman.

Ladd, D. Robert. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

Lampert, Martin D., and Ervin-Tripp, Susan M. 1993. Structured coding for the study of language and interaction. In *Talking Data: Transcription and Coding in Discourse Research*, eds Jane A. Edwards and Martin D. Lampert. 169–206. Hillsdale, NJ: Lawrence Erlbaum.

Leech, Geoffrey, Weisser, Martin, Wilson, Andrew, and Grice, Martine. 1998. Survey and guidelines for the representation and annotation of dialogue. LE-EAGLES-WP4-4. Working paper for the EAGLES Project (http://www.ling.lancs.ac.uk/eagles/).

MacMahon, W. K. C. 1981. Henry Sweet's system of shorthand. In *Towards a History of Phonetics*, eds R. E. Asher and Eugenie J. A. Henderson. 265–81. Edinburgh: Edinburgh University Press.

Marcus, Mitchell P., Santorini, Beatrice, and Marcinkiewicz, Mary Ann. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19.313–30.

MacWhinney, Brian. 1995. *The CHILDES project: Tools for Analyzing Talk*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum.

McEnery, A. M., and Rayson, P. 1997. A Corpus/annotation toolbox. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, eds Roger Garside, Geoffrey Leech, and A. M. McEnery. 194–208. London: Longman.

McEnery, A. M., and Wilson, Andrew. 1997. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mishler, Elliot G. 1991. Representing discourse: the rhetoric of transcription. *Journal of Narrative and Life History*, 1.255–80.

Mylonas, Elli, and Renear, Allen (eds). 1999. The Text Encoding Initiative at 10: not just an interchange format anymore – but a new research community. *Computers and the Humanities*, 33, Issues 1 and 2.

Neidle, Carol, and MacLaughlin, D. 1998. SignStream[tm]: a tool for linguistic research on signed languages. *Sign Language and Linguistics*, 1.111–14.

Ochs, Elinor. 1979. Transcription as theory. In *Developmental Pragmatics*, eds Elinor Ochs and Bambi B. Schieffelin. 43–72. New York: Academic Press.

Parkes, M. B. 1993. *Pause and Effect: An Introduction to the History of Punctuation in the West*. Berkeley: University of California Press.

Pierrehumbert, Janet, and Hirschberg, Julia. 1990. The meaning of intonational contours in the interpretation of discourse. In *Intention in Communication*, eds Philip R. Cohen, Jerry Morgan, and Martha E. Pollack. 271–311. Cambridge, MA: MIT Press.

Pike, Kenneth L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Preston, Dennis R. 1985. The Li'l Abner syndrome: written representations of speech. *American Speech*, 60.328–36.

Psathas, George. 1990. Appendix: transcription symbols. In *Interaction Competence*, ed. George Psathas. 297–307. Washington, DC: University Press of America.

Roach, Peter. 1994. Conversion between prosodic transcription systems – Standard British and ToBI. *Speech Communication*, 15.91–9.

Roach, Peter, and Arnfield, Simon. 1995. Linking prosodic transcription to the time dimension. In *Spoken English on Computer: Transcript, Mark-up and Application*, eds Geoffrey Leech, Greg Myers, and Jenny Thomas. 149–60. New York: Longman.

Schegloff, Emanuel A. 1982. Discourse as an interactional achievement: some uses of "uh huh" and other things that come between sentences. In *Analyzing Discourse: Text and Talk*, ed. Deborah Tannen. 71–93. Washington, DC: Georgetown University Press.

Scollon, Ron. 1982. The rhythmic integrity of ordinary talk. In *Analyzing Discourse: Text and Talk*, ed. Deborah Tannen. 335–49. Washington, DC: Georgetown University Press.

Slobin, Dan I. 1967. *A Field Manual for Cross-cultural Study of the Acquisition of Communicative Competence*. Berkeley: University of California, Department of Psychology.

Steele, Joshua. 1775. *An Essay Towards Establishing the Melody and Measure of Speech to be Expressed and Perpetuated by Peculiar Symbols*. London: J. Almon.

Steele, Joshua. 1779. *Prosodia Rationalis: or, An Essay Towards Establishing the Melody and Measure of Speech, to be Expressed and Perpetuated by Peculiar Symbols*. London: J. Nichols.

Svartvik, J. (ed.). 1990. *The London–Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.

Svartvik, J., and Quirk, Randolph (eds). 1980. *A Corpus of English Conversation*. Lund: C. W. K. Gleerup.

Sweet, Henry. 1892. *A Manual of Current Shorthand, Orthographic, and Phonetic*. Oxford: Clarendon Press.

Tannen, Deborah. 1981. [Review of W. Labov and D. Fanshel, *Therapeutic Discourse: Psychotherapy as Conversation*]. *Language*, 57.481–6.

Tannen, Deborah. 1984. *Conversational Style*. Norwood, NJ: Ablex.

Tedlock, Dennis. 1983. *The Spoken Word and the Work of Interpretation*. Philadelphia: University of Pennsylvania Press.

Tench, Paul. 1996. *The Intonation Systems of English*. New York: Cassell.

Thompson, Henry S., Anderson, Anne H., and Bader, Miles. 1995. Publishing a spoken and written corpus on CD-ROM: the HCRC Map Task experience. *Spoken English on Computer: Transcript, Mark-up and Application*, eds Geoffrey Leech, Greg Myers, and Jenny Thomas. 168–80. New York: Longman.