LE-EAGLES–WP4-4

Integrated Resources Working Group

Survey and guidelines for the representation and annotation of dialogue

Geoffrey Leech, Martin Weisser, Andrew Wilson and Martine Grice

16 Oct 98

Authorial team:

Geoffrey Leech, UCREL, Lancaster University, UK
Martin Weisser, UCREL, Lancaster University, UK
Andrew Wilson, English Language and Linguistics, Chemnitz University of Technology, Germany
Martine Grice, Fachrichtung 8.7 Phonetik, Universität des Saarlandes, Saarbrücken, Germany

EAGLES Coordinators:

Dafydd Gibbon, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany
Jock McNaught, Centre for Computational Linguistics, UMIST, Manchester, UK

Contributing members of the Integrated Resources Working Group:

Kerstin Fischer, Computer Science Department, Universität Hamburg, Germany
Susanne Jekat, Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany
Elisabeth Maier, SBC / IT Camp, Basel, Switzerland
Paul Mc Kevitt, Center for PersonKommunikation, Aalborg University, Denmark
Jean Carletta, HCRC, University of Edinburgh, UK
Joaquim Llisterri, Universitat Autonoma de Barcelona, Spain

We also gratefully acknowledge help from the following:

Anton Batliner, Niels Ole Bernsen, František Cermak, Alain Couillault, Paul Dalsgaard, Mika Enomoto, Ulrich Heid, Arne Johnsen, Magne Johnsen, Andreas Kellner, Klaus Kohler, David Milward, Roger Moore, Norbert Reitlinger, Paul Rogers, and Fernando Sánchez-León.

Contents

1 Introduction **6**

1.1 This chapter 6

1.2 The subject of this chapter: What is meant by 'Integrated Resources'? 7

1.3 Limitations of the task undertaken in this chapter 7

2 A preliminary classification of dialogue corpora **9**

2.1 Dialogue acts 10

2.2 Towards a dialogue typology 10

3 Levels of representation or annotation **14**

3.1 General coding issues 14

3.2 Orthographic transcription 14

3.2.1 Background 15

3.2.2 Documentation on texts 15

3.2.3 Basic text units 16

3.2.4 Reference system 17

3.2.5 Speaker attribution 17

3.2.6 Speaker overlap 18

3.2.7 Word form 19

3.2.7.1 Word partials. 20

3.2.7.2 Orthography, including punctuation. 21

3.2.7.3 Unintelligible speech. 21

3.2.7.4 Uncertain transcription. 21

3.2.7.5 Substitutions. 21

3.2.8 Speech management 22

3.2.8.1 Pauses. 22

3.2.8.2 Quasi-lexical vocalizations. 22

3.2.8.3 Other phenomena. 22

3.2.9 Paralinguistic features 23

3.2.10 Non-verbal sounds 23

3.2.11 Kinesic features 23

3.2.12 Situational features 24

3.2.13 Editorial comment 24

3.2.13.1 Alternative transcriptions. 24

3.2.13.2 General comments. 25

3.2.14 Recommendations 25

3.2.14.1 Highest priority 25

3.2.14.2 Recommended . . . 26
3.2.14.3 Optional . . . 26
3.3 Morphosyntactic annotation . . . 26
3.3.1 Dysfluency phenomena in morphosyntactic annotation . . . 27
3.3.2 Word-classes which are characteristic of speech, but not of writing . . . 28
3.3.2.1 Interjections in morphosyntactic annotation . . . 28
3.3.2.2 Adverbs in morphosyntactic annotation . . . 29
3.3.3 Extending the part-of-speech categories in EAGLES morphosyntactic guidelines 30
3.3.4 Residual problems . . . 30
3.3.5 An alternative solution . . . 31
3.3.6 Recommendations . . . 31
3.4 Syntactic annotation . . . 31
3.4.1 Dysfluency phenomena in syntactic annotation . . . 32
3.4.1.1 Use of hesitators or 'filled pauses' . . . 33
3.4.1.2 Syntactic incompleteness . . . 33
3.4.1.3 Retrace-and-repair sequences . . . 34
3.4.1.4 Dysfluent repetition . . . 34
3.4.1.5 Syntactic blends (or anacolutha): . . . 34
3.4.1.6 Concluding remarks on syntactic annotation and dysfluency . . . 35
3.4.2 Unintelligible speech . . . 35
3.4.3 Segmentation difficulties . . . 36
3.4.4 Recommendations . . . 37
3.5 Prosodic annotation . . . 37
3.5.1 ToBI . . . 38
3.5.1.1 ToBI Tones . . . 39
3.5.1.2 ToBI Break Indices . . . 40
3.5.1.3 Using the ToBI system . . . 40
3.5.1.4 ToBI for other languages and dialects . . . 41
3.5.2 TSM - Tonetic stress marks . . . 43
3.5.3 Conversion between ToBI and the TSM system . . . 44
3.5.4 INTSINT . . . 45
3.5.5 Automatic annotation of prosody in VERBMOBIL . . . 46
3.5.6 Prosodic studies on the ATIS project . . . 47
3.5.7 X-SAMPA and SAMPROSA . . . 48
3.5.8 Recommendations . . . 49
3.6 Pragmatic annotation: functional dialogue annotation . . . 49
3.6.1 'Historical' background . . . 49

3.6.2 Methods of analysis and annotation . . . 50
3.6.3 Segmentation of dialogues . . . 51
3.6.4 Functional annotation of dialogues . . . 52
3.6.5 Utterance tags . . . 52
3.6.5.1 Communicative status . . . 52
3.6.5.2 Information level and status . . . 52
3.6.5.3 Forward-looking communicative function . . . 53
3.6.5.4 Backward-looking communicative function . . . 53
3.6.5.5 General remarks on the above categories . . . 54
3.6.6 Levels of functional annotation . . . 54
3.6.6.1 Micro-level annotation . . . 54
3.6.6.2 Meso-level annotation . . . 55
3.6.6.3 Macro-level annotation . . . 56
3.6.7 Techniques for identifying dialogue acts or topics . . . 56
3.6.7.1 Techniques for identifying dialogue acts . . . 56
3.6.7.2 Techniques for identifying topics . . . 57
3.6.8 Evaluation of Coding Schemes . . . 58
3.6.9 Annotation tools and general coding recommendations . . . 59
3.6.10 Recommendations . . . 60

A TEI paralinguistic features 68

B TEI P3 DTD: base tag set for transcribed speech 69

C A few relevant web links 71

1 Introduction

1.1 This chapter

The main purposes of this chapter are to present a survey of current and developing work in the areas of research covered by WP4 (Work Package 4: Integrated Spoken and Written Language Resources), and to provide preliminary guidelines for the representation or annotation of dialogue in resources for language engineering (see also Gibbon et al., 1998: 146-172).

The terms *representation* and *annotation* have distinct conventional uses in this chapter. 'Representation indexrepresentation' is used for the orthographic transcription of a dialogue, giving the basic information about what was said, by whom it was said, and other necessary details. The term 'annotation', on the other hand, is used for the additional levels of linguistic information which are added to the orthographic transcription. This conventional usage needs some brief preliminary explanation.

In reference to corpora of written language, the distinction is relatively clear: the *representation* of a text is the encoding of the orthographic form of the text itself, either as straight *ASCII* text, or in some mark-up system such as is provided by the TEI (Text Encoding Initiative: see Sperberg-McQueen and Burnard, 1994). On the other hand, *annotation* constitutes additions to that basic representation, providing various levels of linguistic analysis (such as morphosyntactic, syntactic, semantic levels: see Garside et al., 1997: 1-19). However, with a corpus of spoken language, the orthographic transcription does not have the same status of basic representation of the data, being itself a level of linguistic abstraction from the speech signal. (The term *transcription* above corresponds to *representation* in the sense that an orthographic transcription, say, undertakes to represent, as a verbatim record, what was said by the speakers in a dialogue.)

Traditionally, users of the transcription have treated it as a useful substitute for the actual sound recording, in deriving from it the wording and sense of the spoken message. It is clear, however, that this substitute use is not a desirable use of an orthographic transcription in spoken language resources for language engineering (LE). From the point of view of speech analysis, an orthographic transcription is more remarkable for what it excludes than for what it includes. Moreover, it is assumed, with modern technological progress, that all users of a spoken language corpus will have ready access to the sound recording, which can therefore be regarded as the basic record of any spoken language data.

Although this means that the orthographic transcription loses its observational primacy, there is still an important sense in which the orthographic transcription is the primary level of abstraction from the data, involving as little interpretation as possible. A common format for orthographic *representation* of dialogue is therefore highly desirable for the exchange (and automatic processing) of the data. Other levels of information, termed *annotations*, are added to this baseline verbatim record, without which it would be difficult to make sense of them.

This draft chapter contributes to the overall goals of WP4, which are to:

1. Identify and describe linguistic phenomena specific to spoken language and in particular to dialogue, which require special provision for annotation.

2. Survey, compare and analyse methods, solutions and practices proposed to represent and annotate these phenomena.

3. Propose guidelines for annotating the identified dialogue-specific phenomena at various levels.

4. Integrate these recommendations or guidelines, in a coherent way, into the overall annotation guidelines.

The present chapter addresses itself to the second and third of these goals, while not overlooking the other goals where relevant.

1.2 The subject of this chapter: What is meant by 'Integrated Resources'?

In the 1980s, the *speech community* and the *natural language* community' were effectively two research communities working on a common subject matter – human language – but otherwise having little communication with one another. In the 1990s this situation has changed, simply because many of the applications of language engineering (LE) involve both the domains of 'speech' and of 'natural language'. In 1998 it is more evident than ever before that these communities have to pool their specialist knowledge and to strive to become a single research community (see Llisterri 1996, Section 2.2 on the need for such convergence).

The Natural Language community has in the past concentrated (a) on written language processing, and/or (b) on the processing of language at higher levels of analysis (e.g. the syntactic and lexical levels) which apply both to written and spoken language, and where the distinction between the two channels is relatively unimportant. The speech community, on the other hand, has in the past tended to concentrate on 'lower' levels of analysis which relate fairly directly to the spoken signal.

However, it has already become clear that this division of interest can no longer be maintained: many of the most forward-looking and challenging applications of LE today (e.g. high-quality speech synthesis, large-vocabulary speech recognition, speech-to-speech translation, dialogue systems) involve both low-level and high-level processing. A parser, for example, is needed for processing both spoken and written language data. Moreover, current R&D (research and development) is working towards integrated spoken language systems undertaking all levels of speech understanding and speech synthesis, such as are needed for the appropriate understanding and production of speech in dialogue.

1.3 Limitations of the task undertaken in this chapter

Hence *integrated resources for spoken and written language* refers to LE resources which are to be shared by both speech and natural language processing research. They include *corpora*, *lexicons, grammars* and *tools*. For example, lexicons for integrated resources should provide for the integration of lexical information as a common resource relating to both spoken and written language (while allowing for their expedient separation where the need arises). There is also need for integration in a further sense: resources such as lexicons and corpora should be consistent with one another so that information can be easily exchanged between them. Similarly, tools should be capable of processing data in terms of the representations used for other resources. What can be achieved within the scope of this chapter, however, is limited in several ways.

1: In this chapter we restrict our attention primarily to (a) *corpora*, because this is the area in which the need for standardization arises most compellingly. We have not been able to consider (b) lexicons, (c) grammars and (d) tools in any detail. On the other hand, (d) *tools* have been given some attention here (see especially 3.6.9), since the transcription and annotation of spoken corpora are in part constrained by what tools exist or can be developed to facilitate and integrate these tasks.

A **corpus**in this context is a body of spoken language data which has been recorded, has been transcribed (in part or in toto) and documented for use in the development of LE systems, and in principle at least, is available for use by more than one research team in the community. The needs for *standards*, or rather *guidelines*, for the representation and annotation of spoken language data arises primarily because of the need to ensure interchangeability of data, between different sites, in a multilinguistic community such as the EU, so that progress in the provision of resources can be shared and can provide a springboard for further collaboration and advances in the future.

2: Apart from the focus on corpora, there is an additional restriction on the scope of this chapter, which is the decision to limit the treatment of integrated resources to **dialogue** corpora. For the present purposes we define a dialogue as a discourse in which two or more participants interact communicatively, and where at least one of the participants is human. This covers cases of human-machine as well as human-human dialogue. In principle, this can include not only spoken dialogue, but also written dialogue, where (for example) a human participant interacts with a machine via a keyboard. However, in practice, this chapter will mainly focus on spoken dialogue.

The focus on dialogue is timely, in view of the recent emergence of dialogue as an area ripe for rapid development, and the consequent demand for dialogue corpora. In the words of Walker and Moore (1997: 1):

> In the past, research in this area focused on specifying the mechanisms underlying particular discourse phenomena; the models proposed were often motivated by a few constructed examples ... Recently however the field has turned to issues of robustness and the coverage of theories ... this new empirical focus is supported by several recent advances: an increasing theoretical consensus on discourse models; a large amount of on-line dialogue and textual corpora available; and improvements in component technologies and tools for building and testing discourse and dialogue testbeds. This means that it is now possible to determine how representative particular discourse phenomena are, how frequently they occur, whether they are related to other phenomena, what percentage of the cases a particular model covers, the inherent difficulty of the problem, and how well an algorithm for processing or generating the phenomena should perform to be considered a good model.

Research in this field can be either close to or distant from practical commercial or industrial applications. Less applications-oriented studies may concentrate on certain modules or levels of analysis to the exclusion of others. All such studies can, however, be valuable in leading to richer and more precise models of human dialogue behaviour. What is particular significant, in task-oriented dialogue annotation, is that all levels of analysis can be seen as culminating in the pragmatic level, where the communicative function of the dialogue is characterised in terms of dialogue acts. Dialogue, in this perspective, is the nexus which gathers all areas of integrated resources research and development into a practical focus.

3: A third understandable limitation on our study of integrated resources is that we focus attention primarily on applications-oriented task-driven dialogue, bearing in mind that the objective of EAGLES is to promote the setting of standards in LE, rather than more generally in linguistics or social science, in such fields as dialectology, sociolinguistics, discourse analysis or conversational analysis. In recent years, corpora of spoken dialogue have been compiled for a wide variety of reasons. For example, one well-developed initiative is the CHILDES database (MacWhinney, 1991) which sets standards for the interchange of data between researchers in the area of child language acquisition. Another instance of incipient standardization is the spoken subcorpus of the BNC (British National Corpus) (see Burnard, 1995), which contains ca. 10 million words of spoken English, all transcribed and marked up in accordance with the guidelines of the TEI (Text Encoding Initiative) — see Johansson (1995). The need for a standard in this case had to be reconciled with the requirement of a corpus large enough to be usable for dictionary compilation and other wide-ranging fields of linguistic research. Other examples could be added: there can be many reasons for introducing standards/guidelines for representation of dialogue, apart from those which are most salient to the LE community. While it is instructive to take note of these other initiatives, especially where they come to conclusions of value to LE specialists, they should not be treated unquestioningly as a model to be followed in this chapter.

4: A final limitation of this task is the following. We have restricted attention to certain levels or tiers of representation/annotation where there is felt to be a particular need to propose guidelines. The levels for which a representation or annotation of dialogue can be provided are many: see Gibbon et al. (1998: 149 ff) for a reasonably complete list. However, for the present purpose we disregard semantic annotation (which is the concern of another Work Package, WP1), and we also largely ignore phonetic/phonemic and physical levels of transcription, on which considerable standardizing work has been done already (see, for example, Gibbon et al. 1998, 688-731 on SAMPA). We confine our attention to the following levels:

- **general** (Section 3.1) — general coding issues (i.e., SGML, XML, etc.)
- **orthographic** (Section 3.2) — constructing a verbatim record of the dialogue
- **morphosyntactic** (Section 3.3) — part-of-speech or word-class tagging
- **syntactic** (Section 3.4) — treebanks (either partially or fully parsed)
- **prosodic** (Section 3.5) — representation of suprasegmental phenomena such as accentuation and phrasing, using annotation systems such as ToBI, TSM or INTSINT and automatic analysis of acoustic parameters (e.g. fundamental frequency)
- **pragmatic** (Section 3.6) — functional units at macro-, meso- or speech acts levels in dialogue

At the same time, we assume that all the different levels of annotation above need to be integrated in a multi-layer structure, and linked through time alignment to the sound recording.

It has to be admitted that these levels (particularly the orthographic, pragmatic and prosodic) do not yet show a highly developed trend towards standardization. Consequently, this chapter concentrates heavily on surveying current practices, and on identifying those which may be considered good models for others to follow. Inevitably, we will have overlooked some significant current research, and will have also drawn tentative conclusions which others will contest. We look forward to feedback from others both in supplying additional information and in offering alternative analyses and proposals.

2 A preliminary classification of dialogue corpora

Before we turn to the different levels of representation or annotation, it is important to consider the various types of dialogue which have been investigated or modelled for LE purposes. This section contains an outline of some of the different types of dialogues that occur in different research projects and that are to some extent the basis for finding ways of categorising and identifying dialogue acts in Section 3.6 below.

For example, one of the most general types of dialogue concerns airline, train timetable or general travel inquiries. The German VERBMOBIL project specifically deals with *appointment scheduling* and *travel planning* tasks, while the TRAINS corpus developed at the University of Rochester, USA, deals with developing *plans to move trains and cargo* from one city to another. One of the major dialogue projects in the US, the ATIS (Air Travel Information Service) project, deals strictly with *providing air travel information* to customers, and major companies, such as *Texas Instruments* and *AT&T*, have been involved in the collection and evaluation of the corpus.[1] Other dialogue projects involve *furnishing rooms interactively* (COCONUT, University of Pittsburgh), *giving directions on a map* (HCRC Map Task, University of Edinburgh) and *explaining cooking recipes* (Nakatani et al., 1995). These are just a few of the tasks to which dialogue projects have devoted attention up to the present.

As yet, there does not seem to exist any complete or systematic typology of dialogues, which makes it difficult (for example) to establish a complete list of all the goals that might be involved in the annotation and use of dialogue material.[2] Broadly, dialogues can be classified and described by

[1] More information on some of the work that has so far been done on the ATIS corpus can be found in Section 3.5.6.

[2] Useful background for both external and internal aspects of dialogue description are to be found in the sociolinguistic literature of the past 30 years, for which Dell Hymes's work on the 'components of speech' and 'rules of speaking' is a seminal starting point (see Hymes, 1972/1986).

reference to either *external* or *internal* criteria. The former include situational and motivational factors. The latter include formal or structural factors, especially how the dialogue breaks down into smaller units or segments such as dialogue acts (see 3.6 below). However, there seems to be a definite need for such a classification in order to establish a valid list of criteria that are to be used for annotation: one that is based on actual experience and not on pure introspection. Such a list of criteria can then serve as a basic reference model that would need to be expanded only for special purposes that did not fit any of the existing criteria. A starting point for establishing such a typology is suggested below in 2.2.

2.1 Dialogue acts

However, first it will be convenient to introduce here the term *dialogue act*, which will recur in this chapter, and will be more fully explained in Section 3.6. Dialogue acts are the smallest functional units of dialogues, and are utterances corresponding to speech acts such as 'greeting', 'request', 'suggestion', 'accept', 'confirm', 'reject', 'thank', 'feedback'. When considering the overall communicative function of dialogues, it is as well to bear in mind that for annotation as well as for processing purposes, they are seen as decomposable into such basic communicative units.

2.2 Towards a dialogue typology

In principle, we need a typology of dialogues geared towards the needs of LE as they can be foreseen at present. In practice, present research not surprisingly shows a heavy concentration on certain rather straightforward kinds of dialogue: those with the features marked ** below.

<u>A. NUMBER OF PARTICIPANTS</u>

A.1 TWO PARTICIPANTS **
A.2 MORE THAN TWO PARTICIPANTS

Most dialogues in LE research have two participants only (at any one stage).[3] More than two participants greatly complicate the task note only of collecting data, but of modelling all levels of analysis and synthesis. The number of overlaps is likely to increase, thereby influencing the quality and analysability of speech and the complexity of annotation.[4]

<u>B. TASK ORIENTATION</u>

B.1 TASK-DRIVEN **
B.2 NON-TASK-DRIVEN

Almost all dialogues in LE research are task-driven; that is, there is usually a specific task (or possibly more than one task), which at least one participant aims to accomplish with the aid of the other(s). An example is the Edinburgh Map Task Corpus (Anderson et al. 1991) in which one participant guides another to trace a route on a map. Others are the TRAINS corpus (Allen et al. 1996), in which speakers develop plans to move trains and cargo from one city to another and the VERBMOBIL dialogues that deal with appointment scheduling and travel planning. In contrast, most conversational dialogues would be classified as non-task-driven.

<u>C. APPLICATIONS ORIENTATION</u>

[3] An exceptional case is the three-participant dialogue scenarios used in some VERBMOBIL projects, involving two negotiators and an interpreter/intermediary (see Jekat et al., 1997).

[4] It is of interest to mention, however, that large spoken corpora such as the 4.2-million-word demographic component of the BNCBNC (British National Corpus), although of little value to LE, often contain dialogues with many participants (see Burnard 1995).

C.1 APPLICATIONS-ORIENTED **
C.2 NON-APPLICATIONS-ORIENTED

Applications orientation is a relevant parameter particularly among dialogues which are task-driven. The Map Task corpus may be cited as an example of a non-applications-oriented dialogue type. However valuable its contribution to research, it cannot be seen to have direct commercial or industrial applications. In contrast, dialogues which have clear application to useful human-machine interfaces, such as those dealing with airline or hotel reservations, may be classified as applications-oriented.

D. DOMAIN RESTRICTION

D.1 RESTRICTED DOMAIN **
D.2 UNRESTRICTED DOMAIN

Again, most dialogues in LE are restricted to a relatively tightly-defined domain of subject-matter. All three of the examples in 2. above belong to a restricted domain. (On the other hand, an everyday dialogue at the dinner table would be an example of unrestricted domain.)

A typology of domains follows naturally, at this point, under D.1. The following are purely exemplificatory:

D.1 DOMAIN

D.1.1 travel **
D.1.2 transport **
D.1.3 business appointments **
D.1.4 telebanking
D.1.5 computer operating systems
D.1.6 directory enquiry services
D.1.7 (etc.)

Subclassification may also be needed: e.g., under ‘travel’, air travel, hotel bookings, and rail travel are subdomains.

E. ACTIVITY TYPES

E.1 COOPERATIVE NEGOTIATION **
E.2 INFORMATION EXTRACTION **
E.3 PROBLEM SOLVING
E.4 TEACHING/INSTRUCTION
E.5 COUNSELLING
E.6 CHATTING
E.7 (etc.)

Alongside domain, the *activity type* (Levinson 1979) to which the dialogue belongs is another variable defining the type of dialogue, particularly in terms of the constraints on the dialogue roles adopted by participants. For example, under E.1 in the VERBMOBIL three-agent dialogues the participants may be characterized as two ‘negotiators’ and one ‘interpreter/intermediary’. In E.2, the two participants may be characterized as ‘customer’ and ‘service-provider’ . In current dialogue research, there is a major division between two leading paradigms: *cooperative tasks* between human participants (such as negotiating appointments) (E.1) and *information extraction tasks* (such as obtaining information on a computer operating system) in which a human agent interrogates a computer system (or a human surrogate for a computer system) (E.2) (see Gibbon et al., 1998: 598 on 'dialogue strategies'). Other task-driven activity types include problem-solving (as in the Map Task Corpus), teaching/instruction, counselling, chatting and interviewing.

Relations between variables (C.) 'applications orientation' and (E.) 'activity type' are obvious. On the whole, applications-oriented dialogue corpora at present will be characterized as either E.1 or E.2. Similarly, constraints on (D.) domain and (E.) activity type are clearly interrelated variables. They help to delimit the nature of the *task* (see B.1 below). However, they can be considered independently: the Linguistic Data Consortium (LDC) Switchboard Corpus has dialogues in which speakers share a pre-determined topic or domain of discourse; however, the activity type is not constrained in any specific way.

At this point, we turn to a classification of *tasks*, which logically should have been slotted in earlier, after 'B. Task Orientation'. The reason why it has been postponed, is to show the relation of interdependence between, on the one hand, task and domain, and on the other hand, task and activity type.

B.1 TASK

B.1.1 Negotiating appointments and travel planning (VERBMOBIL) **
B.1.2 Answering airline/travel inquiries (ATIS) **
B.1.3 Developing plans for moving trains and cargo (TRAINS) **
B.1.4 Furnishing rooms (COCONUT) **
B.1.5 Giving directions to find a route on a map (Map Task)
B.1.6 (etc.) ...

Distinct tasks can be informally defined by the intention(s) of participants, the illocutionary function(s) of their utterances (Mc Kevitt et al., 1992) or by the end state which defines the successful accomplishment of the task. The number of tasks for which dialogue takes place is very large. Also, the amount of detail which may be specified to define the task for a particular dialogue is open-ended. Hence no closed set of 'task attributes' can be reasonably specified. As an example, consider the following as a succinct definition of the Map Task scenario (Thompson et al., 1995: 168):

> Each participant has a schematic map in front of them, not visible to the other. Each map is comprised of an outline and roughly a dozen labelled features (e.g. 'white cottage', 'Green Bay', 'oak forest'). Most features are common to the two maps, but not all. One map has a route drawn in, the other does not. The task is for the participant without the route to draw one on the basis of discussion with the participant with the route.

It is a sound practice to keep 'task' and 'domain' as separate parameters, recognizing that when a dialogue system has to be built for a particular application, the two parameters need to be intimately combined for the specification of that particular system. The separation of task and domain is particularly useful for the typology both of dialogues and of dialogue acts (see Section 3.6 below): it enables generalizations across indefinitely many different tasks and different domains to be built into the typology, and into the construction of suitably generic dialogue system software.

F. HUMAN/MACHINE PARTICIPATION

F.1 HUMAN–MACHINE DIALOGUE

F.1.1 SIMULATED (WIZARD OF OZ) **
F.1.2 NON-SIMULATED

F.2 HUMAN–HUMAN DIALOGUE

F.2.1 MACHINE-MEDIATED **

F.2.2 NON-MACHINE-MEDIATED

In corpus-driven methodology, there is always a problem of matching the naturally-collected data to the needs of the artificial LE system. One problem of dialogue research where this shows up strongly is in our lack of knowledge of how human beings will behave when conversing with computer dialogue systems. How far will they adapt, when talking to a machine, so that their dialogic behaviour is 'unnatural' by the standards of human-human dialogue? To answer this question, *Wizard of Oz experiments* (see Gibbon et al. 1998: 104-5, 143, 375-9) have been set up to simulate the behaviour of a machine in dialogue with a human being, and to record both the behaviour of the machine and the behaviour of the human being who believes he or she is interacting with a machine.

The other option under F.1, non-simulated human-machine dialogue, is clearly of limited value for R&D purposes, unless the computer system has already attained a basically satisfactory level of functionality. This has been described as a system-in-a-loop method (see Gibbon et al., 1998: 581).

To understand the way in which humans interact with machines is also important because there are many types of machine-mediation that may each influence the way dialogue is conducted in a particular way, both when communicating with the computer and with another human via the computer. Even using the telephone may be considered a form of machine-mediation restricting the transmission channel, although it is something we accept as part of our everyday lives and tend not to consider. Other forms of mediation may include or exclude other channels, such as video-conferencing systems or chat programs on the computer.

G. SCENARIO

G.1 SPEAKER CHARACTERISTICS

G.2 CHANNEL CHARACTERISTICS

G.3 OTHER ENVIRONMENT CONDITIONS

By *scenario* we mean the various practical conditions and attendant circumstances which affected the collection of the dialogue data. Such conditions are important to keep track of, since they might have had an effect (foreseen or unforeseen) on the value of the corpus as a basis for further research and development.

Speaker characteristics are often stored in a speaker database, and include how speakers were sampled; the age and gender of each speaker, the speakers' native language, their geographical provenance, their drinking and smoking habits (see Gibbon et al., 1998: 110 ff); whether speakers are known to one another; whether speakers are practised in the dialogue activity. Speaker characteristics also include (a) what language(s) was/were spoken, and (b) what the native language of each speaker is.

Channel characteristics include use of the spoken versus written medium; recording characteristics (e.g. whether multi-channel recording was used); use or non-use of a telephone line; availability of visual channel; recording in studio vs. recording on location; and so on.

Other environment conditions include not only general contextual factors, but also special design features used in the collection of data and affecting the nature of the outcome: e.g. a signal button was used in some VERBMOBIL recordings to request a turn, thereby eliminating turn overlaps and allowing speakers to formulate their ideas before speaking. Another dialogue manipulation device is the Wizard of Oz scenario mentioned above (under F.1.1).

3 Levels of representation or annotation

3.1 General coding issues

We will shortly turn to the examination and recommendation of representation and annotation practices at the specific levels listed towards the end of 1.3 above. But first, we should give attention to general coding issues which affect all these levels. Perhaps the overriding issue is whether all levels should follow same general encoding standards. There is much to be said for adhering to existing or emerging standardization initiatives, since this would make information exchange or display much easier and reduce the need for (re)-writing individual tools for each application. The best candidates to consider are the SGML-based TEI standardization initiative and the more recent emergence of the XML standard. In principle, they could apply to all levels of transcription and annotation. However, it is premature to be dogmatic on this issue. In the following sections, we discuss and exemplify TEI mark-up where appropriate, but at the same time we illustrate other forms of encoding where the data we are illustrating happen to be in these alternative forms. 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.

Another issue is the degree to which different levels of transcription or annotation make use of information provided by other levels. Here again, it would be premature to insist on too great a degree of conformity. Let us consider briefly the requirement of segmentation or 'chunking' at various levels. The orthographic transcription (3.2) will divide the dialogue up into *turns*, within which further units will typically be signalled, where necessary, by the use of full stops or other punctuation marks. The *'orthographic sentence'*, if indicated at this level, may be regarded as a pre-theoretical unit, arrived at more or less impressionistically by the transcriber, who may not have the expertise to make use of prosodic or other levels of information. At the syntactic level, a similar unit (termed in 3.4 a 'C-unit') may be recognized, but may not correspond one-to-one with the 'orthographic sentence' of the basic transcription. Equally, at the prosodic (3.5) and pragmatic (3.6) levels, segmentation may lead to the delimitation of *tone groups* or *utterances* which are important at those levels. Whereas in the longer run we may anticipate more integration of these units at different levels of analysis, it would be better at this stage to regard them as independent though correlated. The degree to which one level of annotation depends on another rests on factors such as the ordering of the procedures of annotation and the kinds of expertise the transcribers or annotators make use of. For purposes of implementation, however, segmentation at the orthographic, syntactic and/or prosodic levels may be seen as subservient to the task of isolating key pragmatic dialogue-units representing the communicative goals of the participants.

3.2 Orthographic transcription

The aim here is to represent the macro-features of the dialogue, including a verbatim record of what was said. A 'verbatim record' is a useful abstraction for many purposes, but it must naturally not be confused with the speech event itself. Some kind of hierarchy of priority seems to be needed in what kinds of macro-features of the dialogue to represent orthographically, and at what level of detail to represent them: see the Recommendations at the end of this Section.[5]

[5]There is a large literature on both practice and principle in the transcribing and coding of spoken language data. Particularly relevant to this section are the transcription conventions for SPEECHDAT corpora in Gibbon et al. 1998: 824-828. Two collections of studies of transcription more from the point of view of general linguistics and discourse analysis are those of Edwards and Lampert (1993) and Leech et al. (1995).

3.2.1 Background

This section takes account of the recommendations made by Llisterri (1996) and by Gibbon et al. (1998) within the EAGLES framework and of those made by Johansson et al. (1991) for the TEI, now largely codified in P3 (Sperberg-McQueen and Burnard, 1994). The corpus survey on which the following discussion is based comes partly from the document of Johansson et al. (1991) and partly from a fresh extension of it, which pays particular reference both to corpora produced for dialogue projects and to corpora in European languages other than English.

We try in particular to address the issue of integrating spoken and written resources — e.g., making representations of spoken corpora accessible to the language engineering (not just the speech technology) community. For this reason, we sometimes focus on *processibility* of texts (e.g., by stochastic or rule-based taggers and parsers) as an issue.

There is, at present, no strong consensus as to the means of representation, so that, e.g., whilst we may use examples based on the TEI, we do not assume the necessity of TEI conformance. Rather, we concentrate on the *features* that should be represented. However, some forms of representation naturally capture certain phenomena more easily than others: for instance, the start and end tags used in SGML/TEI are particularly useful for indicating the duration of a speech-simultaneous phenomenon such as a non-verbal noise. It might also be noted that, in choosing a representation scheme, individual symbols that could be confused with other markup should perhaps be avoided: for example, the `@` character used by VERBMOBIL to mark overlapping speech could possibly be confused with the SAMPA representation of the schwa character. The use of tags with whole-word representations (e.g., the Spanish `<simultáneo>`) would minimize this kind of confusion. However, with multi-layered 'stand-off' annotation that separates the annotated material from the actual annotation (cf. Thompson 1997), this would be less of an issue. The labels for the various tags can be standardized for any given language, but it is not necessary that a single specified language be adopted as a universal 'metalanguage': tools may be developed to translate between different language versions, where this is necessary for processing (e.g., in multilingual research).

The issue of obligatory vs. recommended vs. optional levels (cf. the recommendations on morphosyntax [Leech and Wilson, 1994]) is one that should also be addressed. Obviously, some applications will require more detailed transcription and analysis than others.

3.2.2 Documentation on texts

There are three primary ways of documenting information about texts:

1. a separate set of documentation — e.g., a manual
2. a header within the text itself, which may be
   (a) structured — e.g., a TEI header
   (b) relatively unstructured — e.g., a few lines of COCOA references.
3. separate documentation files with links (pointers) into the text. Those files may contain
   - pointers into a soundfile
   - a speaker database
   - etc.

Amongst the corpus linguistics community, a header has for some time been considered the minimum requirement for text documentation. An in-text header — as opposed to external documentation — makes it less easy to confuse texts: it can be used as part of an automatic analysis, to output background information; and it enables quick reference, especially when a manual is for some reason not to hand. On the other hand, in-text headers make for redundancy, if the same information has to be repeated in the head of each text using the same transcription scheme. This redundancy can be avoided by including in the header a reference or (better) a link to external documentation, in the form of a manual.

Whether a header or external documentation is used, then, as a bare minimum, it should normally contain an *identifier* for the specific text and basic *information on the speakers*. We recommend that additional information should include:

1. Speaker characteristics
   - number of participants
   - individual speaker attributes — e.g., age; sex; social class; native languages; regional accents
2. Channel characteristics
   - use of telephone line or other channels
   - recording details — e.g., time and date; technical specifications
3. General environmental conditions
   - contextual information — e.g., where the dialogue took place; under what physical conditions
   - human or machine or simulated (Wizard of Oz)
   - etc.
4. Other information
   - activity type (see Section 2 above)
   - degree of spontaneity
   - matters under discussion (domain/task)
   - details of the orthographic transcription
   - details of levels of linguistic annotation
   - contact details for obtaining additional information, for reporting difficulties or errors, etc.

The speech community, especially according to the decisions agreed on during the SAM-project, favours external files which can be distinguished via different extensions and are linked together via pointers (cf. Gibbon et al., 1998: 732 ff). There is good reason, for example, for separating a speech file (containing waveforms only) from associated descriptive files.

3.2.3 Basic text units

The most common text units in dialogue corpora are the *text* (i.e., a self-contained dialogue or dialogue sample with a natural or editorially created beginning and end) and the *turn* (or contribution). Tone groups are also sometimes marked. 'Orthographic sentences' (that is, units delimited by conventional written punctuation) are also often present (see 3.2.7.2), but these should probably be viewed as artefacts of transcription, rather than as real observable text units *per se*.

We suggest that the text and turn should be the basic text units in orthographic transcription, together with the intuitively-identified 'orthographic sentence'. There is no reason to include tone groups in orthographic transcription, as these are difficult to identify reliably (see Knowles, 1991): any marking of tone groups belongs to the interpretative stage of prosodic markup (Llisterri's [1996]

S3 level). Similarly, there is no reason to include *utterances*, whose identification belongs rather to the level of dialogue act annotation (see 3.6). The notion of turn is itself not wholly unproblematic, since interruptions and overlaps can occur, but there are methods for representing these aspects (see, e.g., 3.2.6 below). As noted, 'orthographic sentences' are often used in transcription for greater intelligibility and processibility (e.g., by taggers that assume the sentence as the basic processing unit), but it should be emphasized that the turn is a basic unit of *spoken dialogue* transcription, and that the 'orthographic sentence', delimited by turn boundaries and/or sentence-final punctuation, is merely a convenient impressionistic unit providing useful preliminary input to other levels of annotation.

3.2.4 Reference system

A reference system — i.e., a set of codes that allow reference to be made to specific texts and locations in texts — may be absent from transcribed spoken corpora. This is partly due to the fact that multiple versions of spoken corpora often exist, with a basic transcription being stored as one file and a time-aligned version being stored as a different file. A time-aligned file has, in essence, already a reference system, in that the time points can be used to refer to specific locations in the dialogue. Nevertheless, it is both useful and straightforward to introduce a basic reference system into ordinary orthographic transcriptions also. The references may be encoded either as a separate field, as in the TRAINS corpora:

```
58.3  : load the tanker
58.4  : then go back
```

or merged with speaker codes as in VERBMOBIL:

```
TIS019: gut  , bin mit einverstanden , dann ist das klar .
HAH020: danke sch"on <A> .
```

3.2.5 Speaker attribution

Speaker attribution is most often indicated by a letter code at the left-hand margin, but may sometimes be inferred from the turn, especially if there are only two participants in the dialogue. The code may or may not be enclosed in some kind of markup. Also, a speaker's turn may or may not be closed by an end tag. Sometimes, the code may be longer than a single letter; in VERBMOBIL, it also includes digits to indicate the turn number — see 3.2.4 above. Some examples are:—

FROM TRAINS:

```
57.1 M: puts the OJs  in the   tanker
58.1 S:       +southern route+
```

BASED ON THE TEI RECOMMENDATIONS:

```
<u who=A> Have you heard that she is back?</u>
<u who=B> No.</u>
```

FROM CREA[6]

[6]CREA is the Corpus de Referencia del Español Actual, a 10-million-word corpus containing a million words of transcribed speech compiled at the Real Academia Española. The corpus is SGML encoded and follows closely the conventions of the TEI and CES (Corpus Encoding Standard: see Ide et al., 1996). Further information can be obtained from joaquim.llisterri@cervantes.es or mpino@crea.rae.es.

```
<u who="anat00001.PER002" trans="smooth">Ha llamado.</u>
<u who="anat00001.PER001" trans="smooth">No, la hemos llamado nosotros.</u>
<u who="anat00001.PER002" trans="smooth">Bueno.</u>
```

The speaker identification codes used, such as
`<u who="anat00001.PER002" ...>`,
relate to information already given in the text header or accompanying documentation.

Cases where there is more than one speaker, or where the transcriber is unsure who is speaking, are normally explicitly indicated. The TEI, for instance, recommends the following practices:—

- for uncertainty:
  `<u who=A1 uncertain=medium>`
  where `uncertain` can take various values such as a comment on the degree or cause of uncertainty.

- for multiple speakers:
  `<u who='A1 B1 C1'>`

- for unknown speakers:
  `<u who=unknown>`

The same features can be marked with slightly different conventions in non-TEI markup schemes.

3.2.6 Speaker overlap

Speaker overlap, i.e., synchronous speech by more than one participant in the dialogue, is one of the most important issues in dialogue transcription. An examination of existing corpora demonstrates that the most common method of indicating overlapping speech is by '*bracketing*' the relevant segments of both interlocutors' speech, although the choice of bracketing characters varies considerably (e.g., @ preceded or followed by an overlap identifier number in VERBMOBIL, plus signs in TRAINS, SGML *tags* in the Corpus of Spoken Contemporary Spanish [Marcos-Marín et al., 1993 — hereafter 'CSCS']). Sometimes, the speech of only one of the two or more overlapping interlocutors is bracketed, although this is potentially less clear than the marking of *all* overlapping speech.

Three other methods of handling overlap may also be encountered:—

1. Vertical alignment, as in a musical score, of overlapping segments (widely used in conversation analysis and sociolinguistic transcription).

2. Reorganization of overlaps into separate turns, without representing where overlaps occur (as used, e.g., in the Czech national corpus).

3. The TEI practice of using time pointers, for example:—

```
<timeLine>
     <when id=P1 synch='A1 B1 C1'>
     <when id=P2 synch='A2 C2'>
</timeLine>
     ...
<u who=A>this is <anchor id=A1> my <anchor id=A2> turn</u>
<u who=B id=B1>balderdash</u>
<u who=C id=C1> no <anchor id=C2> it's mine</u></u>
```

The first alternative is technically problematic, as it often does not delimit with markup the precise stretches of speech that overlap: often only the start of an overlap is marked. Thus this information can easily be lost, especially when different display or print fonts are used that alter the visible alignment. The second is simply an idealization: it falsifies what is happening and obliterates any evidence of overlap in favour of neat, drama-like turns. The third (TEI) option is less objectionable, and has the advantage of dealing very well with multiple overlaps: e.g. where three speakers are talking simultaneously, and cross-bracketing would otherwise occur. For most purposes, it is perhaps a little too cumbersome in comparison with bracketing; however, a multi-layered approach to transcription and annotation — e.g., Thompson's (1997) suggestions using eXtensible Markup Language (XML) — can make it far less cumbersome for human users.

Occasionally, overlap bracketing crosses turns. In the CSCS, for example, a single overlap tag encloses the stretch of overlapping speech across speaker boundaries:—

```
<H1> <simultáneo>Sí, sí.
<H2> ...había</simultáneo> sido mucho más compleja la posición
```

This is, however, perhaps less clear than if the overlap markup were nested within the turns, thus:—

```
<H1> <simultáneo>Sí, sí.</simultáneo>
<H2> <simultáneo> ...había</simultáneo> sido mucho más compleja la posición
```

CREA uses `<overlap> ...</overlap>` tags, as has already been seen in the preceding section.

3.2.7 Word form

Most corpora transcribe speech using the standard (or dictionary) forms of words, regardless of their actual pronunciation. The use of standard word forms has a huge advantage, in that annotation and retrieval tools, for example, may be applied relatively unproblematically to speech as well as to writing.

Furthermore, everything (including numbers) is typically written out in full. Thus it is important to distinguish different ways of saying the same numeral: in German *2* may be pronounced as either *zwei* or *zwo.* Similarly, in English there are different ways of saying the same string of numerals: *1980* can be said as 'nineteen eighty' (the year) or as 'one nine eight oh' (a telephone number) or as 'one thousand nine hundred and eighty' (an ordinary number). Units of time, currency, percentages, degrees, and so on are normally transcribed in full to capture their pronunciations — e.g., *two hundred dollars and fifty cents* rather than *$200.50*; or *ten to twelve* rather than *11.50*. However, in some cases, it may be more straightforward to transcribe numbers simply in arabic numerals: for example, in a restricted domain such as airline travel dialogues, the majority of numerical expressions may be flight numbers, which will conform to a uniform system of pronunciation. A further possible argument in favour of the more 'simplified' form of transcription (e.g., *$200.50*) is that the actual pronunciation may be represented at another (phonemic) level, if a multi-layered form of transcription and annotation is employed.

Common contractions and merges that are also encountered in written texts (e.g., *can't, gonna*) are usually allowed, but otherwise dictionary forms are used, with special pronunciations indicated instead by editorial comments (see 3.2.13 below). In projects such as the BNC, a supplementary list was drawn up of those common allowable contractions, etc., that were not included in a standard dictionary. Spelling of interjections (e.g. the choice in English between *okay* and *O.K.*) can also be a problem: see Section 3.2.8.2 below. In practice, all lexical items that appear in a corpus should also appear in a lexicon, be it either an external, pre-existing standard dictionary or a lexicon specially generated from the corpus.

In some languages, compounds are also an issue for transcription. This is not a problem for languages such as German, but it is a problem for languages such as English, which, historically, have a more flexible approach to the representation of compounding. For instance, in English, one may find *keyring*, *key ring* or *key-ring*. It would be difficult, if not impossible, to lay down strict rules for the representation of compounds. The key essentials, therefore, are *internal consistency* of practice in representing compounds and *explicit documentation* of the practice adopted. If compounds *are* represented as multi-word units, it is possible to tag them as compounds at the morphosyntactic level (see 3.3.4).

Pseudo-phonetic/modified orthographic transcription tends to be reserved for oddities such as non-words or neologisms that have no true dictionary form. Letters of the alphabet that are pronounced individually are normally demarcated by spaces, to distinguish, for example, the two different pronunciations of *VIP* — /vɪp/ vs. /vi: aɪ pi:/. In CREA, the tag <distinct> is used for spelled-out words, with the attribute 'dele' (for 'deletreado'):

```
<distinct type='dele'>pe-e-erre-erre-o uve-e-erre-de-e</distinct>
```

(Here the speaker spells out the two words *perro verde*.) It is probably sufficient to separate these with spaces (e.g., `V I P`), but sometimes additional markup is encountered, as in VERBMOBIL: `$V $I $P`.

It has been suggested that a standard dictionary should be employed for each language as an arbiter, wherever needed, for these dictionary forms. The Duden has already been used in this way for German in VERBMOBIL, and the dictionary of the Real Academia Española has similarly been used for CREA. However, this may be a little too idealistic. Often, dictionaries present more than one possible spelling of a word — e.g., *analyze* vs. *analyse*. Also, it is difficult to conceive of transcribers checking spellings in a standard dictionary, when they feel confident of how to spell something. It may be that a style guide, such as Hart's *Rules* for English (Hart 1978), would help with restricting common variant spellings. For languages with less spelling variation and/or one standard 'academy' dictionary, the situation could be more straightforward. Where available, a better alternative would be to use special dictionaries that have already been developed during projects in the speech community. These tend to be based on experience and actual requirements for systems, and normally take into account all the problems encountered during system development.

For example, to reduce error rates in testing and training signal recognition systems based on a particular language model, frequently occurring assimilations between individual words have to be integrated into the dictionary because the system has to read and understand the transcriber's representation of the utterance, e.g. in German the spoken form *hamwanich* vs. the written form *haben wir nicht*.

3.2.7.1 Word partials. Word partials, also known as unfinished or truncated words, are typically transcribed as follows: as much of the word as is pronounced is transcribed, followed by a 'break-off' character — for instance a dash or an asterisk. Sometimes a tag is used instead of a special character, e.g., `<distinct type='titu'>` (for Spanish 'titubeo') in CREA. For example:

```
<distinct type='titu'>es*</distinct> estamos
```

In this case, an asterisk (*) is added to the end of the incomplete word.

Some guidelines (e.g., the Gothenburg corpus of spoken Swedish) also allow for word-final partials, in which case the 'word partial' character may occur at the beginning rather than the end of a string. Most transcriptions of word partials use standard or modified orthography, but this can be confusing in cases like the English digraph *po*–, which may represent either the diphthong of *poll* or the simple vowel of *pot*. It may thus be better to use some form of phonetic representation, such as SAMPA, for word partials; however, if there is a further level of phonemic transcription, then this is unnecessary.

An interesting aspect of the guidelines used by the TRAINS project is that an interpretation (or expansion to full form) of word partials is added where possible. This has both advantages and disadvantages. Where a partial is not part of a repeated sequence that includes a full form, it enables more content to be extracted for language understanding and so on, but, on the other hand, it may be argued that to interpret such partials — even when they seem unambiguous — is to read additional (and perhaps unwarranted) information into the transcript beyond what needs to be represented. Such interpretative information should preferably not appear at the level of orthographic transcription. Furthermore, word partials may also at times serve a communicative function, indicating that the speaker has changed his/her mind about what to say next or how to interpret something, and expanding them may thus lead to misinterpretation.

3.2.7.2 Orthography, including punctuation. As to the more general form of transcription, the use of a basic subset of the standard orthography is both normal and desirable. Sentence-initial capitals may be omitted, but, otherwise, normal capitalization and at least full stops tend to be used. This improves readability for the human user and improves processibility for taggers, parsers, and so on. Obviously, it is understood that such standard orthography is, to a considerable extent, interpretative when applied to speech, but its advantages outweigh its disadvantages. The use of punctuation characters other than full stops is an open question, but commas may sometimes have certain advantages as well. In English, for example, using a comma before a tag question is unambiguous and may actually help to identify the purpose of this particular phrase type as communicating a possible request for feedback: e.g. *Two o'clock, is it*. There is also a case for using question marks where the transcriber clearly perceives an utterance as a question. This can be useful especially where the structure of the utterance does not mark it as interrogative. There are many questions in which lack such marking (e.g. *Next week?*), and their import is not clear to a reader who does not have access to the prosodic level of annotation.

Whatever punctuation scheme is adopted, the general rule must be to explain it in the text documentation, e.g. in the header. For example, if punctuation has been used, it should be explicitly stated which punctuation marks have been employed, and how they have been assigned (whether impressionistically or otherwise).

3.2.7.3 Unintelligible speech. It is sometimes impossible to decipher — at least in part — what a participant is saying, because of unclarity in the recording. Normally a single code is used — e.g, `<inintelligible>` in the CSCS or `<%>`, added directly to the word, in VERBMOBIL. Sometimes a form of bracketing is employed instead, with the number of unintelligible syllables given. An estimate of the number of unintelligible syllables is desirable, but it is emphasized that this estimate can only be approximate.

3.2.7.4 Uncertain transcription. In other cases, the transcriber can hazard a guess as to what was said, but wishes to indicate the existence of uncertainty. Normally, such uncertain transcriptions are bracketed in some way, but with conventions different from those used for truly unintelligible speech. Here are two examples of ways of marking uncertain transcriptions:—

- Uncertain *syllables or sounds*: in the CSCS, these are bracketed within the word, thus: `burri<(t)>o`.
- Uncertain *words and phrases*: in the TEI, these are placed inside a set of start and end tags, e.g., `<unclear>burrito</unclear>`. The TEI tag shown here also has an optional attribute `reason`.

3.2.7.5 Substitutions. Also to be considered under this heading are those cases where words — normally proper nouns — are to be replaced for confidentiality or other reasons. These may be marked with codes, since this makes it more clear where an original text word has been replaced. The practice of simply substituting an alternative name without comment is sometimes encountered, but should perhaps be avoided; however, a replacement could be used if it is commented, e.g., by the use of a TEI regularization tag:—

```
<reg>Bert</reg>
```

Obviously, in circumstances of confidentiality, the `orig` attribute, which normally encodes the original form of words, cannot be used.

3.2.8 Speech management

By 'speech management' we understand the use of phenomena such as quasi-lexical vocalizations, pauses, repairs, restarts, and so on.

Although speech management is normally an issue for transcription, it should be noted that sometimes phenomena included under this heading are instead *annotated* at a separate level of processing — cf. the so-called *dysfluency annotation* of the Switchboard corpus in the Penn Treebank project.[7]

3.2.8.1 Pauses. Unfilled pauses (by which we mean *perceived* pauses, rather than silence in the speech signal) are typically marked with suspense dots (...) or some other special punctuation such as an oblique slash. It is important to distinguish short pauses from longer pauses or silences, which may indicate an interruption by some non-conversational event, activity, etc. The Gothenburg Swedish corpus uses various numbers of slashes (/, //, or ///) to give an impression of the length of a pause. Sometimes a tag is used instead of punctuation — e.g., `<P>` in VERBMOBIL. Both methods may allow additional comments to be added as to the length of a pause.

3.2.8.2 Quasi-lexical vocalizations. Most corpora make some attempt to standardize the transcription of quasi-lexical vocalizations, such as interjections and filled pauses such as *um, uh-huh, oi, ooh* and *ah*. In contrast, the CSCS avoids the use of invented/idealized word forms and instead uses markup to indicate where quasi-lexical vocalizations occur. The down side of this, however, is that such features can confuse transcription with dialogue-actdialogue act annotation: they require an interpretation of the function of a vocalization (e.g., agreement, negation). A possible third way, which is mentioned as an option by the TEI guidelines, would be to merge the two systems, so that quasi-lexical vocalizations have standardized forms but occur in the form of markup to indicate that standardization has occurred. For example:—

```
<vocal type=quasi-lexical desc=uh-huh>
```

However, this approach may be found to be too verbose and cumbersome. It may be better simply to use a standard list of orthographic forms for these phenomena, without any additional markup, and this approach is also sanctioned by the TEI. Whichever approach is adopted, it is useful to draw up a standardized and generally acceptable list of these quasi-lexical forms for each language, so that unwanted variants do not proliferate, causing retrieval problems.

3.2.8.3 Other phenomena. Many corpora do not explicitly identify repetitions, repairs, etc. However, for the purpose of activities such as part-of-speech tagging or speech recognition (cf. section 3.5.6), it may be important to do so, so that, for example, repetitions do not form part of the dialogue model and therefore disturb the working or training of a Markov model of category transitions. If repetitions and so on are identified in the transcription, it is probably desirable that one full-word transcription should be retained in the main running text and the rest marked up

[7] For an example, see http://www.cis.upenn.edu/~treebank/switch-samp-dfl.html; for the manual, see ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.

with some kind of bracketing. The TEI's <del> tag is one possible way of representing this and allows the various types of phenomena to be noted:—[8]

```
<del type=truncation>s</del>see
<del type=repetition>you you</del>you know
<del type=falseStart>it's</del>he's crazy
```

3.2.9 Paralinguistic features

By 'paralinguistic features' we mean those concomitant aspects of voice such as laughter, tempo, loudness, and so on that occur during speech. We exclude features that do not accompany speech but rather occur in isolation (e.g., laughter not superimposed on speech), for which see 3.2.10 below.

Paralinguistic features tend to be encoded with a finite set of standard features, but sometimes also free comment is allowed. A standard list of codes will enable features to be retrieved and counted in concordancing software, etc. Unconstrained comment tags should perhaps be avoided as much as possible. The TEI has already produced a basic list of paralinguistic features, which can be used or amended for EAGLES purposes; these are reproduced in Appendix A of this document.

The use of balanced start and end tags will enable the duration of a paralinguistic phenomenon to be encoded more clearly.

3.2.10 Non-verbal sounds

Non-verbal sounds are typically transcribed as a form of comment. Sometimes, a standard set of codes is defined in place of free comment.[9] However, it may be advisable for at least one more general feature to be retained (e.g., `noise`), to allow for unattributable sounds or those for some reason omitted from the standard list. It is possible, following the practice of the CSCS, to combine standard features and free comment, so that additional information is available as well as a basic indication of broadly what kind of noise has occurred.

Minimally, four types of non-verbal sound might be differentiated:—

1. non-verbal but *vocal* utterances attributable to the speaker (e.g., a laugh, or audible intake of breath)
2. non-verbal but *vocal* utterances not attributable to the speaker (e.g., an unattributed grunt)
3. non-vocal noises attributable to the speaker (e.g., snapping fingers)
4. non-vocal noises not attributable to the speaker, including noises that are not humanly produced (e.g., a dog barking, a doorbell ringing)

Again, as with paralinguistic features, the use of start and end tags allows a continuous noise to be represented.

3.2.11 Kinesic features

Kinesic features comprise what is, in informal speech, termed 'body language' — e.g., eye contact, gesture, and other bodily movements. Few corpora represent these features, since transcription is

[8]But see 3.2.7.1 above for a preferred method of transcribing truncations (phonetic representation rather than orthographic characters).

[9]Good starting points for a typology of non-verbal noises would be the two noise databases, 'Noise-ROM-0' and 'Noisex': see Gibbon et al. (1998: 8)

typically from audio rather than from video data or a live performance. In the past, kinesic features have been of less relevance to natural language and speech research than have the other features discussed in this document; however, as work on audio-visual speech synthesis progresses, they are likely to become much more relevant. But, since these have been investigated by the Multimodal Working Group of EAGLES, guidelines on such features belong to another chapter. We may note, however, that in an auditory transcription they can be included as editorial comments or using the TEI's `<kinesic>` tag, which has attributes to indicate the 'actor', a description of the action, and whether or not it is a repeated action.

3.2.12 Situational features

Basic information about the context of a dialogue (e.g., the participants, location, etc.) tends to be included in the text header or equivalent descriptive documentation (see Section 3.2.2). More 'short-term' information, such as the arrival or departure of a participant, is normally introduced as editorial comment. For these features the TEI suggests a special comment tag (`<event>`) , with the same attribute set as `<kinesic>`.

3.2.13 Editorial comment

Editorial comment comprises a number of cases where an interpretative information needs to be added over and above the transcription of the phenomena described above. These include:

3.2.13.1 Alternative transcriptions. Pseudo-phonetic or modified orthographic transcription is largely avoided as a general rule. However, in at least some cases, it may be desirable to indicate, separately from a full phonetic/phonemic transcription, how a word or phrase was pronounced, e.g., because it is a dialect form or a homograph. Modified orthography in the transcription itself may cause difficulty in concordancing or processing the text and may, in any case, be misleading — e.g., for non-native speakers using the corpus. An approach similar to that adopted by VERBMOBIL might be preferable, namely that alternative transcriptions should be enclosed within markup brackets. A similar approach is recommended by the TEI using the `<reg>` tag:—

```
<reg sic='booer'>butter</reg>
```

If more than one standard orthographic word is included in a variant pronunciation, VERBMOBIL also adds a number indicating how many of the standardly transcribed words are represented by a given pronunication. This feature is not part of the TEI syntax for `<reg>`, but might be a optional addition. It would be less important in a TEI representation than in VERBMOBIL, since VERBMOBIL does not use start and end tags to bracket the stretch of speech. If using a number, *whatcha* in English, for example, might be represented with something like:—

```
<reg words=3 orig='whatcha'>what are you</reg>
```

In view of the development of the SAMPA conventions for encoding phonetic (IPA) transcriptions in 7-bit ASCII, it might be possible to represent alternative pronunciations in SAMPA format rather than in an idiosyncratic modified orthography:—

```
<reg orig='bU?@'>butter</reg>
```

Since many computers still use a 7-bit character set, it is probably advisable, for the time being, to stick with SAMPA rather than attempting to use richer forms of encoding such as Unicode.

3.2.13.2 General comments. General in-text comments are typically introduced within some form of distinctive bracketing. In addition to the comment itself, the Gothenburg corpus of spoken Swedish encloses the stretch of text to which the comment refers. Comments in this scheme can also be numbered. We feel that enclosing the text commented on does make the comments more transparent. Numbers are probably not essential (in the Gothenburg corpus, comments occur on a different line to transcribed text, which is why they are used there). In an SGML (but non-TEI conformant) representation, this would look something like the following:—

```
That is what <note comment="Which one?">Geoff</note> said.
```

3.2.14 Recommendations

As proposed at the beginning of this section, we will conclude with recommendations regarding the priority of information to be included in the orthographic representation of a dialogue. We provide for three levels of priority: ‘Highest priority’, ‘Recommended’ and ‘Optional’. These lists are by no means exhaustive, and features may be added to them, or moved from one list to another, according to the needs of this or that project.

3.2.14.1 Highest priority

- Text header (or equivalent documentation) with text identification and identification of speakers (3.2.2)
- Text header documentation to include information on (a) speaker characteristics, (b) channel characteristics and (c) environmental conditions, as recommended at the end of Section 2 (3.2.2)
- Dialogue divided into turns (3.2.3)
- Speaker of each turn made explicit (3.2.5)
- Standard spellings used wherever possible (deviations from standard spelling practices to be justified and documented) (3.2.7)
- Numbers, currency expressions, dates, clock times etc. written out in full (except where there is no risk of ambiguity, and where there are overriding reasons for economy) (3.2.7)
- Spelled-out letters (e.g. V I P) to be separated by spaces (3.2.7)
- Normal use of capitalization and full stops (but sentence initial capitals optional) – avoid use of abbreviatory stops (3.2.7.2)
- Overlapping speech in indexed brackets or tag pairs, these to be closed within turns (3.2.6)
- Word partials marked, with use of phonetic representation where needed (3.2.7.1)
- Quasi-lexical vocalizations transcribed using standard representations (3.2.8.2)
- Unintelligible speech tagged as such (3.2.7.3)
- Uncertain transcriptions tagged as such (3.2.7.4)

3.2.14.2 Recommended

- Text header (or an independent but linked document) to specify transcription conventions (3.2.2)
- Pauses tagged, and long and short pauses distinguished (3.2.8.1)
- Repetitions and false starts tagged (3.2.8.3)
- Paralinguistic features tagged using standard list of features (3.2.9)
- Non-verbal sounds tagged using standard list of features (3.2.10)

3.2.14.3 Optional

- Comments tagged (3.2.13.2)
- Pause lengths marked (3.2.8.1)
- Alternative pronunciations tagged and represented with SAMPA (3.2.13.1)
- Kinesic features tagged (3.2.11)
- Punctuation other than full stops (usage to be explained in header) (3.2.7.2)
- Short-term ituational features to be tagged in-text where appropriate (3.2.12)

3.3 Morphosyntactic annotation

Morphosyntactic annotation is also known as word-class tagging, POS (part-of-speech) tagging, or grammatical word tagging. It takes the form of associating a word-class label with each word token in a corpus. The set of *tags* used for labelling words in a particular language and in a particular corpus is known as a tagset. The list of tags, together with their definitions and the guidelines needed to map them on to a corpus, is known as a *tagging* scheme.

Previous work on morphosyntactic annotation within the EAGLES framework has primarily focussed on written language corpora and their relation to lexicons. Although in practice only a few European languages have been exemplified, in intention the framework adopted has been multilingual and language- and application-indepedendent. A number of EAGLES or EAGLES-related documents are relevant. Leech and Wilson (1994/1996) provides a set of preliminary recommendations for the morphosyntactic tagging of corpora; exemplary tagsets are provided for Italian and for English. This document has been closely coordinated with work on another document, Monachini and Calzolari (1994), which proposes a set of morphosyntactic guidelines for both lexicons and corpora, and which exemplifies tagsets in some detail for Dutch, English, Italian and Spanish. Three documents which provide draft morphosyntax guidelines for Italian, English and German respectively are Monachini (1995), Teufel (1996) and Teufel and Stöckert (1996). Of these, the German scheme (Teufel and Stöckert) is worked out in considerable detail.

Morphosyntactic information can typically be represented as a type hierarchy, with features and their values. The major 'pos' (part of speech) feature has such values as noun, verb, adjective, pronoun, adverb and interjection. More peripheral word categories are included under the values 'unique/unassigned' (e.g. infinitive and negative markers) and 'residual' (e.g. formulae, foreign words). Each of these values (except 'interjection', which tends to be undifferentiated) is then represented as a hierarchy table within which subcategories are shown as subsidiary features and values. For example, for nouns, the following features and values may commonly occur: Type (common, proper); Number (singular, plural); Case (nominative, genitive, dative, etc.); Gender (feminine, masculine, etc.). The range of features and values can obviously vary from one language to another, as can their hierarachical dependencies. But it is proposed that the morphosyntactic inventory for each language should be mappable into an **intermediate tagset** (Leech and Wilson 1994/1996, Section 4.3), which shows what is common between languages, while enabling the differences to be captured by optional extensions and omissions.

The actual formal representation or encoding adopted for morphosyntactic annotation can vary from one tagging scheme to another. One proposal for tagging within the TEI guidelines is found in the CDIF implementation for the BNC (Burnard, 1995; Garside et al., 1997: 19-33). Another, known as CES has been put forward for implementation as a general EAGLES standard by Ide et al. (1996: Section 5.2). The follow example illustrates the SGML-based CDIF tagging scheme for the BNC:

```
<w AV0>Even <w AT0>the <w AJ0>old <w NN2>women <w VVB>manage
<w AT0>a <w AJ0>slow <w UNC>Buenas<c PUN>, <w AV0>just <w CJS>as
<w PNP>they<w VBB>'re <w VVG>passing <w PNP>you<c PUN>.</PUN>
```

In this model, the primary textual data and the annotations are combined in a single file, the annotations being encoded as SGML tags. However, in the Corpus Encoding Standard (CES) model of Ide et al. (1996), preference is given to the mechanism of placing annotations in a separate file, with its own document type definition (DTD). In this case, cross-reference between the text itself and the annotation document is achieved by using HyTime-based TEI addressing mechanisms for element linkage. In effect, the text document and the annotation documents associated with it are handled as a single hyper-document (Ide et al., Section 5.0).

Our particular concern here, however, is with the linguistic decisions involved in morphosyntactic annotation of dialogue. It could be argued that this is not a special problem area for dialogue corpora, since the same word-class categories are likely to appear in both spoken and written texts. (Even 'ums' and 'ers' occur in fictional dialogue.) That there is no great difficulty here is suggested by the fact that the whole of the BNC, for example, has been tagged using the same tagset for the spoken data (c.10 million words) as for written texts (c.90 million words).

However, most tagsets have been devised primarily for written language, and the fact that the same tagset can be applied to spoken and written data should not lead us to ignores the fact that frequency and importance of word categories varies widely across the two varieties of data. Interjections and hesitators (or filled pauses) (*um, er* etc.) are vastly more frequent in speech than in writing. There are, in fact, two aspects of morphosyntactic tagging which need to be considered in adapting a tagset from written to spoken language:

(a) dysfluency phenomena:

  (i) How to tag pause fillers (*um, er,* etc.);
  (ii) How to tag word partials (e.g. where a speaker is interrupted in mid-word).

(b) Word-classes which are characteristic of speech, but not of writing:

  (i) How to tag discourse markers etc.
  (ii) How to tag peripheral adverbials

3.3.1 Dysfluency phenomena in morphosyntactic annotation

There are two problems to consider under this heading. The first is how to tag hesitators, i.e. filled pauses such as *um* and *er* in English. The second is how to tag word-partials (see 3.2.7.1 above) which result from repairs and incomplete utterances. In the so-called intermediate tagset proposed in EAGLES preliminary guidelines (Leech and Wilson, 1994/1996), there is, as already noted, a 'catch-all' peripheral part-of-speech category U ('unique' or 'unassigned') that can be used for these quasi-lexical phenomena. The guidelines also allow for the subdivision of this category U into subcategories such as Ux 'hesitator' and Uy 'word-partialword-partial' (where x and y are digits). It is highly recommendable that the morphosyntactic annotation of spoken language make use of such subcategories. Alternatively, the guidelines would allow the I (interjection) part-of-speech category to be subclassified to include hesitators (see (ii) below). Hence in this respect, although the existing morphosyntactic annotation guidelines are adequate, devising optional extensions such as the inclusion of new subcategories should be seriously considered.

On the other hand, an alternative solution is not to assign morphosyntactic tags to these items at all, but to mark them in the orthographic transcription as non-word vocalizations comparable to laughs and snorts (see 3.2.10 above). This solution is in tune with the proposal, discussed further in 3.4.1 below, to treat dysfluency phenomena as extraneous to the grammatical annotation of speech.

3.3.2 Word-classes which are characteristic of speech, but not of writing

Tagsets may need to be augmented to deal with spoken language phenomena such as discourse markers (*well, right*), pragmatic particles (*doch, ja*), and various kinds of adverbs (especially stance or modal adverbs and linking adverbs) which are strongly associated with the spoken language. Most of these forms might in a very general sense be termed 'adverbial' in that they are peripheral to the clause or sentence, are detachable from it, and may often occur in varying positions, particularly initial or final, in relation to any larger grammatical structures of which they are a part. They tend to have an important role in marking discourse functions and therefore in providing criteria for dialogue act classification (see Section 3.6 below).

3.3.2.1 Interjections in morphosyntactic annotation The interjection part of speech category (I) is badly served in the current EAGLES documentation, since no subcategories are recommended. However, analysis of spoken language corpora reveals the high frequency of a number of rather clear subcategories which are also relatively distinct in their syntactic and discoursal distribution. It is suggested, therefore, that these might be distinguished by different tags all beginning with the part-of-speech prefix I. Something like this proposal, put forward in two earlier articles (Stenström, 1990 and Altenberg, 1990) was adopted by Sampson (1995: 447-8) in his seminal discussion of the grammatical annotation of spoken English. His subcategory tags (which begin with U rather than I) include, in addition to familiar exclamatory interjections such as *oh* and *wow* (tagged UH), the following:

UA	Apology	(e.g. *pardon, sorry, excuse_me*)
UB	Smooth-over	(e.g. *don't_worry, never_mind*)
UE	Engager	(e.g. *I_mean, mind_you, you_know*)
UG	Greeting	(e.g. *hi, hello, good_morning*)
UI	Initiator	(e.g. *anyway, however, now*)
UL	Response Elicitor	(e.g. *eh, what*)
UK	Attention Signal	(e.g. *hey, look*)
UN	Negative	(e.g. *no*)
UP	*please*	as discourse marker
UR	Response	(e.g. *fine, good, uhuh, OK, all_right*)
UT	Thanks	(e.g. *thanks, thank_you*)
UW	*well*	as discourse marker
UX	Expletive	(e.g. *damn, gosh, hell, good_heavens*)
UY	Positive	(e.g. *yes, yeah, yup, mhm*)

Table 1: Sampson's subcategories for interjections.

This list is simply presented here as an illustration, showing that the **interjection** category in spoken language may be seen as much broader and more variegated than is allowed for in traditional grammar. This should not be worrying in that the Latin etymology of *interjection* suggests that it is something 'thrown between', in a sense that applies more or less happily to all the items above. They are grammatically 'stand-alone' items, capable of occurring on their own in a turn, or else of being loosely attached (prosodically speaking) to a larger syntactic structure, normally either at the beginning or, less commonly, at the end.

3.3.2.2 Adverbs in morphosyntactic annotation Like interjections, adverbs are dealt with cursorily by existing EAGLES guidelines and practices. Leech and Wilson (1994/1996) simply include recommended subcategories for base, comparative and superlative forms, as well as for interrogative adverbs such as *when, where* and *how.* Apart from these, various syntactico-semantic functions of adverbs (such as place, frequency and manner) can easily be recognized through optional extensions. On the whole, however, tagset makers have avoided subcategorising adverbs, on the following grounds. Adverbs constitute a loosely organized word class, in which even well-known subcategories, such as time, place, degree, manner and stance adverbs, are notoriously difficult to distinguish by hard-and-fast criteria, and certainly difficult to recognize and tag automatically. Yet it is worth noting that two tagsets for English which have been devised with spoken corpora in mind do subcategorize adverbs in considerable detail. These are the London-Lund Corpus tagset (Svartvik and Eeg-Olofsson, 1982) and the International Corpus of English tagset (Greenbaum and Ni, 1996). The following table with brief extracts from the London-Lund Corpus tagset gives an impression of how the adverb part of speech can be usefully subcategorized for spoken language.

TAG	CATEGORY	SUBCAT	SUBSUBCAT or ITEM	EXAMPLE
AApro	adverb	adjunct	process	*correctly*
AAspa	adverb	adjunct	space	*outdoors*
AAtim	adverb	adjunct	time	*how*
...	...	...	...	...
AQgre	adverb	discourse item	greeting	*goodbye*
AQhes	adverb	discourse item	hesitator	*now*
AQneg	adverb	discourse item	negative	*no*
AQord	adverb	discourse item	order	*give over*
AQpol	adverb	discourse item	politeness	*please*
AQpos	adverb	discourse item	positive	*yes, [mm]*
AQres	adverb	discourse item	response	*I see*
...	...	...	...	...
Asemp	adverb	subjunct	emphasiser	*actually*
ASfoc	adverb	subjunct	focusing	*mainly*
ASint	adverb	subjunct	intensifier	*a bit*
...	...	...	...	...

Table 2: Some adverb subcategories from the London-Lund Corpus.

Again, this partial list is not intended as a model to be recommended, but it does illustrate something of the diversity and importance of adverbial components in speech, and the need to consider carefully the addition of subcategories to the tagset before undertaking a morphosyntactic tagging of spoken data.

3.3.3 Extending the part-of-speech categories in EAGLES morphosyntactic guidelines

Returning to the *interjection* category, one tentative proposal is for an extended use of the I ('interjection') POS category in the EAGLES morphosyntactic guidelines (Leech and Wilson, 1994), with the following subcategories, based on those in Biber et al. (forthcoming 1999), Ch. 14:

TAG	CATEGORY	EXAMPLES (English)
I1	exclamations	*oh, ah, ooh*
I2	greetings/farewells	*hi, hello, bye*
I3	discourse markers	*well, now, you know*
I4	attention signals	*hey, look, yo*
I5	response elicitors	*huh? eh?*
I6	response forms	*yeah, no, okay, uh-huh*
I7	hesitators/filled pauses	*er, um*
I8	polite formulae	*thanks, sorry, please*
I9	expletives	*God, hell, shit*

Table 3: Extended interjection POS categories.

These subcategories cover the major 'interjection' phenomena which occur in spoken English generally. However, there is one major caveat over their use in morphosyntactic annotation: many of the words in these classes are liable to occur in more than one of the subcategories, so that ambiguity can be a major headache for automatic tagging, or even for manual tagging. For example, *oh*, classified above as an exclamation, in many instances behaves more like a discourse marker; *okay*, classified as a response form, can also occur as a response elicitor and as a discourse marker. A way out of this problem is to regard all the subcategory names in the table as preceded by the word 'primarily': e.g. *oh, ah,* etc. are designated as 'primarily exclamations', leaving any ambiguities at this level unresolved.

3.3.4 Residual problems

The sections on interjections and adverbs above illustrate two further difficulties to bear in mind when tagging spoken data.

One is the extremely unclear boundary between these two peripheral parts of speech. We note, in fact, that the two tagsets above, that of Sampson for the SUSANNE Corpus, and that of Svartvik and Eeg-Olofsson for the London-Lund Corpus, are somewhat inconsistent with one another in where they draw the boundary: whereas Sampson places greetings such as *good-bye*, response forms such as *yes* and the politeness marker *please* among interjections, Svartvik and Eeg-Olofsson place them among adverbials. This is an area where drawing the line between categories appears to be little more than an arbitrary decision.

Another phenomenon of spoken language illustrated above is the tendency for multi-word expressions such as *I see, I'm sorry, thank you* and *sort of* to occur with greater density than in written texts. It might be argued that this phenomenon of **multi-words** can be ignored, if one really wants to, in tagging written language (as indeed it is ignored by some well-known taggers). But it can scarcely be ignored in tagging spoken language. The problem, for morphosyntactic annotation, is whether these expressions should be decomposed into their individual orthographic words for tagging purposes, or whether they should be assigned a single tag labelling the whole expression, as in the lists above. If a single **multi-tag** is used, this raises the question of how to represent, in the formal encoding of morphosyntactic tags, this discrepancy of 'more than one orthographic word = one morphosyntactic word' (see Garside et al., 1997: 20-22).

3.3.5 An alternative solution

An alternative solution is to argue that the different kinds of 'interjection' in 3.3.3 above really differ on the functional plane, and that therefore these distinctions belong not to the level of morphosyntactic annotation, but to that of pragmatic annotation (see further 3.6.6.1 below). The rationale for this approach is provided by Fischer (1996), Fischer (1998) and Fischer and Brandt-Pook (1998), where it is shown that a broad class of **discourse particles** can be differentiated functionally and distributionally in a way that facilitates the automatic analysis of dialogue. The discourse functions which these particles perform comprise a limited list: *take-up, backchannel, frame, repair marker, answer, check, modal* and *filler*. On the morphosyntactic level, however, Fischer suggests (personal communication) that a broad differentiation between conjunctions, modal particles and discourse particles may be sufficient (with the possible addition of multi-word categories of speech routines (e.g. *you know*) and pragmatic idioms (e.g. *good-bye*). Of these, conjunctions (e.g. *but*) are connective, being outside the sentential unit themselves, while modal particles (e.g. *schon*) are integrated into the sentential unit and the intonation contour, and discourse particles (e.g. *okay* are not grammatically integrable, but are able to constitute entire utterances.

3.3.6 Recommendations

It would be premature to lay down hard and fast guidelines for the morphosyntactic tagging of dialogue. The most that can be done for the present is to recommend to dialogue corpus creators that they consult existing EAGLES or EAGLES-related documentation relating to morphosyntactic annotation (especially Leech and Wilson, 1994/1996 and Monachini and Calzolari, 1994). At the same time, they should bear in mind that the EAGLES standard for morphosyntactic annotation is still evolving, and that, in particular, there is need to augment and otherwise adapt existing guidelines to the annotation needs of spontaneous dialogue.

3.4 Syntactic annotation

Syntactic annotation has up to now taken the form of developing **treebanks**(see e.g. Leech and Garside 1991, Marcus et al., 1993) or corpora in which each sentence is assigned a tree structure (or partial tree structure). Treebanks are usually built on the basis of a phrase structure model (see Garside et al., 1997: 34-52); but dependency models have also been applied, especially by Karlsson and his associates (Karlsson et al., 1995). Until very recently, little spoken data has been syntactically annotated. There is an EAGLES document (Leech et al., 1996) proposing some provisional guidelines for syntactic annotation, but this again, while acknowledging their existence, omits to handle the special problems of syntactically annotating spoken language material.

With syntactic annotation, as with tagsets, the inventory of annotation symbols has been generally drawn up with written language in mind. An example of syntactic annotation of written language is the following sentence from a Dutch journal, encoded minimally according to the recommended EAGLES guidelines of Leech et al. (1996):

```
[S[NP Begin juni NP] [Aux worden Aux] [VP[PP in [NP het Scheveningse
Kurhaus NP]PP] [NP de Verenigde Naties NP-Subj] [AdvP weer AdvP]
nagespeeld VP]. S]

(At the beginning of June the United Nations will again be enacted in
the Scheveningen 'spa'.)
```

The following is an example of a different syntactic annotation scheme, that of the Penn Treebank (ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/), applied to a spoken English sentence:

```
( (CODE SpeakerB3 .))
( (SBARQ (INTJ Well)
         (WHNP-1 what)
         (SQ do
             (NP-SBJ you)
             (VP think
                 (NP *T*-1)
                 (PP about
                     (NP (NP the idea)
                         (PP of
                             ,
                             (INTJ uh)
                             ,
                             (S-NOM (NP-SBJ-2 kids)
                                    (VP having
                                        (S (NP-SBJ *-2)
                                           (VP to
                                               (VP do
                                                   (NP public service work))))
                                        (PP-TMP for
                                                (NP a year))))))))))
         ?
         E_S))
```

Just as with morphosyntactic annotation (see Section 3.3), we note that in early development of syntactic annotation (especially the IBM-Lancaster treebank, 1987-1991 — see Leech and Garside 1991), there seemed to be nothing seriously inappropriate in the use of syntactically-annotated **written texts** on a large scale as a training corpus for **speech** recognition applications.

Recently, the development of treebanks including or comprising spoken language has confronted a number of research groups with the same problem of adapting syntactic annotation practices to spontaneous spoken language. Four research groups which have been tackling this problem for English data are:

- UCREL, Lancaster (see Eyes, 1996) working on a sample treebank of the BNC
- Marcus and his associates working on the Penn Treebank [10]
- Sampson and his associates working on the CHRISTINE corpus at Sussex[11] (Sampson wrote an anticipatory Chapter 6 on treebanking spoken data in Sampson 1995, which reports on the earlier SUSANNE treebank of written data.)
- Greenbaum, Nelson, and others working on the International Corpus of English at University College London (Greenbaum 1996; Nelson 1996)

3.4.1 Dysfluency phenomena in syntactic annotation

Again as with morphosyntactic annotation, the adaptation of syntactic annotation is most needful to deal with dysfluency. The main phenomena requiring special treatment are:

- Use of hesitators or 'filled pauses'

[10] http://www.cis.upenn.edu/treebank/home.html .
[11] http://www.cogs.susx.ac.uk/users/geoffs/RChristine.html .

- Syntactic incompleteness
- Retrace-and-repair sequences
- Dysfluent repetition
- Syntactic blends (or anacolutha)

In considering what solutions may be applied to the syntactic annotation involving these kinds of dysfluency, we will mainly refer to solutions adopted by Sampson (1995: Ch.6) and for the UCREL syntactic annotation scheme by Eyes (1996). The other two research initiatives mentioned above (the Penn Treebank and the International Corpus of English) have taken a different approach, which bypasses the problem of syntactic annotation of dysfluencies entirely. They have adopted schemes for explicitly annotating dysfluencies. These features may then, if necessary, be excluded from the syntactically annotated material, by applying syntactic annotation only to a normalized version of the data. This normalized version may be represented, alongside a record of the dysfluent material, by the use of mark-up devices like the TEI deletion or regularization tags (see, e.g., 3.2.8.3 above). The approach of Sampson and of UCREL, on the other hand, is to include the dysfluent material in the syntactically annotated material, by means of a set of guidelines devised for that purpose.

3.4.1.1 Use of hesitators or 'filled pauses' Hesitators such as *um* and *er* can be handled relatively unproblematically (in Sampson's terms) by treating them as equivalent to unfilled pauses. In syntactic annotation of written corpora, generally, punctuation marks are incorporated into the syntactic tree, being treated as terminal constituents comparable to words. For the training of corpus parsers, this is a useful strategy, since punctuation marks generally signal syntactic boundaries of some importance. Similarly, for spoken language, it is an advantage to adopt the same strategy, and to treat pause marks like punctuation, as in effect 'words' in the parsing of a spoken utterance. This strategy is then extended to filled pauses or hesitators.[12] The general guideline adopted by UCREL and by Sampson (SUSANNE) is that punctuation marks are attached as high in the syntactic tree as possible; i.e. they are treated as immediate constituents of the smallest constituent of which the words to the left and to the right are themselves constituents. This policy generalises very naturally to hesitators, regarded as vocalized pause phenomena.

3.4.1.2 Syntactic incompleteness Syntactic incompleteness occurs where the speaker fails to complete an utterance, owing to self-correction, to interruption, or to some other disruption of the speech production process. In 3.3.1 above we discussed the case of word-partialsword-partial (incomplete or truncated words) as a problem for morphosyntactic annotation. On the syntactic level there is a comparable problem of non-terminal constituent partials, where a constituent is interrupted before its completion:

```
<pause> [NP you NP][VP 're [NP/ a British NP/]V] <pause>
```

This example from the BNC guidelines illustrates the use of a special marker (in this case a slash following the non-terminal constituent label) to indicate that the constituent is incomplete. In Sampson's scheme, instead, a marker is inserted *within* the incomplete constituent, to indicate the locus of the interruption:

```
[S [NP she ] [VP was going ] [PP into [NP the # ] ] ]
```

[12]In a similar spirit, in the International Corpus of English morphosyntactic annotation, hesitators are tagged with a 'negative' label UNTAG, which signifies that the item so tagged cannot be assigned to any part-of-speech category (Greenbaum and Ni, 1996).

(adapted from Sampson 1995: 454)

It should incidentally be noted here that, as a matter of principle as well as of practice, the issue of the (un)grammaticality of syntactically incomplete sentences does not generally arise with treebanks (see Sampson, 1987). In written data, as well as in spontaneous speech, ungrammaticality (by the standards of formally defined rule-driven parsers) is found to be of frequent and routine occurrence. Therefore any automatic syntactic annotation of spoken or written data has to cope with this phenomenon - for example, by the adoption of robust probabilistic parsing algorithms which will provide an adequate syntactic annotation for every sentence or utterance. No special dispensation is required for spoken data containing dysfluencies.

3.4.1.3 Retrace-and-repair sequences We will use this term to refer to frequent cases (also known as 'false starts') where a speaker 'interrupts' the production process by discontinuing the construction of the current constituent, returning to an earlier point of the same utterance (thereby notionally deleting the sequence 'retraced'), and restarting from there. Sampson proposed the use of a marker (again #) to signal the interruption point, and the inclusion of both the retrace and the repair within a minimal superordinate constituent:

```
and that [NPs any bonus [RELCL he ] # money [RELCL he gets over that ] ]
is a bonus
```

This example, liberally adapted from Sampson (1995: 453), uses the minimum bracketing needed to demonstrate the point. The labels adopted are those in the EAGLES preliminary syntactic annotation guidelines (Leech et al., 1996). The example shows how, on either side of the interruption point #, two relative clauses, the former incomplete, are handled as co-constituents of the same noun phrase.

3.4.1.4 Dysfluent repetition Repetition, as a manifestation of dysfluency, occurs where the speaker shows hesitation by repeating the same word, or the same sequence of words, before proceeding with the normal production process. The repetition can be iterated. In Sampson (1995) this repetition is again handled by the intervening use of the interruption-point marker #. It is treated, in effect, as a special case of a retrace-and-repair sequence, where the retrace and the repair are identical:

```
[O Oh [S [NP I ] [VP don't think ] # [NP I ] [VP don't think ] [NCL I ever
went to see mine ] S] O]
```

(This is again adapted from Sampson (1995: 457), with use of labelled bracketing in accordance with EAGLES syntactic annotation guidelines, to illustrate the point.)

3.4.1.5 Syntactic blends (or anacolutha): These occur where, in the course of an utterance, a speaker changes tack, failing to complete the syntactic construction with which the utterance began, and instead substituting an alternative construction. E.g. the switch to a non-matching tag question in: *And there's an accident up by the Flying Fox, is it?* (example from the BNC). Since no test of grammaticality is generally applied to treebank annotations, the annotation of cases like the one above causes no problem and probably needs no special annotation. More drastically incoherent sentences, however, do occur quite frequently in spontaneous speech. An example (from the BNC) is:

```
(1)    And this is what the, the $<$unclear$>$ what's name now
       now $<$pause$>$ that when it's opened in nineteen ninety-two
       $<$pause$>$ the communist block will be able to come through
       Germany this way in.
```

In this utterance, punctuated as a single sentence, there appear to be three word sequences between which there is no common superordinate constituent, and so a minimal analysis of the following general form is adopted according to the BNC guidelines (# is again added to indicate interruption points):

```
(1a)   [ And this is what the #, the <unclear> ] # [ what's
       name now # now  ] # <pause> [ that when it's opened in
       nineteen ninety-two <pause> the communist block will be
       able to come through Germany this way in ].
```

This example illustrates the effect of what the BNC guidelines call a 'structure minimization principle', which specifies that a syntactic annotation should not contain more information than is warranted in the context. A possible source of inconsistent parsing practice is that different grammarians will interpret the incoherent sentence differently — one reading into the sentence a particular structure, and another another. This can be avoided if annotators err on the side of omission rather than inclusion of uncertain information. In example (1a) above, there is no clear warranty for making the three major segments fit into a single overarching constituent. Similarly, it may be felt unwarranted to give particular syntactic labels to these segments. One option which is allowed in the BNC guidelines (again in line with the 'structure minimization principle') is the omission of labels where there are no clear criteria for the assignment of a particular label. This option is followed in (1a) above. On the other hand, there are arguable grounds for labelling the three segments as sentence (S), sentence (S) and nominal complement clause (NCL) respectively. Hence the following is an alternative, slightly fuller annotation:

```
(1b)    [S And this is what the #, the <unclear> S] # [S what's name
        now # now  S] # <pause> [NCL that when it's opened in nineteen
        ninety-two <pause> the communist block will be able to come
        through Germany this way in NCL].
```

3.4.1.6 Concluding remarks on syntactic annotation and dysfluency At present the syntactic annotation of spontaneous spoken language is in a pioneering stage, and the practices shown above should be regarded as highly tentative and incomplete. With this serious reservation, the above illustrations do show how syntactic annotation practices may be adapted to cope with dysfluent features of spontaneous speech. The two major methods employed — that of normalization by excluding dysfluencies and that of stretching syntactic annotation to include the parsing of dysfluencies — have complementary advantages. The normalization option enables spoken data to be automatically parsed with relatively little need to customize software for spontaneous spoken input, since major dysfluencies can be edited out. On the other hand, the inclusion option is preferable to the extent that is provides some parsing information even for incompleteness and repair phenomena. It can be pointed out, also, that the normalization procedure cannot be applied to some markedly dysfluent utterances such as example (1) above. Here it is not at all clear what a normalized version of the utterance would be.

3.4.2 Unintelligible speech

Another problem related to that of syntactic incompleteness arises in dialogue when the circumstances of speech production or of recording leave passages of speech unclear or unintelligible (cf.

3.2.7.3 and 3.2.7.4 above). Example (1) in Section 3.4.1.5 above shows how an unintelligible passage (tagged <unclear>) may be incorporated into a syntactic phrase marker. The general treatment of unintelligibility is parallel to that of incomplete constituents. Just as a marker # was introduced by Sampson (see 3.4.1.2) to signal the location of a point of interruption, so a marker such as <unclear> may signal the point where the parsing information cannot be recovered because of unintelligibility.

The tag <unclear>, unlike <pause>, refers to a verbal sequence. The only problem is that the annotators do not know which words the speaker used. The strategy here, then, is to *include* <unclear> within parse brackets wherever this appears appropriate, in order to 'complete' an otherwise incomplete constituent. Examples:

```
So [NP all these [ families and <unclear> ]NP]

No but [S <unclear> [NP twenty one NP]S] [S aren't you S]?
```

In the first case, it is obvious that <unclear> fills the gap in an otherwise incomplete coordinate construction. In the second case, the incompleteness arises from a gap at the *beginning* of the main clause. We can guess that the unclear words are *you are* or *you're*, because of the tag question which follows. So we have some warrant to include <unclear> within the [S ... S]. However, on the principle of minimising structure, we refrain from inserting any further brackets.

3.4.3 Segmentation difficulties

The syntax of spoken dialogue may seem fragmentary or disorderly for reasons other than dysfluency or unintelligibility. Some reasons are:

(i) The canonical sentence of written language, as a structure containing a finite verb, is far from the being a satisfactory basis for the segmentation of speech into independent syntactic wholes. According to one count by Leech (Biber et al., forthcoming, 1999, Ch.14), ca. 39% of the independent syntactic units of conversational dialogue have no finite verb: many are single-word utterances typically consisting of a single interjection in the extended sense of 3.3.2.1. The practice in the compilation of treebanks has often been to use parse brackets (conventionally [S ... S]) to enclose the whole parsable unit, but to make no assumption that what occurs within those brackets should have the structure of a canonical sentence. Thus a stand-alone noun phrase unit, such as *No problem*, should be parsed simply [S [N *No problem* N] S]. The [S ... S] brackets may be interpreted as 'sentence' or, say, as '(syntactic) segment', according to the annotator's or user's preference. For our present purpose, the term **C-unit**[13] will be used for a segment parsed as an [S ... S] which is not part of another [S ... S].

(ii) The criteria for what counts as a C-unit in speech are difficult to determine, and may have to rely on prosodic separation (for example the boundary of a major tone group or intonation phrase).

(iii) There are utterance turns in dialogue where one speaker completes a syntactic construction begun by another speaker.

There appear to be four methods of segmenting a dialogue into C-units:

[13] In educational linguistics, the term C-unit has evolved on the model Hunt's (1966) T-unit as a measure of syntactic complexity in children's written language. It is an attempt to define a 'maximal parsable unit' for speech. One attempted definition (Chaudron, 1988: 45) begins with a definition of the T-unit as 'any syntactic main clause and its associated subordinate clauses' and goes on to define a C-unit as 'an independent grammatical predication; the same as a T-unit, except that in oral language, elliptical answers to questions also constitute complete predication'. Although this definition is still influenced by written norms, the concept of a maximally parsable unit of spoken language underlies it.

(a) The C-unit should be delimited by criteria internal to syntax. That is, where no syntactic link can plausibly be established between one parsable unit and another, they are treated as independent. This solution, however, does not address point (ii) above.

(b) The C-unit should be delimited by prosodic criteria, either alone, or in conjunction with syntactic criteria where these are clear. This solution, obviously, depends on the existence and quality of a prosodic level of annotation.

(c) The C-unit should be delimited by orthographic criteria: that is, by treating sentence-final punctuation marks (specifically periods and question marks) as boundaries. This is the simplest method to apply, assuming that the orthographic transcription is so punctuated. On the other hand, it is the most arbitrary, since punctuation marks are artefacts of the transcription, and do not have a warranted linguistic function.

(d) The C-unit should be delimited by pragmatic, functional or discoursal criteria. Apart from the turn boundary, which is no doubt the clearest delimiter one can use for parsing, pragmatic and discoursal criteria are probably no clearer in determining C-units than internal syntactic criteria. However, in the development of language engineering dialogue systems, a considerable effort has been invested in the recognition of functionally-defined segments corresponding to dialogue acts. Moreover, in this context, the importance of syntactic annotation is in facilitating the automatic recognition and delimitation of such functional units, rather than parsing as an end in itself. Hence there is much to be said for relying on functional criteria as the most valuable guide to segmentation for purposes of dialogue annotation.

3.4.4 Recommendations

As for morphosyntactic annotation, it would be premature at the present stage to lay down hard and fast guidelines for the syntactic annotation of dialogues. In fact, syntactic analysis is more complex than morphosyntactic analysis, and its technology is at a less advanced stage of evolution. Moreover, there is relatively little consensus even on basic syntactic information. Here, as in Section 3.3, our policy must be to recommend to dialogue corpus builders that they consult the existing EAGLES provisional recommendations (in Leech, Barnett and Kahrel, 1996), and bear in mind the need to extend and modify these recommendations in the light of the needs of syntactic analysis for spoken dialogue.

We expressly avoid making any recommendations for defining a maximally parsable unit. In this area, the limits of syntax remain unclear, and there may be specific reasons why an annotator may need to align major syntactic boundaries with other boundaries, such as prosodic (see 3.5) or pragmatic (see 3.6.3) units.

3.5 Prosodic annotation

Prosodic labelling remains one of the major problem areas in the annotation of spoken data generally, and spoken dialogue in particular. This section takes the section on prosody in the *EAGLES Handbook* Gibbon et al. (1998: 161 ff) as its starting point, and brings it up to date in the light of recent work in the field.

In written text, as already noted in 3.2.7.2, use is sometimes made of punctuation marks to signal broad intonational distinctions, such as a question mark to indicate a final rise in pitch or a full stop to signal a final fall. Since it is well established that there is no one-to-one mapping between prosodic phenomena and syntactic or functional categories, it is important for a prosodic annotation system to be independent. In Southern Standard British English, for example, a rise in pitch may be used with a syntactically marked question, but this is not necessarily, and in fact not usually the case. On the other hand, questions with no syntactic marking often take a final rise, as, apart from context, it is the only signal that a question is being asked. A fully independent prosodic annotation allows for investigations into the co-occurrence of prosodic categories with dialogue annotations at other levels, once the annotations are complete.

Prosodic annotation systems generally capture two main types of phenomena: (i) those which lend *prominence*, and (ii) those which divide the speech up into *chunks* or *units*. Words are made prominent by the accentuation of (usually) their lexically stressed syllable. Many Western European languages have more than one accent type. It is thus necessary to capture not only on which word an accent is realised but also which kind of accent is used. Since in some cases the accent may occur on a syllable other than the primary lexical stress of a word, some annotation systems tag explicitly the syllable (or the vowel in the syllable) upon which an accent occurs, rather than the word as a whole. Such a representation, however, requires a finer annotation of the corpus at a non-prosodic level than simple orthography, e.g. a segmentation into syllables or phoneme-sized units.

Common to all annotation systems is the division of utterances into prosodically-marked units or phrases, where prosodic marking may include phenomena such as *audible pause* (realised as either actual silence or final lengthening), *rhythmic change*, *pitch movement* or *reset*, and *laryngealisation*. Dividing an utterance into such units is usually the first step taken when carrying out a prosodic annotation, as many systems place restrictions on their internal structure. However, the size and type of prosodic units proposed by the systems described below differs considerably.

It is currently common practice for a manual prosodic annotation to be carried out via auditory analysis accompanied by visual inspection of a time-aligned speech pressure *waveform* and *fundamental frequency* (F_0)) track. This is the case for the *ToBI* annotation system described in 3.5.1 below. Additional information, e.g. spectrogram or energy, may also be available. Despite this, we report on one system, *Tonetic Stress Marks* (TSM) in 3.5.2, which originally used to rely entirely on auditory analysis, since it is a well-established system which has been used for the annotation of a digitally available database.

Phenomena occurring across prosodically defined units, such as current pitch range, are not symbolically captured by any of the systems described below. A number of systems incorporate a means by which such information can be retrieved from the signal. For example, ToBI has a special label for the highest F_0 in a phrase. The F_0 value at this point may be used to give an indication of the pitch range used by the speaker at that particular point in time. *INTSINT* marks target points in the F_0 curve which are at the top and bottom of the range. However, the range is determined for a whole file which might be one or more paragraphs long. Register relative to other utterances is only captured in cases where the beginning of a unit is marked relative to the end of the previous one (e.g. in INTSINT). However, none of the manual annotation methods capture structures at a more macro level than the intonation phrase or its equivalent.

All existing representation systems for intonation have drawbacks. For a list and description of some of those systems, see Gibbon et al. (1998: 161 ff).

3.5.1 ToBI

The ToBI (Tones and Break Indices) system is an established standard for the prosodic annotation of digital speech databases in General American English. It has been successfully applied to Southern Standard British and Standard Australian English, but, since it is an adaptation of a phonological model, it is not claimed to be applicable as it stands to other varieties of English.

It has been made clear in the ToBI documentation that ToBI does not cover varieties of English other than those listed above, and that modifications would be required before it could be used for their transcription. In the ToBI guidelines it is stated that "ToBI was not intended to cover any language other than English, although we endorse the adoption of the basic principles in developing transcription systems for other languages, particularly languages that are typologically similar to English" (Beckman and Ayers Elam, 1997: section 0.4). The implication in Silverman et al. (1992)

that ToBI aimed to meet the need for a suprasegmental equivalent to the IPA is therefore to be ignored. It is the basic principles behind ToBI, rather than a set of phonologically-motivated categories, which allow its adaptation to other languages.

A ToBI transcription consists of a speech signal and F_0 record, along with time-aligned symbolic labels relating to four types of event. The two main event types are tonal, arranged on a *tone tier* and junctural, arranged on a *break index tier*. There is additionally a *miscellaneous tier* for the annotation of non-tonal events such as voice quality or paralinguistic and extralinguistic phenomena, and a further tier containing an orthographic transcription, the *orthographic tier*. The tone and break index tiers are discussed below.

3.5.1.1 ToBI Tones As far as the tonal part of ToBI is concerned, the basic principles are taken from the phonological model of English intonation by Pierrehumbert (1980). This model has given rise to a substantial number of studies within what has been termed by Ladd (1996) as the *autosegmental-metrical* framework. Some of these studies have developed into similar ToBI systems for other languages. Others lay down the groundwork for such an adaptation, but have not yet been applied to the annotation of large-scale corpora.

Within the autosegmental-metrical framework, tones are used in two major ways: they can be part of an *accent* or they can be involved in the signalling of a *boundary.* Tones may be *high* (H) or *low* (L). Accents may contain one or more *tones.* If there is more than one tone in an accent, it is important that the tone which aligns with the prominent syllable be marked as such. This is done by means of an asterisk (or star) diacritic. By default, monotonal pitch accents have the star on their only tone. The inventory of pitch accents is language or dialect specific.

Tones signalling the boundaries of prosodically defined phrases may occur at their left or right edges. Whether a tone (or, in principle more than one tone) may occur at a boundary of a given domain is, again, specific to individual languages or dialects, as is the number and types of domain which allow for tonal marking.

The ToBI inventory for General American English (more recently referred to as E_ToBI) has five basic pitch accents, the glosses are taken from Beckman and Hirschberg (ToBI annotation conventions):

H*	'peak accent'
L*	'low accent'
L+H*	'scooped accent'
L*+H	'rising peak accent'
H+!H*	'clear step down onto the accented syllable'

Table 4: ToBI Pitch Accents

All of the H tones in the above inventory may be marked with a '!' diacritic which indicates that they are downstepped relative to the immediately prior H tone. The downstep diacritic is obligatory in the H+!H* accent. The others, if downstepped would be transcribed !H*, L+!H*, L*+!H, and, in principle, !H+!H*. The prerequisite for using a ! diacritic is that there must be at least one H tone prior to the downstepped tone from which it can be stepped down.

There are two domains at the right edge of which there is an obligatory tone: the *intermediate phrase* and the *intonation phrase*. Intonation phrases contain at least one intermediate phrase. The tones available at the right edge of the intermediate phrase are:

(1) L-

(2) H-

The right edge of an intonation phrase is automatically the right edge of an intermediate phrase. It is customary to label the sequence of tones at these two right edges together. Since there is also the choice of H or L tone at the intonation phrase boundary, there are four combinations to choose from:

(1) L-L%

(2) L-H%

(3) H-H%

(4) H-L%

The '-' diacritic is used for intermediate phrase boundaries and '%'for intonation phrase boundaries. One problematic aspect of the transcription of boundaries is the fact that the phonetic implementation of the tone sequences is far from transparent. The H% or L% is raised by an automatic 'upstep' if it follows H-. This means that H-H% symbolises a high rising boundary reaching a level very high in the current pitch range (H% is upstepped), and H-L% symbolises a high level boundary (the L% is upstepped to the same value as the previous H- tone).

One further edge tone may optionally be used. This is an intonation phrase initial boundary tone, transcribed: %H.

3.5.1.2 ToBI Break Indices In the current ToBI system, there are five levels of perceived juncture, referred to as *break indices*, between words transcribed on the orthographic tier. They are numbered from 0 to 4. The lowest degree of juncture between two orthographic words is level 0, where the words are grouped together into a 'clitic group', e.g. between 'did' and 'you' pronounced as 'didya'. Level 1 is the default boundary between two words in the absence of any other prosodic boundary. Levels 3 and 4 correspond to intermediate phrase and intonation phrase boundaries. Since these latter two break index levels are linked to the tonal representation, the system might be argued to be circular. However, there is provision in the system for signalling cases where there is a mismatch between the tonal boundary transcribed and the perceived juncture. This is provided by a '-' diacritic, as in '4-', and by level 2. Level 2 can be used where there is tonal evidence to indicate a level 3 or 4 boundary but a lower degree of perceived juncture. Alternatively, it can also indicate a high degree of separation between the words without the corresponding tonal evidence.

It is important to point out here that the break indices are perceptual categories. In order to assign them, transcribers need make use of auditory information only.

3.5.1.3 Using the ToBI system A major advantage of the ToBI system is that there are extensive training materials and well-developed tools for carrying out the annotation. For instance, a transcriber can hear cardinal examples of all of the pitch accent and boundary types at any point during transcription.[14]

The fact that the majority of ToBI users are also users of ESPS/waves+^{TM} has been a distinct advantage to users of this system over others in a number of respects: regarding access to training materials, which were initially only available in digital form in ESPS format (the alternative being audio cassette with paper records), exchange of data with other transcribers, and the availability of transcription tools within ESPS including phrase-internal syntax checkers.[15] However, the fact that ESPS/waves+^{TM} is a commercial product has been an obstacle for those with alternative software

[14] The ToBI labelling guide, including electronic text and accompanying audio example files, is available at http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html .

[15] The current ToBI transcription tool, 'tobitool', for transcribing with xwaves can be obtained from http://www.entropic.com/tobi.html .

wishing to learn and use the ToBI system, or for research institutions that do not have sufficient funding.

There have been recent attempts to address this imbalance, in that training materials are now available over the world wide web with incorporated audio files and time-aligned transciptions, F_0 tracks and speech waveforms. A .au/.gif format version of the Guide is currently available in beta version at the ToBI homepage URL. Furthermore, a public domain program, 'fish', which uses Tcl/Tk running under Unix, has been developed by members of the German ToBI group.[16] It supports data exchange using Esprit SAM formats. Provided that this public domain software continues to be available, the ToBI system can be recommended, as long as it is not adopted wholesale for a dialect or language for which it has not already been adapted.

3.5.1.4 ToBI for other languages and dialects J_ToBI is a transcription standard for Standard (Tokyo) Japanese, developed in collaboration between linguists at Ohio State University, USA, and speech engineers at *ATR* Interpreting Telecommuncations Research Laboratories, Japan.[17]

GToBI is a consensus transcription system for German developed by a multi-site group including Universities in Saarbrücken, Braunschweig, Stuttgart, Erlangen and Munich.[18]

[16] It is currently available at the following address: http://sbvsrv.ifn.ing.tu-bs.de/reyelt/ .

[17] An HTML version of the training materials containing audio (.au) and graphics (.gif) is available at: http://ling.ohiostate.edu/Phonetics/J_ToBI/jtobi_homepage.html . From here there is a link to an ftp site containing a postscript version of the Guide, audio files in ESPS and SUN .au format, and eps, .gif, and .ps files of F_0 track, waveform, and labels. A hard copy is also available (Venditti 1995).

[18] Information about the standard and a .ps version of the training materials (Benzmüller and Grice, 1997) is available at the following address: http://www.coli.uni-sb.de/phonetik/projects/Tobi/gtobi.html .

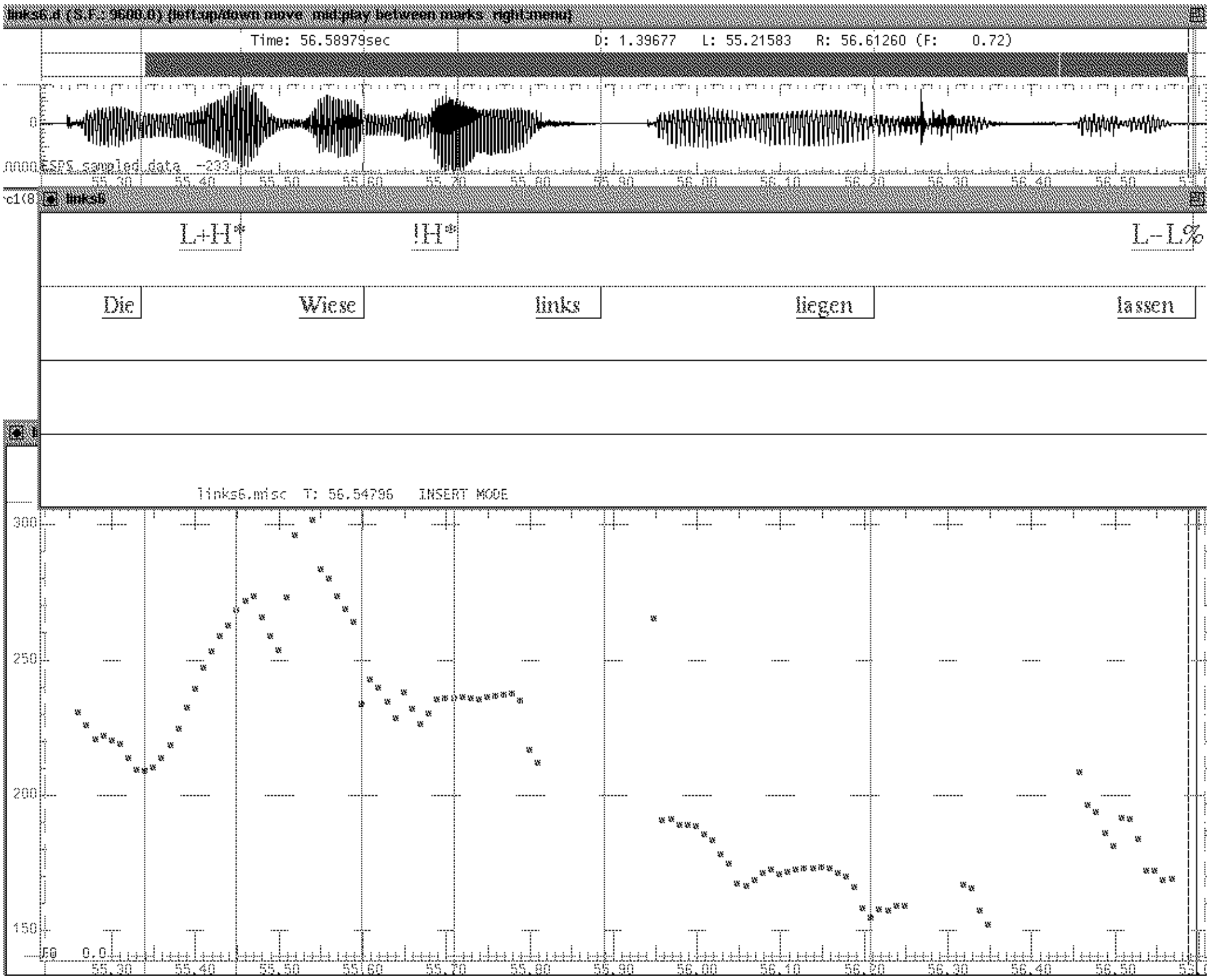

Figure 1: An example of GToBI transcription, time-aligned with an F_0-track.

The training materials introduce basic pitch accents and edge tones along with tonal modifications such as upstep and downstep. For training purposes schematic diagrams and lists of important criteria for each category are provided, along with pointers to speech files containing canonical examples. The speech signal files, available in headerless binary Unix and ESPS formats are available on demand at the address on the page. More on the GToBI system can be found in Gibbon et al. (1998).

Inter-transcriber agreement ratings are reported in Reyelt et al. (1996) and Grice et al. (1996). Results show that GToBI is already adequate for large-scale database annotation with labellers of differing expertise at multiple sites.

In addition to the existence of ToBI systems for Japanese and German, an adaptation to the English ToBI has been made for the transcription of western Scottish (Glasgow) English, GlaToBI, (Mayo et al., 1997). Although no training materials are available, the system has been used in cross-transcriber consistency tests. The adaptations made include an L*H accent, representing a rise (rather than, say, a L valley as in L*+H) which is aligned with the accented syllable, and the elimination of automatic upstep of boundary tones after a H- intermediate phrase tone. In GlaToBI, H-L% represents a fall, rather than a level stretch as in E_ToBI.

It has been argued (Nolan and Grabe, 1997) that ToBI, by which E_ToBI is meant, is too phonological for the comparison of dialects of English. This is to be expected, since it was not designed to do this. The adaptation necessary for GlaToBI illustrates this point.

Autosegmental-metrical analyses have been carried out to a greater or lesser degree in a great many languages. These are, amongst others, Dutch (Gussenhoven, 1984/1993; Gussenhoven and Rietveld, 1991), Bengali (Hayes and Lahiri, 1991), American Spanish (Sosa, 1991), Greek (Mennen and den Os, 1993; Arvaniti, 1994), Italian (Grice, 1995; Avesani, 1990; D'Imperio, 1997), French (Post, 1993), European Portuguese (Frota, 1995).

3.5.2 TSM - Tonetic stress marks

The *tonetic stress marks system*, as used for the transcription of the SEC corpus (see Knowles et al., 1996: 51-57 for a critical account) is based on the British school style of auditory intonation analysis. The TSM transcription system has two levels of intonation phrasing: the *major tone group*, the end of which is marked with a double bar, '‖', and the *minor tone group*, the end of which is marked with a single bar '|'. The TSM system indicates the presence and tonal characteristics of every accent by means of a diacritic before the accented syllable.

There is no internal structure to the major or minor tone groups, except that they must contain at least one accented syllable. The tones in the TSM inventory are:

(1) level

(2) fall

(3) rise

(4) fall-rise

(5) rise-fall,

each of which may be *high* or *low*, where *high* means that the starting point of the tone is higher than the previous pitch and *low* that the starting point is lower.

If an accented syllable is final in a tone group, marking it with a given tone determines the pitch from the beginning of that syllable up to the tone group boundary. The domain includes all syllables up to but not including the next accented syllable or end of tone group.

The corpus which has been auditorily transcribed using this method is the Lancaster IBM Spoken English Corpus (SEC), which has been digitised and is now also available as the MARSEC (MAchine Readable Spoken English Corpus)[19]. The original SEC, transcribed by Briony Williams and Gerry Knowles, was completed in 1987 and comprises five different versions:

(1) Spoken recording

(2) Unpunctuated transcriptions

(3) Orthographic transcriptions

(4) Prosodic version

(5) Grammatically tagged version

The MARSEC, developed by Peter Roach, Simon Arnfield and Gerry Knowles, contains a time-aligned version of the original corpus including annotations. Most of the files are in Entropics/waves+^{TM} format although there also versions of the original .sig files in PC format, which can be converted to ESPS format by means of a shell script.

The British school type of analysis, using TSM at least for nuclear tones, has successfully been adapted to a number of languages. However, it is not, as far as the authors of this document are aware, currently being used for database annotation in any of these.

[19]Available from http://midwich.reading.ac.uk/research/speechlab/marsec/marsec.html .

3.5.3 Conversion between ToBI and the TSM system

An attempt to devise a system for automatically converting nuclear intonation contours in the TSM transcription into E_ToBI has been made by Roach (1994). Although ToBI and the TSM system both have two levels of phrasing, these two levels do not map onto each other in a straightforward way. The minor tone group corresponds to the intonation phrase. There is no equivalent in ToBI of the major tone group. Furthermore, there is no equivalent in the TSM system of intermediate phrase boundary marking. However, as a first approximation, Roach suggests placing an intermediate phrase boundary after every kinetic (i.e. non-level) tone.

Of note is that the conversion uses only a subset of ToBI tones: those with the starred tone in initial position (i.e. H*, L*, and L*+H). This is because nuclear tones in the British system capture the pitch from the beginning of the accented (nuclear) syllable up to the end of the tone group. This precludes, in Roach's view, the use of leading unstarred tones (L in L+H* for instance) in the conversion.

Roach's conversion table, slightly modified, is as follows:

TSM description	at intermed. boundary	at inton. boundary
low level	(no level tones here)	L* L-L%
high level	(no level tones here)	H* H-L%
rise-fall	L*+H L-	L*+H L-L%
high fall-rise	?	H* !H-H%
high fall	H* L-	H* L-L%
low fall	!H* L-	!H* L-L%
high rise	H* H-	H* H- H%
low rise	L* H-	L* L-H%
low fall-rise	?	!H* L-H%

Table 5: Conversion between TSM and ToBI, according to Roach (1994)

The main problem Roach finds is where fall-rises are transcribed in tone-unit medial position, which was converted into intermediate phrase final position. Here the ToBI system cannot capture the fall rise. It would need a sequence of HLH, and since the final H would have to be the boundary, then the pitch accent would have to be H*+L, an accent which is missing in the English inventory, falls being usually captured by a combination of H* and one or more low phrase tones.

Ladd points out that "it is pointless to attempt to state a complete correspondence" (1986: 82) between Pierrehumbert's analysis (the model upon which ToBI is based) and the British school. However, he does give a table of correspondences which differs from Roach's in a number of respects. Two major differences are as follows.

Ladd gives more than one equivalent for certain British-style nuclear tones as he also makes use of leading unstarred tones. For example, he lists L+H* L-L% as corresponding to a rise-fall, and L*+H L-L% as corresponding to an emphatic version of this tone. Roach on the other hand specifically rejects the possibility of using L+H* L-L% as a rise-fall because "perceptually the effect of rise-fall is of a pitch movement with strong prominence at the onset" (1994: 96).

Roach uses downstepped tones as equivalents of the low versions of the tones. This is understandable, as the definition of the 'low' variants of the tones in the SEC TSM system is that they begin lower than a previous syllable. However, there are problems with this analysis, since in ToBI downstep can only be used on a non-initial H tone in a phrase. This means that a low fall which is the only accent in a phrase, would be converted into !H* L-, which would be ruled out as illegal. Ladd does not use downstepped H tones as equivalents of the beginnings of low nuclear tones. Instead, he takes other options, such as L* L-L% to represent the low fall.

A short look at the differences in the correpondence tables leads to the conclusion that caution must be taken if any conversion is attempted in either direction. However, perhaps the mere fact that correspondences have been sought is an indication that of all the systems described here, the two most compatible are TSM and ToBI.

3.5.4 INTSINT

INTSINT (International Transcription System for Intonation) aims at providing a system for cross-linguistic comparison of prosodic systems. It has been developed by Daniel Hirst, based on a stylization procedure of the F_0 (Fundamental frequency) contour built up from interpolation between points.[20]

Transcription in INTSINT is based on prosodic target points aligned with an orthographic or phonetic transcription. It can be used at different levels of detail, allowing a narrow as well as a broad phonetic transcription. Although it is conceived as a system for cross-language comparisons, language-specific subsets of elements can be recommended.

INTSINT is based on the postulate that "the surface phonological representations of a pitch curve can be assumed to consist of phonetically interpretable symbols which can in turn be derived from a more abstract phonological representation" (Hirst, 1991: 307). The pitch contour - or pitch curve - can be represented as a sequence of pitch target points that can be interpolated by a function. In favour of this approach to the representation of pitch curves, Hirst (1991) quotes evidence from acoustic modelling studies showing that pitch targets account better for the data than pitch changes and from perceptual studies claiming that pitch patterns are predominantly interpreted in terms of pitch levels. INTSINT aims therefore at the symbolization of pitch levels or prosodic target points, each characterising a point in the fundamental frequencyfundamental frequency curve.

The symbolization of prosodic target points is made by means of arrow symbols corresponding to different pitch levels. Higher, Upstepped, Lower, Downstepped or Same are tonal symbols describing relative pitch levels defined in relation to a previous pitch target or to the beginning of an intonation unit. Top or Bottom are tonal symbols describing absolute pitch levels described in relation to the operative range of the intonation unit; Mid is assumed to occur only at the beginning of an intonation unit, and is then considered unmarked.

Hirst, Nicolas & Espesser (1991) have shown that, at least for French, the prosodic targets can be defined with respect to the speaker's F_0 (Fundamental frequencyfundamental frequency) mean - Mid-, to one point fixed at a half-octave interval above the mean - Top - and to one point fixed at a half-octave interval below the mean - Bottom -. The F_0 modelling is carried out automatically by a program called MOMEL (Hirst & Espesser, 1991) that, after F_0 detection, provides the best fit for a sequence of parabolas, dividing the F_0 curve into a microprosodic and a macroprosodic profile. The microprosodic component is caused by the individual segmental elements of the utterance, and the macroprosodic component reflects the intonation patterns produced by the speaker (Hirst & Espesser, 1991). The output of the programme is a sequence of target points with a time value in ms. and a frequency value in Hz. Target points can be then automatically coded into INTSINT symbols, once the position of the intonation unit boundaries has been manually introduced.

An experiment comparing listener's evaluation of a synthesized text using original target points and INTSINT-coded target points has shown that the INTSINT version attained more than 80% of the score attributed to the version synthesized with the original target points (Hirst, Nicolas & Espesser, 1991).

Within the MULTEXT project a tool is planned for the automatic symbolic coding of F_0 target points using INTSINT. A preliminary description of such an algorithm is given in Hirst (1994) (see also Hirst et al., 1994) which attempts to provide an optimal INTSINT coding of a given curve by seeking to minimise the mean squares error of the predicted values from the observed values.

[20] Information on INTSINT can be obtained from http://www.lpl.univ-aix.fr/~hirst/intsint.html .

Absolute pitch values Top, Mid and Bottom are modelled by their mean values and Relative pitch levels are modelled by a linear regression on the preceding target point.

One major difference between INTSINT and other models described so far is that symbols are aligned simply with a point in the signal. In the TSM system, a nuclear tone begins on a stressed syllable and is transcribed immediately before this syllable. In ToBI a tone is marked with a star to signal alignment with the lexical stress of a given word, allowing for the capture of timing differences such as that between L+H* and L*+H where the rise is earlier in the first than the second. ToBI also uses diacritics to signal alignment with a given boundary (although only loosely in the case of intermediate phrase edge tones). In INTSINT, on the other hand, target points are simply coded for their height, of which there are five categories (as opposed to two in the ToBI and TSM systems). Information as to the alignment of the target point with a given constituent can be retrieved, if there is a parallel analysis of the utterance into such constituents. Distinctions regarding the timing of target points in relation to accented syllables (such as L+H* and L*+H above, or *early*, *medial* and *late* peak (see Kohler, 1987)) are not captured in the tonal annotations. Again, actual alignment information is not explicitly coded, but retrievable through the linking up of different levels of annotation, assuming that they are available.

3.5.5 Automatic annotation of prosody in VERBMOBIL

Details of the prosodic annotation employed in the VERBMOBIL project are given in Gibbon et al (1998: 165-168). Verbmobil has two types of manual annotation, KIM and ToBI (as reported on in Reyelt et al., 1996 and Grice et al., 1996). The prosodic labelling system PROLAB, based on the Kiel Intonation Model (KIM), is described in Kohler (1995). The model itself is described in Kohler 1991 and 1996. Here we deal with automatic annotation, which is carried out separately.

Prosodic information is currently being used in the following analysis modules in VERBMOBIL: syntactic analysis, semantic construction, dialogue processing, transfer, and speech synthesis. Clause boundaries, for example, are successfully detected at a rate of 94%.

A word hypotheses graph (WHG) and the speech signal serve as input for the prosodic analysis, which then enriches the WHG with prosodic information based on "the relative duration [...]; features describing F_0 and energy contours like regression coefficients, minima, maxima, and their relative positions; the length of the pause (if any) after and before the word; the speaking rate; [...]" (Niemann et al., 1997b: 2). Probabilities for accent on the word, clause (or sentence) boundaries and sentence mood are computed and used to facilitate syntactic analysis at clause or sentence level, to disambiguate sentence particles like *noch* ('still' vs. 'another') on the semantic level, to segment dialogue acts through the use of prosodic boundaries, to enable transfer from German to English by taking into account the sentence mood, and to 'imitate' the voice of the original speaker in speech synthesis by adapting pitch level and speaking rate.

Based on the results of this kind of prosodic analysis, the number of possible parse trees in the syntactic analysis can be reduced by 96% and processing time sped up by 92%. Below, we give one example each of prosodic disambiguation on the syntactic and the semantic level:

(1a) "*Vielleicht. Am Montag bei mir. Paßt das?*"
"*Maybe. On Monday, at my place. Is that OK?*"

(1b) "*Vielleicht am Montag. Bei mir paßt das.*"
"*Maybe on Monday. That's possible for me.*"
(Niemann et al., 1997b: 2)

(2a) "*Dann müssen wir noch einen <u>Termin</u> ausmachen.*"
"*Then we still have to fix a <u>date</u>.*"

(2b) "*Dann müssen wir <u>noch</u> einen Termin ausmachen.*"
"*Then we have to fix <u>another</u> date.*"
(Niemann et al., 1997b: 3)

In (1), identifying the clause boundaries prosodically helps to delimit the utterances automatically and to classify them according to dialogue acts. In (2), disambiguation of the particle *noch* is achieved by identifying the presence (1b) or absence (1a) of primary stress/accent on it.

3.5.6 Prosodic studies on the ATIS project

Various studies have been conducted on the corpus data collected on the ATIS project in order to determine whether prosodic features can be exploited in the automatic analysis of human-machine interaction. These studies deal partly with finding ways of automatically identifying structural elements of discourse (Wang and Hirschberg, 1992) and partly with developing strategies for identifying and correcting dysfluencies (cf. section 3.3.1) on the basis of prosodic information (Nakatani and Hirschberg, 1994).

Wang and Hirschberg's study on the automatic clasification of intonational phrase boundaries had the explicit aim of detecting in which way structural/prosodic information predicted from the text can serve as a first step towards identifying the structural elements of texts, and determining how this information can be augmented and made more reliable by exploiting observed prosodic information in order to improve speech recognition and synthesis.

In order to achieve this aim, they used *classification and regression tree* (CART) techniques (Riley, 1989) to determine the most salient features, after having first manually annotated the data prosodically according to Pierrehumbert's (1980) model of intonation.

Based on their analysis, they identified a combination of prosodic and (morpho)syntactic features that can be used to detect prosodic boundaries more reliably:

1. Similar length of adjacent (preceding & current) phrases.
2. General length of a phrase. The occurrence of a boundary becomes more likely if a phrase is longer than $2\frac{1}{2}$ seconds, but less likely if the resulting phrase is less than half the length of the preceding phrase.
3. Accentuation, i.e. a boundary is more likely to occur after an accented word.
4. Syntactic constituency, e.g. the relative inviolability of an NP.
5. Word-class. A boundary is less likely to follow after function words other than *to, in* or a conjunction.

However, the results of their analysis also show that the success of automatic detection of phrase boundaries drops when dysfluent uttterances in the data are not 'normalised'. They therefore conclude that dysfluent boundaries cannot be phonologically categorised in the same way as fluent boundaries and may present a problem for automatic analysis. This is especially important as:

> The quality of the ATIS corpus is extremely diverse. Speakers range in fluency from close to isolated-word speech to exceptional fluency. Many utterances contain hesitations and other disfluencies, as well as long pauses (greater than 3 sec. in some cases). (Wang and Hirschberg, 1992: p. 12)

The problem that dysfluent utterances present for speech recognition or, more precisely, *spoken language understanding systems*, is treated in Nakatani and Hirschberg (1994). In their study, they try to identify *repair cues* and how those, in turn, may be used to detect and correct repairs efficiently in order to facilitate the analysis of spontaneous speech. They define *repair* as "...the self-correction of one or more phonemes (up to and including sequences of words) in an utterance." (Nakatani and Hirschberg, 1994: p. 7)

To illustrate how dysfluent speech can cause problems for a speech recognition system, they give the following examples of ill-recognised speech from the ATIS corpus:

> (1) *Actual string*: What is the fare **fro-** on American Airlines fourteen forty three
> *Recognised string*: With fare **four** American Airlines fourteen forty three
> (2) *Actual string*: Show me all **informa-** information about aircraft type, Lockheed L one zero one one
> *Recognised string*: Show meal **of make** information about aircraft flight Lockheed L one zero one one
> (3) ...Delta leaving Boston seventeen twenty one arriving Fort Worth **twenty two** twenty one forty and flight number ... (Nakatani and Hirschberg, 1994: p. 2)

While all three examples here represent the occurrence of *false starts*, examples (2) and (3) present represent dysfluent *speech fragments*, whereas example (3) is clearly different from the first two with respect to the fact that the utterance may be correctly recognised, but is nevertheless not correctly interpretable. Based on the frequent occurrence of fragments in dysfluent speech, they conclude that:

> ...the interruption of a word is a sure sign of repair, and so we expect the that the ability to distinguish word fragments form non-fragments would be a significant aid to repair detection. (Nakatani and Hirschberg, 1994: p. 9)

In order to classify and help to correct the different types of repair, they set up a *Repair Interval Model* (RIM), based on their analysis, using CART techniques. This model distinguishes between three sub-intervals, each interval possibly containing a number of features that may aid in the detection of repairs:

1. **Reparandum Interval**: Covers the lexical material that is to be repaired. May consist of word fragments, unfragmented words that are repeated or even (noun) phrases that are respecified. Fragmentation seems to occur more frequently in content words and most of fragments appear to be one syllable or less in length. Glottalisation may accompany fragmentation and when it does, seems to be distinct from *creaky voice*. One further distinction between fluent phrase boundaries and non- fluent ones is the absence of *final lengthening* in the latter.

2. **Disfluency interval (DI)**: Extends from the Interruption Site (IS) to the point of resumption of fluent speech. Characterised partly by silent, rather than filled pauses which are generally shorter than fluent pauses, whereby they tend to be even shorter for fragment repairs than non-fragment ones. However, pausal duration alone does not appear to be a reliable indicator of repairs and has to be examined in conjunction with other factors, such as a possible increase in F_0 and amplitudes from the last accented syllable of the reparandum to the first syllable of the correcting material and the possible occurrence of matching spectral-time or lexical patterns.

3. **Repair interval**: Contains correcting material.

The implications of their study are that, in order to detect different types of repair, different methods of analysis, such as spectral-time pattern matching, the analysis of pausal duration, the use of phone-based recognisers, etc. might be employed in conjunction with one another in order to improve the detection and subsequent correction of dysfluent utterances.

3.5.7 X-SAMPA and SAMPROSA

X-SAMPA[21] is a computer-compatible version of the International Phonetic Alphabet, including all diacritics, and symbols for prosody and intonation. It is well established in the phonetics and speech

[21] http://www.phon.ucl.ac.uk/home/sampa/x-sampa .

technology fields for the transcription and annotation of phoneme-sized segments. However, one of the main weaknesses of the IPA, and by extension also of its computer-compatible equivalent, is the provision for the transcription of prosody and intonation. The fact that there are many models currently in use, both in basic and applied research, makes standardisation an impossible task. It is not simply a matter of choosing which symbol to use, but rather of choosing which phenomena are to be captured. It is therefore necessary to have a computer-compatible alphabet for prosodic annotation which attempts to cover the breadth of the field.

An attempt to meet this need is SAMPROSA[22], which was designed for application in multi-tier transcription systems. SAMPROSA requires that intonational annotations be transcribed on an independent tier from other transcriptions or representations of the signal. It is argued that symbolic representations on different tiers may be related in two different ways. They may be related through association between prosodic and segmental units such as those on a phone, syllabic or orthographic tier. This is the autosegmental-metrical approach used in the ToBI system, and to some extent in the TSM system. Alternatively, they may be related by synchronisation: "The symbols may be assigned to the signal as tags or annotations; the temporal relations between symbols are then given empirically (extensionally) via their position with respect to the signal" (see footnote on SAMPROSA). This is the approach taken by the INTSINT system.

It is important to point out that neither X-SAMPA not SAMPROSA are transcription systems as such. They are computer-compatible codes for use in transcription, once a model has been selected. Alternatively, they can be used for computer-coding extensions to existing models, leading to improved readability across the different approaches.

3.5.8 Recommendations

It is not possible to make absolute recommendations in the field of prosodic annotation. The ToBI transcription system is to be recommended, if it is to be used for languages or dialects for which there is already a standard. However, it is not to be adopted wholesale for a new language or dialect. Rather it is to be adapted, where possible referring to existing autosegmental-metrical work in the literature. The INTSINT method of annotating intonational phenomena is a method which requires little adaptation for a new language, and can be recommended as an alternative, although the phenomena covered are not the same as those covered by ToBI. Although it was originally designed to be used on a purely auditory basis, the TSM system, as long as it is supported by making recourse to an F_0 track, provides a third, albeit possibly outdated, alternative.

Since the field is rapidly developing, it is advisable that anyone wishing to undertake prosodic annotation consult the links provided in this document before beginning work.

3.6 Pragmatic annotation: functional dialogue annotation

3.6.1 'Historical' background

From a historical perspective, it should be mentioned that since the 1960s, there has developed a considerable body of linguistic research on the communicative structures and components of dialogue. On the one hand, linguistic philosophers such as Austin (1962) and Searle (1969, 1980) developed the concepts 'illocutionary act' and 'speech act', to explore and define the range of functional meanings associated with utterances. On the other hand, in sociolinguistics and discourse analysis, various segmental models of dialogue behaviour have been developed by Sinclair and Coulthard (1975), Ehlich and Rehbein (1975), Stubbs (1983) and Stenström (1994), among others. Such studies have often assumed that dialogues can be exhaustively segmented into units, and that these units can be reliably assigned a particular functional interpretation. Some have assumed that

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

there is a hierarchy of such dialogue acts, analogous to the hierarchy of units (word, phrase, clause, etc) in syntax. These approaches have sometimes influenced the assignment of dialogue acts for automatic speech processing, and provide a foundation for general studies of dialogue analysis.

Another historical influence on dialogue research has been the work of the philosopher H.P. Grice on the understanding of spoken communication in terms of the intentions of the speaker (see Grice, 1969). However, this is probably of little relevance to the applications-oriented R&D which is the focus of this chapter.

In the more immediate context of LE, much of the work on dialogue analysis and annotation has up to now been done by the members of the Discourse Resource Initiative (DRI) and many links can be found on its homepage.[23] The DRI holds annual workshops in an attempt to unify previous and ongoing annotation work in dialogue coding. Out of the first workshop of the DRI, there evolved a coding scheme, called DAMSL (Dialog Act Markup in Several Layers), which served as a basis for annotation of the 'homework' material assigned to participants for the second workshop at Schloß Dagstuhl, Germany [24]. Since then the DAMSL scheme has been revised to incorporate at least some of the suggestions made by the participants of the workshop. [25] Further recommendations, especially with regard to the coding of higher-level discourse structures, are to be expected as the outcome of the third DRI workshop in May 1998 in Chiba, Japan (see Nakatani and Traum, 1998). [26]

The DRI workshops may be seen as 'milestones' in the development of dialogue coding and represent a concerted effort to establish international standards in this field. Most of our recommendations are, at least to a considerable extent, based upon their workshop materials and reports.

3.6.2 Methods of analysis and annotation

The pragmatic annotation of dialogues constitutes a special case. Whereas the coding of all other levels of representation/annotation discussed so far may to an extent be performed independently, ideally pragmatic annotation makes use of information from all other levels.

Within LE projects, two different methods for the segmentation, annotation and analysis of dialogue are employed. Dialogues are segmented and annotated either automatically (VERBMOBIL, TRAINS) or manually using online marking tools (Instructions for Annotating Discourses, TRAINS, HCRC Map Task). None of the projects seem to rely on purely 'manual' annotation schemes, i.e. without the support of any online annotation tools, such as coders or SGML markup tools. Note that the term *segmentation* is sometimes used to refer to either structural or functional units, an ambiguity which is probably best avoided. We use the term unambiguously to refer only to the structural/textual level and not the functional one.

One of the main problems in analysing discourse is to separate form from content, in other words to distinguish the structural from the functional level. Although, for example, a speaker's turn may correspond to only one sentence on the structural/syntactic level, on the functional level it may correspond to more than one speech act or form only one part of a larger functional unit (see Section 3.6.4 for more details). This duality may sometimes lead to confusion if the same term is used to refer to both a structural and a functional unit within the dialogue, e.g. the term *turn* being used synonymously with *speech act*. In the context of this document, 'structural' may be understood as 'utilising information available from the orthographic, syntactic or prosodic levels of representation/annotation'.

[23] http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html

[24] http://www.dag.uni-sb.de/ENG/Seminars/Reports/9706/

[25] http://www.cs.rochester.edu/research/trains/annotation/RevisedManual/RevisedManual.html .

[26] However, at the moment of writing, we have only had very cursory information as to the outcome of this workshop, so that we can only to give very sketchy details in the appropriate sections. We assume that more detailed information will be made available at the DRI website in due course.

3.6.3 Segmentation of dialogues

Before analysing any dialogue according to its functional elements, it is first necessary to segment it into textual units that serve as a basis for its representation and annotation. This may have to be done manually, but in most cases will nowadays be done automatically according to the criteria outlined in Section 3.2. Within the turn (see 3.2.3 above), the most commonly used basic unit for this is the *structural utterance*, which will often, on the syntactic level, correspond to what we called in Section 3.4 a *maximally parsable unit* or *C-unit*. This may correspond to a traditional sentence, or, in many cases, to a single 'stand-alone' word or phrase. Note that some documents on dialogue coding may actually refer to structural utterances as *phrases* (see Nakatani et al., 1995). However, we recommend using the term *structural utterance* as *utterance* is the most commonly used term within the LE community.[27]. However, we think that using it without the attribute *structural* may lead to confusion as the same term is often used to identify **functionally relevant** items as well and therefore propose a two-way distinction between *structural* and *functional utterances*.

In order to segment the turns of a dialogue into individual structural utterances, it seems to be more or less common practice to use mainly syntactic clues or pauses, sometimes supplementing them by making recourse to intonational clues. In fact, assuming that an orthographic transcription has already been undertaken (see Section 3.3), a pre-interpretative segmentation of the text will have been undertaken already, using such clues in the marking of full stops (see 3.2.7.2) or other punctuation marks. In this case, it will be the dialogue act annotator's task to refine those structural utterance units already tentatively identified in the orthographic text representation, splitting or merging such units where necessary.

When prosodic clues are used, they are still in practice usually based upon the transcriber's auditory interpretation and not on actual physical evidence. Two notable exceptions here are the VERBMOBIL project and some of the studies done on the ATIS corpus, which use pattern-matching techniques based on the F_0-contour and other prosodic features to establish structural utterance units, (see 3.5.5 and 3.5.6 above for more detail). Work of a similar kind is being undertaken within the framework of the TRAINS project as well.

Various different techniques are employed to represent structural utterances in the text. Most projects will initially make use of some kind of orthographic transcription as outlined in 3.2 and may later refine it according to more functional criteria. Some researchers prefer to store each functional utterance (no matter how short it may be) on one line by itself, whereas others group utterances according to 'intuitive sentences' and separate individual structural utterances from each other by using such symbols as a forward slash (/) (Condon and Cech, 1995). However, important as the structural analysis may be, it may be seen as no more than a preliminary to functional annotation and great care has to be taken not to overemphasise the importance of structural elements such as line breaks, so that they may inadvertedly be confused as having functional significance.

As already noted, apart from the utterance, there is only one higher-order structural unit, which is generally referred to as *turn* (see 3.2.3). (It is also sometimes referred to as a *segment*; however, the use of the term 'segment' here may be slightly problematic, as it may be confused with segments identified at the phonetic level.) A turn generally comprises the sequence of utterances produced by a single speaker up to the point where another speaker takes over. However, cases of overlap also have to be taken into account. Turns which totally overlap with another turn need to be coded separately since they may have functional significance, for example as expressions of (dis)agreement on the part of the interlocutor. In contrast to the structural utterance discussed immediately above, it is more important to mark turns at the pragmatic level because it is always important to be clear about who is speaking at any given time.

[27] To add to the potential confusion, 'utterance' is sometimes used (e.g. in the TEI encoding of spoken texts) as equivalent to a *turn* (see 3.2.3)

3.6.4 Functional annotation of dialogues

The functional annotation of dialogues, sometimes also referred to as *dialogue act annotation*, is a means of capturing and encoding different levels of discourse structure, and identifying how they relate to one another at the pragmatic level. Previously, there had been some debate as to whether this type of coding should try to capture information about a *speaker's intention* or the pragmatic *effect on the dialogue*, but this issue seems to have been resolved at the third DRI workshop at Chiba, in favour of coding with regard to the latter as a speaker's intention may not always be clear to the coder. Functional annotation plays an increasingly important role in current LE applications such as automatic translation systems, generation of summaries of dialogue content, etc. (see, for example, Alexandersson et al., 1997). Functional annotation will be examined first from the point of view of individual utterances (in Section 3.6.5) and secondly from the point of view of multi-level annotation (in Section 3.6.6).

3.6.5 Utterance tags

To characterise the function of individual utterances, the annotator may apply *utterance tags* that characterise the role of the utterance as a dialogue act. The revised DAMSL manual identifies four different dimensions according to which utterances may be classified: (1) 'Communicative Status', (2) 'Information Level', (3) 'Forward-Communicative-Function' and (4) 'Backward-Communicative-Function'. One additional dimension, that is not included in the DAMSL manual, but was discussed at the Dagstuhl conference, is that of 'Coreference'. This, however, may be regarded as a more general aspect of discourse annotation (including the annotation of written texts) and, as such, is beyond the scope of this document. We thus end up with a general four-way distinction for classifying dialogues (slightly expanded with respect to the DAMSL categories), which is discussed in more detail below.

3.6.5.1 Communicative status Communicative status refers to whether an utterance is intelligible and has been successfully completed. If this is not the case, then the utterance may be tagged as either

(1) *Uninterpretable*,

(2) *Abandoned*

or

(3) *Self-talk*.

3.6.5.2 Information level and status Information level gives an indication of the semantic content of the utterance and how it relates to the task at hand. The revised DAMSL manual offers a four-way distinction between

(1) *Task* ('Doing the task'),

(2) *Task-management* ('Talking about the task'),

(3) *Communication-management* ('Maintaining the communication')

and

(4) *Other* (a dummy category for anything that is relevant, but cannot be categorised according to (1) - (3)).

The members of the Dagstuhl conference, however, decided that a three-way distinction would probably be more practical and proposed two alternative classifications:

a. (1) *Task*, (2) *About-task*, (3) *Non-relevant*

b. (1) *Task*, (2) *Communication*, (3) *Non-relevant*

Information status distinguishes between whether the information contained in an utterance contains *old* or *new* information. This distinction is not included in the DAMSL manual, but was discussed at Dagstuhl, where four alternative schemes where considered:

a. Retain a simple distinction between (1) *old* and (2) *new*,

b. Add a category (3) *irrelevant*,

c. Subdivide *old* into (a) *repetition* (including anaphora), (b) *reformulation* (or paraphrase) and (c) *inference* (to bridge anaphora)

d. Define four categories: (1) *repetition* (2) *reformulation* (3) *inference* and (4) *new*.

3.6.5.3 Forward-looking communicative function Dialogue utterances that may be tagged as having forward-looking communicative function are those utterances that could constrain future beliefs and actions of the interlocutors and thus affect the subsequent discourse.

The four categories of the DAMSL manual are:

(1) *Statement*: e.g. *assert, reassert, other-statement*, etc.,

(2) *Influencing-addressee-future-action*: e.g. *request, question, directive*, etc.,

(3) *Committing-speaker-future-action*: e.g. *offer, commit*, etc.

and

(4) *Other-forward-(looking-)function*: dummy category for fixed, relatively rare functions like *conventional-opening, conventional-closing*, etc.

No particularly noteworthy differences from the DAMSL manual emerged from the Dagstuhl conference, but note that category (4) may possibly be subsumed under information level category (3) *communication-management*.

One issue that has been raised at the Chiba workshop is the role of acknowledgements and dysfluency phenomena (see 3.3.1, 3.4.1 and 3.5.6) with regard to their possible forward-looking functions and how they may be integrated into a coding scheme.

3.6.5.4 Backward-looking communicative function In contrast to those utterances that have a forward-looking communicative function, utterances that relate to previous parts of the discourse may be annotated as backward-looking. The DAMSL categories for this are:

(1) *Agreement*: e.g. *accept, maybe, reject, hold*, etc.,

(2) *Understanding*: e.g. *backchanneling, signal-non-understanding, signal-understanding*, etc.,

(3) *Answer*: generally signals compliance with a request for information,

(4) *Information-relation*: utterances expressing explicitly how an utterance relates to the previous one,

and

(5) *Antecedents*: any utterance may be marked as relating to more than just the preceding one.

3.6.5.5 General remarks on the above categories The two final categories in 3.6.5.3 and 3.6.5.4 above do not seem to be mutually exclusive as there can be some overlap between them, i.e., it is sometimes difficult to decide whether an utterance is completely forward-looking or backward-looking. It might therefore be better to think of them as '**Primarily** Forward-looking (Communicative) Functions' and '**Primarily** Backward-looking (Communicative) Functions'. However, the DAMSL manual does not exclude the possibility of assigning multiple tags for forward- or backward-looking communicative functions and this concept was again reconfirmed at the Chiba workshop. Also, whereas the former two categories *communicative status* and *information level and status* primarily relate to the micro level of dialogue structure, the latter two can be seen as the building blocks for the higher-level structures discussed below.

3.6.6 Levels of functional annotation

In addition to a sequence of individual utterances, it is common to posit a hierarchy of dialogue units of different sizes. In conversational analysis the term *adjacency pair* (Sacks 1967-1972) has been commonly used for a sequence of two dialogue acts by different speakers, the second a response to the first. Similarly, in discourse analysis the term *transaction* has been used for a major unit of dialogue devoted to a high-level task, and the term *exchange* for a smaller interactive unit, not dissimilar to the adjacency pair (see Sinclair and Coulthard, 1975; Stubbs, 1983; see also Gibbon et al., 1998: 568-9 on the application of these concepts to dialogue systems). It has further been proposed that such hierarchical groupings of dialogue acts can be modelled in terms of a *dialogue grammar* (see Gibbon et al., 1998: 185).

Multi-level functional annotation may be undertaken by determining the dialogue function of individual (meaningful) utterances and grouping them according to three different levels, the *micro*, the *meso* and the *macro levels*, although not all researchers make use of such a three-level distinction. These will be discussed in 3.6.6 below.

3.6.6.1 Micro-level annotation Micro-level annotation seeks to identify the minimal meaningful functional units within the dialogue and to determine their functional value for the dialogue by assigning utterance tags to them. The annotation may be performed automatically as in the VERBMOBIL project or - most commonly - manually.

Both in the automatic and manual annotation of functional utterances/dialogue acts, we encounter similar problems, which were discussed in detail at the Dagstuhl & Chiba conferences. They are briefly outlined below and some recommendations as to their solution will be given in Section 3.6.7. [28] Since these problems concern annotation of content rather than of form, we shall refer to them as problems of *functional annotation*. They are related to, yet (at least in principle) distinct from, the problems of syntactic segmentation discussed under 3.4.3).

- *Pragmatic particles, discourse markers and interjections* (e.g. *well, okay, alright*): e.g. Where are these to be treated as utterances in their own right, and where as parts of others? (On 'interjections' used in a broad sense relevant here, see 3.3.2.1).
- *Hesitations*: What role do hesitations have in the delimitation of utterances?
- *Coordinated sentences* (e.g. sentences linked by *and*): When are coordinators like *and* to be regarded as beginning a new utterance?
- *Subordinate sentences* (e.g. *... so; ... because*): The same question arises with subordinated as with coordinated sentences.

[28]Note that the Dagstuhl paper refers to them as problems of *segmentation*, but that, in line with our earlier reservation regarding the term *segment*, we prefer to avoid it here.

- *Reformulations*: e.g. Are they to be treated as constituting a different utterance or dialogue act from the utterances they reformulate?
- *Suggestions and requests for their confirmation*: e.g. Should they be regarded as separate utterances?

Members of the Dagstuhl conference essentially identified the following three types of *functional boundaries*:

(1) *regular utterance-token boundaries* (suggested mark-up: @) correspond to what are referred to as *utterances* above.

(2) *weak utterance-token boundaries* (suggested mark-up: *) are optional sub-units.

(3) *drop-in utterance-token boundaries* (suggested mark-up: $) serve to delimit phenomena such as self-repair and hesitations, which can interrupt other segments and do not have a functional role in relation to what precedes or follows.

However, category (3) is not necessarily to be taken at face value, since *self-repairs* or *hesitations* may actually fulfil functional roles, as pointed out earlier (see Section 3.2.7.1), and may therefore better be included under (1) or (2).

Based upon the above categories, a set of five *annotation rules* was proposed:

(1) Annotate utterances that serve to perform an illocutionary function (@)

(2) When in doubt as to whether to annotate or not, **do not** annotate.

(3) If there are strong indicators, e.g. prosodic boundaries such as a long pause, annotate (@). (Note: but only in cases which are compatible with Rule (1).)

(4) Even when speakers collaborate in the completion of a unit, annotate at locations of speaker change (@).

(5) Optional: Annotate smaller units using weak boundaries (*) where the resulting sub-units serve the same illocutionary function.

In accordance with (3), Nakatani and Traum (1998) recommend treating discourse particles or cue phrases as separate utterance tokens, but note that this may not always be advisable for the former as they can sometimes be difficult to distinguish from other word classes, e.g. German *schon*, which may be used as either a discourse particle or an adverb. Some general remarks on the identification of utterances/dialogue acts are provided in Section 3.6.7 and on the coding of boundaries/utterances in Section 3.6.9.

3.6.6.2 Meso-level annotation Meso-level annotation groups individual functional utterances into higher-order units directly above the micro-level of individual utterances/dialogue acts. There currently seem to exist two slightly different major approaches to treating meso-level structures: those exemplified by the HCRC Map Task Corpus (see Carletta and Taylor, 1996) and by the Draft Coding Manual (see Nakatani and Traum, 1998) [29] that is to serve as a basis for discussion at the third DRI conference.

The HCRC approach starts by identifying specific *initiating* dialogue acts, called *moves*, such as instructions, explanations, etc., taking them as the starting point for *(conversational) games*. Those games, in turn, then encompass all functional utterances up to the point where the purpose specified by the initiating act has either been fulfilled or is abandoned (see Carletta et al., 1995: 3).

[29] http://www.cs.umd.edu/users/traum/DSD/mtman.ps

In contrast to this, the approach suggested by Nakatani and Traum (1998) groups functional utterances according to *Common Ground Units (CGUs)*, which, at a more abstract level, represent all those units that are relevant to developing mutual understanding of the participants. CGUs may be cancelled, modified or corrected in retrospect.

Both schemes are based on initiating elements and responses to them and allow for nesting of games/CGUs within other units continued at a later stage. However, the main difference, and potential danger, in the latter scheme is that it also allows for explicit exclusion of functional utterances like 'self-talk' which are deemed as being irrelevant for the dialogue. We suggest, however, that no such elements be excluded until a later stage of the analysis: elements can always be 'flagged' or tagged as being irrelevant and consequently be ignored, but only when it has been firmly established that they actually **are** irrelevant.

3.6.6.3 Macro-level annotation Macro-level annotation is concerned with identifying higher-order structures immediately below the level of the actual dialogue. In order to illustrate it, we shall be referring to the same two approaches as for meso-level annotation.

After having established games at meso-level, the Map Task approach groups those games into *transactions*, encompassing sub-dialogues that represent the achievement of one major step in the task.

The Nakatani and Traum scheme, again, seeks to capture relations between CGUs at a more abstract level by grouping them into *I-Units*. The 'I' in this term may stand for either '*informational*' or '*intentional*'. However, there seems to have been some controversy at the Chiba workshop as to how to encode CGUs in general and especially as to the usefulness of I-Units.

The VERBMOBIL scheme of functional annotation for negotiative telephone calls (Alexandersson et al., 1997) does not include a meso-level, but has a macro-level consisting of the following phases of the dialogue:

1. H - Hello
2. O - Opening
3. N - Negotiation
4. C - Closing
5. G - Goodbye

This is the canonical ordering of the phases, but some variation is allowed for.

3.6.7 Techniques for identifying dialogue acts or topics

In this section, we give a brief outline of some of the techniques that may be used for the automatic or manual segmentation of dialogues into dialogue acts or the identification of topics. As above, this cannot be an exhaustive account of all the possibilities, as some may depend heavily on the nature of individual tasks.

3.6.7.1 Techniques for identifying dialogue acts As already indicated above (see Section 3.6.3), segmentation of dialogues into individual utterances is mainly performed by looking at a combination of syntactic clues, pauses and intonational information. It is clear that syntax is important, e.g. in signalling questions, but often in dialogues such syntactic clues are absent, and reliance has to be placed on lexical and prosodic information. Below, we shall give a short, very tentative list of items that may signal certain functional categories, and (where possible) point out how they may be disambiguated by taking intonational clues into account. Note that our reference to certain intonational features here only represents a set of rough-and-ready guidelines that may help identifying functional categories, mainly when there is no additional prosodic information available.

(1) Discourse particles/conjunctions/(linking) adverbs:

- *okay, alright, yes, yeah, right, good*, etc. in sentence-initial position,
- *but, however*, etc. as adversatives introducing possible disagreement,
- *well, maybe*, etc. as unclear or expressing doubt or reservation.

(2) Syntactically unmarked questions[30] used as:

- request for information + rise; e.g. *Next Monday?*
- suggestions + rise; e.g. *Tuesday okay?*

(3) Commands & instructions: the former may be indicated by a strong falling intonation and added stress, and the latter will also, in many cases, exhibit a fall.[31]

(4) Negation (in responses): *no, not, don't*, etc. as indicators of possible rejection or disagreement or emphatic agreement, i.e. *not at all*, etc.

(5) Repetition ('uptake') of previous speaker's wording as:

- stating agreement,
- a request for clarification in questions,
- an expression of possible disagreement/incredulity in 'echo' questions with a strong rise.

(6) Backchanneling: *hm, yes, right, I see*, etc. + rise-fall + lengthening.

(7) Openers: *hello, hi*, etc. + (usually) rising intonation.

(8) Closing 'tags': *bye, goodbye*, etc. + fall-rise or fall on monosyllabic words.

While these examples may work for some varieties of English (and possibly for some other European languages as well), one has to bear in mind that they would probably need to be adapted for many other languages and indeed for other accents of English.

Techniques of this kind are used extensively in the VERBMOBIL project, especially with regard to discourse particles and sentence boundaries that are automatically disambiguated prosodically before the actual analysis of dialogue acts is undertaken (see Alexandersson et al., 1997: 71ff and Niemann et al., 1997b;Batliner et al., 1997).[32]

3.6.7.2 Techniques for identifying topics Identification of dialogue acts may depend considerably on recognizing the topic or domain being discussed in a particular part of the dialogue. Techniques for identifying topics, sometimes also referred to as *topic spotting*, rely heavily upon word-spotting, as well as knowledge about both the task and the domain (see Section 2). Once sufficient information is available about individual dialogue acts that may form the building blocks for task-specific interactions (games) and frequently occurring or topic-specific words are identified,

[30]Note that even though questions in RP and many other dialects and languages are generally intuitively assumed to end in a rise, this does not always have to be the case and may depend on the speaker's intentions and status. For further detail, see Knowles (1987).

[31]However, expression, especially of the latter, may be highly dependent on the domain. For example, instructions that take the form of long lists as in the Map Task corpus may well end on a high tone as signals of non-finality.

[32]For more information see our Section 3.5.5 .

a list of 'closed-class' items can be created. Based on this list, the dialogue can be analysed and (probable) functional utterance tags can be assigned automatically.

One possible way of arriving at such a list is creating a concordance of key-words and listing them according to their frequency after having eliminated non-topic-specific high-frequency words like articles, etc. by means of a stop-list. In the domain of travel arrangements, for example, likely candidates for such a topic list are place-names, means of transport, references to dates, time adverbials, etc.

In fully computer-based systems like the VERBMOBIL system, topic spotting may be performed at either the word or the sub-word level (see Niemann et al., 1997 for more detail).

3.6.8 Evaluation of Coding Schemes

In order to assess the validity of coding schemes for dialogue annotation, researchers have in the past looked at inter-rater consistency. However, the notion of being able to evaluate such schemes in this way, without taking the amount of chance agreement into account is being increasingly challenged:

> Research was judged according to whether or not the reader found the explanation plausible. Now, researchers are beginning to require evidence that people besides the authors themselves can understand and make the judgements undelying the research reliably. This is a reasonable requirement because if researchers can't even show that different people can agree about the judgements on which their research is based, then there is no chance of replicating the research results. (Carletta, 1996: p. 1)

Based on experiences in research in experimental psychology, there is now an increasing tendency amongst researchers to try and take the element of chance agreement into account by computing the *Kappa coefficient* (see Carletta, 1996; Flammia and Zue, 1995) for inter-rater agreement:

> The kappa coefficient (K) measures pairwise agreement among a set of coders making category judgements, correcting for expected chance agreement.
>
> $$K = \frac{P(A) - P(E)}{1 - P(E)}$$
>
> where $P\ (A)$ is the proportion of times that the coders agree and $P\ (E)$ is the proportion of times that we would expect them to agree by chance, ... When there is no agreement other than that which would be expected by chance K is zero. When there is total agreement, K is one. (Carletta, 1996: p. 4)

For detailed information on how to compute the K coefficient, see Siegel & Castellan (1988: pp. 284-291).

But even if computing the K coefficient somewhat 'objectifies' determining the validity of individual coding schemes, it still remains difficult to compare the efficiency and reliability of different sets of schemes:

> ..., although kappa addresses many of the problems we have been struggling with as a field, in order to compare K across studies, the underlying assumptions governing the calculation of chance expected agreement still require the units over which coding is performed to be chosen sensibly and comparably. (Carletta, 1996: p. 4)

3.6.9 Annotation tools and general coding recommendations

As already indicated above, most projects in dialogue annotation make use of some form of annotation tool. Below, we shall give a brief list of some of the existing tools, some of which are freely available and can be downloaded from the respective web-sites. One thing that nearly all of them have in common is that they can produce fully, or at least partly, SGML-conformant output files.

(1) Flammia's Nb

- available from: http://sls-www.lcs.mit.edu/flammia/Nb.html
- status: free
- output format: similar to SGML; can be converted with supplied Perl script nb2sgml.pl
- platform(s): Unix and Windows95/NT
- other requirements: TCL/Tk; Perl; platform must support long filenames

(2) dat (TRAINS)

- available from: http://www.cs.rochester.edu/research/trains/annotation/
- status: free
- output format: SGML
- platform(s): Windows and Unix
- other requirements: Perl 5.003, Perl Tk package, Perl FileDialog widget

(3) Python-based tools (Map Task)

- available from: http://www.ltg.ed.ac.uk/software/[33]
- status: free
- output format: SGML
- platform(s): Windows and Unix
- other requirements: Python

(4) Alembic Workbench:

- available from: http://www.mitre.org/cgi-bin/get_alembic/
- status: free
- output format: SGML
- platform(s): SunOS/Solaris, Linux, Windows95/NT, Macintosh [34]
- other requirements: none

(5) LT XML:

- available from: http://www.ltg.ed.ac.uk/software/xml/
- status: free
- output format: XML
- platform(s): Windows95/NT, Linux, various other flavours of UNIX, Macintosh
- other requirements: none

(5) XED:

[33] At the time of writing, the tools were not actually available yet, but supposed to become available in the near future.

[34] Not yet available at the time of writing.

- available from: http://www.cogsci.ed.ac.uk/~ht/xed.html
- status: free for educational/research purposes; commercial licence also available
- output format: XML
- platform(s): Windows95/NT, Solaris
- other requirements: based on LT XML

(6) Speech Analyzer/Speech Manager :

- available from: http://www.jaars.org/icts/software.html
- status: free
- output format: Wave file/MS Access^{TM} database
- platform(s): Windows 3.11/95/NT
- other requirements: none

While items (1)-(4) are graphical user interfaces, LT XML is a set of pre- and post-processing tools for handling XML documents. However, XED can be used to set up and manipulate those documents interactively, although it is more of a text editor than (1)-(4).

Item (6) does not fall into any of the above categories. It differs from the rest of the tools described here in at least two respects. For one thing, it was originally designed as a tool for phonetic analysis, rather than specifically for the annotation of corpora. For the other, it does not actually produce any SGML or XML compatible output, but annotations are first written into the wave file by the Speech Analyzer and may then be extracted by the Speech Manager and stored in a relational database. However, a relational database presents a highly efficient mechanism for storing and analysing data, and it would easily be possible to create SGML or XML annotated output from within the database.

The above selection of available tools shows that nowadays it should be no problem to create annotated dialogue material that is SGML- or even XML-encoded. The major obvious advantage of such an approach is that markup languages make it easy to separate form from content during the annotation. In other words, it should be(come) possible to annotate one's data according to functional criteria and then leave it up to the software to group and display categories according to the requirements of the (research) purpose. One added advantage is that additional items of information can easily be incorporated by making use of hyperlinking facilities. A very good example of how such an approach can be put to good use is the HCRC web-interface to the Map Task Corpus[35], which allows the user to look at individual turns produced by each speaker and to play them back across the web/network.

As far as tools are concerned, though, one thing does remain a problem. Even though some of the tools already allow one to play back parts of dialogues associated with individual utterances, there are still only very few publicly available tools (apart from tools such as the above mentioned Speech Analyzer) that actually allow the transcriber/annotator to look at prosodic information from within the annotation tool. Therefore we still have no way of making use of all the available parameters needed to extract information relevant to the interpretation of the dialogue.

3.6.10 Recommendations

(1) Always try to separate form from content, at least conceptually.

(2) Try to integrate as many levels of analysis as possible, preferably all.

(3) Do not exclude any information at an early stage as it might prove relevant later.

[35] http://www.hcrc.ed.ac.uk/dialogue/public_maptask/

(4) Code your dialogues so that they are exchangeable and allow different users to create different views of them, preferably in SGML or, better yet, XML. XML is on its way to becoming a standard, as already major software producers such as Microsoft and IBM have committed themselves to providing support for it in the future.

References

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Maier, E., Reithinger, N., Schmitz, B. and Siegel, M. (1997). Dialogue Acts in VERBMOBIL-2. VM-Report 204, DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken.

Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog Act Markup in Several Layers.

Allen, J.F., Bradford, W.M., Ringger, E.K. and Sikorshi, T. (1996). A robust system for natural spoken dialogue. In *Proceedings of the Annual Meeting.* Association for Computational Linguistics, pp. 62-70.

Altenberg, B. (1990), Spoken English and the dictionary, in Svartvik, J. (1990) (ed.), *The London-Lund Corpus of Spoken English: Description and Research*, Lund Studies in English 82, Lund: Lund University Press, pp.275-86.

Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R. (1991) The HCRC Map Task Corpus. Language and Speech, 34(4), 351-366.

Arvaniti, A. (1994). Acoustic features of Greek rhythmic structure. in: *Journal of Phonetics* 22: pp. 239-68.

Aston, G. (ed.) (1988), *Negotiating service: Studies in the discourse of bookshop encounters: the Pixi project.* Bologna: CLUEB.

Austin, J. L. (1962). *How to do things with words*. Oxford: Clarendon Press.

Avesani, C. (1990). A contribution to the synthesis of Italian intonation. *Proc ICSLP* 90, vol. 1, pp. 833-836. Kobe, Japan.

Batliner, B., Block, H.U., Kießling, A., Kompe, R., Niemann, H., Nöth, E., Ruland, T. and Schacht, S. (1997). Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries. VM-Report 210. F.-A.-Universität Erlangen-Nürnberg/Siemens AG, München.

Beckman, M. and Ayers Elam, G. (^{3}1997). Guidelines for ToBI Labelling. Ohio State University.

Beckman, M. and Hirschberg, J. (1994). The ToBI Annotation Conventions. Ohio State University.

Benzmüller, R. and Grice, M. (1997). Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI. *Phonus* 3. Institute of Phonetics: University of the Saarland. pp. 9-34.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (eds.), (forthcoming, 1999) *The Longman grammar of spoken and written English.* London: Longman.

Burnard, L. (ed.) (1995), *Users' reference guide for the British National Corpus version 1.0.* Oxford: Oxford University Computing Services

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249-254.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Newlands, A., Doherty-Sneddon, G. and Anderson, A. (1995). HCRC Dialogue Structure Coding Manual. Human Communication Research Centre, 2 Buccleugh Place, Edinburgh EH8 8LW, Scotland.

Carletta, J., Isard, A., Isard, S., Kowtko, J.C., Doherty-Sneddon, G., Anderson, A. (1997), The reliability of a dialogue structure coding scheme, *Computational Linguistics*, 23(1), 13-32.

Carletta, J., Dahlbäck, N., Reithinger, N. and Walker, M. (1997). Standards for Dialogue Coding in Natural Language Processing. Seminar No. 9706, Report No. 167, Schloß Dagstuhl, internationales Begegnungs- und Forschungszentrum für Informatik.

Carletta, J. and Taylor, J. (1996). The SGML representation of the HCRC Map Task Corpus. Human Communication Research Centre, 2 Buccleugh Place, Edinburgh EH8 8LW, Scotland.

Chaudron, C. (1988). *Second Language Classrooms: Research on Teaching and Learning*. Cambridge: Cambridge University Press.

Condon, S. and Cech, C. (1992). Manual for Coding Decision-Making Interactions. Université des Acadiens.

D'Imperio, M. (1997). Narrow focus and focal accent in the Neapolitan variety of Italian. in: *Proc. ESCA Workshop: Intonation: Theory, Models and Applications*. Athens, Greece. pp. 87-90.

Edwards, J. and Lampert, M. D. (eds.) (1993). *Talking data: transcription and coding in discourse research*. Hillsdale, New Jersey: Erlbaum.

Ehlich, K. (ed.) (1994). *Diskursanalyse in Europa*. Frankfurt am Main: Peter Lang.

Ehlich, K. and Rehbein, J. (1975), Zur Konstitution pragmatischer Einheiten in einer Institution: Das Speiserestaurant. In Wunderlich, D. (ed.) *Linguistische Pragmatik*. Frankfurt/M.: Athenäum, 209-254.

Eyes, E.J. (1996). The BNC Treebank: syntactic annotation of a corpus of modern British English, M.A. dissertation, Department of Linguistics and Modern English Language, Lancaster University.

Fischer, K. (1998). *A Cognitive Lexical Pragmatic Approach to the Polysemy of Discourse Particles*. PhD thesis, University of Bielefeld.

Fischer, K. and Brandt-Pook, H. (1998). Automatic disambiguation of discourse particles. *Proceedings of the Workshop on Discourse Relations and Discourse Markers*, COLING-ACL, Montreal.

Fischer, K. (1996). Distributed representation formalisms for discourse particles. In Gibbon (1996), pp. 212-224.

Flammia, G. and Zue, V. Empirical Evaluation of Human Performance and Agreement in Parsing Discourse Constituents in Spoken Dialogue. Proc. *Eurospeech 95, 4th European conference on speech communication and technology*. Madrid, Spain September 1995, Vol. 3., 1965-1968.

Garside, R., Leech, G., and McEnery, T. (eds). (1997), *Corpus annotation: Linguistic information from computer text corpora*. London: Longman.

Gibbon, D. (ed.) (1996). *Natural Language Processing and Speech Technology*. Berlin: Mouton de Gruyter.

Gibbon, D., Moore, R. and Winski, R. (1998). *Handbook of standards and resources for spoken language systems*. Berlin: Mouton de Gruyter Paperback in 4 vols; Hardback in 1 vol.

Greenbaum, S. (ed.) (1996), *English worldwide: the International Corpus of English*. Oxford: Clarendon Press.

Greenbaum, S. and Ni, Y. (1996). About the ICE tagset, in Greenbaum (1996), pp.92-109.

Grice, M. (1995). *The intonation of interrogation of Palermo Italian: implications for intonation theory*. Tübingen: Niemeyer.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J. and Batliner, A. (1996). Consistency in Transcription and Labelling of German Intonation with GToBI. *Conference on Spoken Language Processing, Philadelphia*. pp. 1716-1719.

Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents*. Dordrecht: Foris.

Gussenhoven, C. (1993). The Dutch foot and the chanted call. *Journal of Linguistics* 21: pp. 37-63.

Gussenhoven, C. and Rietveld, T. (1991). An experimental evaluation of two nuclear-tone taxonomies. *Linguistics* 29: pp. 423-49.

Hart, H. (1978). *Hart's rules for compositors and readers at the University Press Oxford*. 38th revised edition. Oxford: Oxford University Press.

Hayes, B. and Lahiri, A. (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory* 9: pp. 47-96.

Hirschberg/Nakatani. 1994. A Corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, (3) 1995. p. 1603-1616.

Hirst, D.J. (1991). Intonation models: Towards a third generation. in: *Actes du XIIème Congrès International des Sciences Phonétiques. 19-24 août 1991, Aix-en-Provence, France.* Aix-en-Provence: Université de Provence, Service des Publications. Vol. 1 pp. 305-310.

Hirst, D.J. and Di Cristo, A. (eds.) (forthcoming). *Intonation Systems. A Survey of 20 Languages*. Cambridge: CUP.

Hirst, D.J. and Di Cristo, A. (forthcoming). A survey of intonation systems. in: Hirst, D. and Di Cristo, A. (eds.) *Intonation Systems. A Survey of Twenty Languages*. Cambridge: CUP.

Hirst, D.J., Di Cristo, A., Le Besnerais, M., Najim, Z., Nicolas, P. and Roméas, P. (1993). Multilingual modelling of intonation patterns. in: House, D. and Touati, P. (eds.). *Proceedings of an ESCA Workshop on Prosody. September 27-29, 1993, Lund, Sweden.* Lund University Department of Linguistics and Phonetics, Working Papers 41. pp. 204-207.

Hirst, D.J. and Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix* 15: 71-85.

Hirst, D.J., Ide, N. and Véronis, J. (1994). Coding fundamental frequencyfundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project. *Proceedings of the ESCA/IEEE Workshop on Speech Synthesis, New York, September 1994.*

Hirst, D.J., Nicolas, P. and Espesser, R. (1991). Coding the F0 of a continuous text in French: An experimental approach. in: *Actes du XIIème Congrès International des Sciences Phonétiques. 19-24 août 1991, Aix-en-Provence, France.* Aix-en-Provence: Universit de Provence, Service des Publications. Vol. 5 pp. 234-237.

Hymes, D. (1972/1986). Models of the interaction of language and social life. In Gumperz, J.J. and Hymes, D. *Directions in sociolinguistics: The ethnography of communication.* (Originally published by Holt, Rinehart and Winston, 1972) Oxford: Blackwell, 1986, pp. 35-71.

Ide, N., Priest-Dorman, G. and Véronis, J. (1996) *EAGLES recommendations on corpus encoding*. EAGLES Document EAG–TCWG–CES/R–F. Version 1.4, October, 1996.

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., Quantz, J. (1995). Dialogue Acts in VERBMOBIL. VM-Report 65, DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken

Jekat, S., Tappe, H., Gerlach, H., Schöllhammer, T. (1997) Dialogue Interpreting: Data and Analysis. VM-Report 189, University of Hamburg.

Johansson, S. (1995), The approach of the Text Encoding Initiative to the encoding of spoken discourse, in: Leech, G., Myers, G. and Thomas, J. (eds.) (1995), *Spoken English on computer: Transcription, mark-up and application.* London and New York: Longman, pp. 82-98.

Johansson, S. et al. (1991). Text Encoding Initiative, Spoken Text Work Group: Working paper on spoken texts (October 1991). Manuscript.

Karlsson, F., Voutilainen, A., Heikkilä, J. and Anttila, A. (1995) (eds) *Constraint Grammar, a language-independent system for parsing unconstrained text*. Berlin and New York: Mouton de Gruyter

Knowles, G. (1987). *Patterns of Spoken English*. London: Longman.

Knowles, G. (1991). Prosodic labelling: the problem of tone group boundaries. In: S. Johansson and A.-B. Stenström (eds.). *English computer corpora: selected papers and research guide*. Berlin: Mouton de Gruyter, pp. 149–63.

Knowles, G., Wichmann, A. and Alderson, P. (1996). *Working with Speech: Perspectives on research into the Lancaster/IBM Spoken English Corpus*. London and New York: Longman.

Kohler, K. (1987). Categorical Pitch Perception. in: *Proc. IX ICphS*, Tallin. Vol. 5, pp. 331-333.

Kohler, K. (ed.). (1991): Studies in German Intonation. *Arbeitsberichte* nr 25, Universität Kiel.

Kohler, K. (1995). PROLAB - the Kiel system of prosodic labelling. in: *Proc. ICPhS 95*. Stockholm, pp 162-165.

Kohler, K. (1996). Parametric Control of Prosodic Variables by Symbolic Input in TTS synthesis. in: van Santen et al. (eds.) *Progress in Speech Synthesis*. New York: Springer, pp. 459-475.

Ladd, D. R. (1996). *Intonational Phonology.* Cambridge: CUP.

Leech, G. and Garside, R. (1991), Running a grammar factory: the production of syntactically analysed corpora or 'treebanks'. In: Johansson, S. and Stenström, A.-B. (eds) (1991), *English computer corpora: Selected readings and research guide*, Berlin and New York: Mouton de Gruyter, pp.15-32.

Leech, G. and Wilson, A. (1994/1996). EAGLES Morphosyntactic annotation. EAGLES Report EAGCSG/IR-T3. 1. Pisa: Istituto di Linguistica Computazionale, 1994. Reissued (Version of Mar. 1996) as: *Recommendations for the morphosyntactic annotation of corpora.* EAGLES Document EAG-TCWG-MAC/R.

Leech, G., Myers, G. and Thomas, J. (eds.) (1995), *Spoken English on computer: Transcription, mark-up and application.* London and New York: Longman.

Leech, G., Barnett, R. and Kahrel, P. (1996). Guidelines for the standardization of syntactic annotation of corpora. EAGLES Document EAG-TCWG-SASG/1.8.

Levinson, S. (1979). Activity types and language, *Linguistics* 17.5/6, pp. 356-99.

Llisterri, J. (1996). EAGLES preliminary recommendations on spoken texts. EAGLES document EAG–TCWG–SPT/P.

MacWhinney, (1991), *The CHILDES project: tools for analyzing talk.* Hillsdale, NJ: Lawrence Erlbaum.

Marcos-Marín, F., Ballester, A. and Santamaría, C. (1993). Transcription conventions used for the Corpus of Spoken Contemporary Spanish. *Literary and Linguistic Computing* 8:4, 283-92.

Mayo, C., Aylett, M. and Ladd, D.R. (1997). Prosodic Transcription of Glasgow English: an Evaluation Study of GlaToBI. in: *Proc. ESCA Workshop on Intonation: Theory, Models and Applications. Athens, Greece, September 18-20.*

Mennen, I. and den Os, E. (1993). Intonation of Modern Greek sentences. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam* 17: pp. 111-28.

Monachini, M. (1995), ELM-IT: An Italian incarnation of the EAGLES-TS. Definition of lexicon specification and guidelines. Pisa: Istituto di Linguistica Computazionale.

Nakatani, C.J., Grosz, B.J., Ahn, D.D. and Hirschberg, J. (1995). Instructions for Annotating Discourses. Cambridge, MA: Center for Research in Computing Technology, Harvard University.

Nakatani, C.J., Grosz, B.J., and Hirschberg, J. (1995). Discourse Structure in Spoken Language: Studies on Speech Corpora. *AAAI-95 Spring Symposion on Empirical Methods in Discourse Interpretation and Generation.*

Nakatani, C.J. and Hirschberg, J. (1994). A Corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 95 (3), pp. 1603-1616.

Nakatani, C. and Traum, D. (1998). Draft: Discourse Structure Coding Manual. http://www.cs.umd.edu/users/traum/DSD/ntman.ps .

Nelson, G. (1996). Markup systems, in: Greenbaum (1996), pp.36-53.

Niemann, H., Nöth, E., Harbeck, S. and Warnke, V. (1997). Topic Spotting Using Subword Units. VM-Report 205. F.-A.-Universität Erlangen-Nürnberg.

Niemann, H., Nöth, E., Kießling, A., Kompe, R. and Batliner, A. (1997). Prosodic Processing and its Use in Verbmobil. München: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp. 75-78.

Nolan, F. and Grabe, E. (1997). Can 'ToBI' Transcribe Intonational Variation in British English? In Botinis, Kouroupetroglou and Carayiannis (eds), *Intonation: Theory, Models and Applications.* Proceedings of the ESCA Workshop, Athens, Greece.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. PhD thesis: MIT (published 1988 by Indiana Univerity Linguistics Club).

Pino, M. (1997) Transcripción, codificación y almacenamiento de los textos orales del corpus CREA. Versión 1.2. Internal Report. Madrid: Instituto de Lexicografa, Real Academia Española.

Post, B. (1993). A phonological analysis of French intonation. MA thesis: University of Nijmegen.

Riley, M. (1989). Some applications of tree-based modelling to speech and language. in: *Proceedings of the Speech and Natural Language Workshop*, Cape Cod MA. DARPA, Morgan Kaufmann.

Reyelt, M., Grice, M., Benzmüller, R., Mayer, J. and Batliner, A. (1996). Prosodische Etikettierung des Deutschen mit ToBI. in: Gibbon, D. (ed.). *Natural Language and Speech Technology: Results of the third KONVENS conference, Bielefeld.* Berlin: Mouton de Gruyter. pp. 144-155.

Roach, P. (1994). Conversion between prosodic transcription systems: 'Standard British' and ToBI. *Speech Communication* 15, 91-99.

Sampson, G. (1987) Probabilistic models of analysis, in: Garside, R., Leech, G. and Sampson, G. (eds.) *The computational analysis of English*. London: Longman. pp.16-29.

Sampson, G. (1995). *English for the computer*. Oxford: Clarendon Press.

Sampson, G. (1997). Web pages on the CHRISTINE project. http://www.cogs.susx.ac.uk/users/geoffs/RChristine.html

Sacks, H. (1967-72). *Unpublished Lecture Notes*. University of California.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: CUP.

Searle, J. R. (1980). *Expression and meaning*. Cambridge: CUP.

Siegel, S. and Castellan, J. Jr. (^{2}1988). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992). ToBI: a standard for labeling English prosody. *Proceedings of the Second international Conference on Spoken Language Processing* 2. Banff, Canada. pp. 867-70.

Sinclair, J. McH. and Couthard, R. M. (1975). *Towards an analysis of discourse*. Oxford: OUP.

Sperberg-McQueen, C.M. and Burnard, L. (1994). Guidelines for text encoding and interchange (TEI P3). Chicago and Oxford: ACH-ACL-ALLC Text Encoding Initiative.

Stenström, A.-B. (1990), Lexical items peculiar to spoken discourse, in: Svartvik, J. (ed.), *The London-Lund Corpus: Description and Research*, Lund Studies in English 82, Lund: Lund University Press, pp.137-76.

Stenström, A.-B. (1994). *An introduction to spoken interaction*. London: Longman.

Stubbs, M. (1983). *Discourse analysis: The sociolinguistic analysis of natural language*. Oxford: Blackwell.

Svartvik, J. and Eeg-Olofsson, M. (1982), Tagging the London-Lund Corpus of Spoken English. In Johansson, S. (ed.). *Computer corpora in English language research*, Bergen: Norwegian Computer Centre for the Humanities, pp. 85-109.

Teufel, S. (1996). EAGLES specifications for English morphosyntax. Draft Version. [ELM-EN] University of Stuttgart. ftp://ftp.ims.uni-stuttgart.de/pub/eagles/

Teufel, S. and Stöckert, C. (1996). EAGLES specifications for German morphosyntax. [ELM-DE] University of Stuttgart. ftp://ftp.ims.uni-stuttgart.de/pub/eagles/

Thompson, H. (1997). Towards a base architecture for spoken language transcript{s,tion}. COCOSDA meeting, Rhodes.

Thompson, H., Anderson, A. and Bader, M. (1995), Publishing a spoken and written corpus on CD-ROM: the HCRC Map Task experience, in: Leech, G., Myers, G. and Thomas, J., *Spoken English on Computer: Transcription, mark-up and application*, pp. 168-80.

Traum, D. (1996). Coding Schemes for Spoken Dialogue Structure. University of Geneva.

Venditti, J.J. (1995). Japanese ToBI Labelling Guidelines. in: Ainsworth-Darnell, K. and D'Imperio, M. *Ohio State Working Papers in Linguistics* 50, pp. 127-162.

Wang, M. and Hirschberg, J. (1992). Automatic Classification of Intonational Phrase Boundaries. *Computer Speech and Language*, 6 (1992), pp. 175-196.

Wells, J. C. (n.d.). Computer-coding the IPA: a proposed extension of SAMPA. London: UCL.

Wells, J.C., Barry, W., Grice, M., Fourcin, A., and Gibbon, D. (1992). Standard Computer-compatible transcription. Esprit project 2589 (SAM), Doc. no. SAM-UCL-037. London: Phonetics and Linguistics Dept., UCL.

A TEI paralinguistic features

Tempo

- fast
- very fast
- getting faster
- slow
- very slow
- getting slower

Loudness

- loud
- very loud
- getting louder
- soft
- very soft
- getting softer

Pitch range

- high pitch range
- low pitch range
- wide pitch range
- narrow pitch range
- ascending
- descending
- monotonous
- scandent (each successive syllable higher than the last, generally ending in a falling tone)

Tension

- slurred
- lax, a little slurred
- tense
- very precise
- staccato, every stressed syllable doubly stressed
- legato, every stressed syllable more-or-less equally stressed

Rhythm

- beatable rhythm
- arhythmic, particularly halting
- spiky rising, with markedly higher unstressed syllables
- spiky falling, with markedly lower unstressed syllables
- glissando rising, like spiky rising but the unstressed syllables also rise in pitch relative to each other
- glissando falling, like spiky falling but the unstressed syllables also fall in pitch relative to each other

Voice quality

- whisper
- breathy
- husky
- creaky
- falsetto
- resonant
- unvoiced laugh or giggle
- voiced laugh
- tremulous
- sobbing
- yawning
- sighing

B TEI P3 DTD: base tag set for transcribed speech

```
<!-- teispok2.dtd:  written by OddDTD 1994-09-09            -->

<!-- 11:  Base tag set for Transcribed Speech               -->
<!-- Text Encoding Initiative: Guidelines for Electronic    -->
<!-- Text Encoding and Interchange. Document TEI P3, 1994.  -->
<!-- Copyright (c) 1994 ACH, ACL, ALLC. Permission to copy  -->
<!-- in any form is granted, provided this notice is        -->
<!-- included in all copies.                                -->
<!-- These materials may not be altered; modifications to   -->
<!-- these DTDs should be performed as specified in the     -->
<!-- Guidelines in chapter "Modifying the TEI DTD."         -->
<!-- These materials subject to revision. Current versions  -->
<!-- are available from the Text Encoding Initiative.       -->
```

```
<!-- 11.2.7:  Components of Transcribed Speech                -->
<!ENTITY % u 'INCLUDE' >
<![ %u; [
<!ELEMENT %n.u;         - -  ((%phrase | %m.comp.spoken)+)      >
<!ATTLIST %n.u;              %a.global;
                             %a.timed;
          trans              (smooth | latching | overlap |
                             pause)              smooth
          who                IDREF               %INHERITED
          TEIform            CDATA               'u'            >
]]>

<!ENTITY % pause 'INCLUDE' >
<![%pause; [
<!ELEMENT %n.pause;     - O  EMPTY                              >
<!ATTLIST %n.pause;          %a.global;
                            %a.timed;
          type               CDATA               #IMPLIED
          who                IDREF               #IMPLIED
          TEIform            CDATA               'pause'        >
]]>

<!ENTITY % vocal 'INCLUDE' >
<![ %vocal; [
<!ELEMENT %n.vocal;     - O  EMPTY                              >
<!ATTLIST %n.vocal;          %a.global;
                             %a.timed;
          who                IDREF               %INHERITED
          iterated           (y | n | u)         n
          desc               CDATA               #IMPLIED
          TEIform            CDATA               'vocal'        >
]]>

<!ENTITY \% kinesic 'INCLUDE' >
<![ %kinesic; [
<!ELEMENT %n.kinesic;   - O  EMPTY                              >
<!ATTLIST %n.kinesic;        %a.global;
                             %a.timed;
          who                IDREF               %INHERITED
          iterated           (y | n | u)         n
          desc               CDATA               #IMPLIED
          TEIform            CDATA               'kinesic'      >
]]>

<!ENTITY % event 'INCLUDE' >
<![ %event; [
<!ELEMENT %n.event;     - O  EMPTY                              >
<!ATTLIST %n.event;          %a.global;
                             %a.timed;
          who                IDREF               %INHERITED
          iterated           (y | n | u)         n
          desc               CDATA               #IMPLIED
          TEIform            CDATA               'event'        >
```

```
]]>

<!ENTITY % writing 'INCLUDE' >
<![ %writing; [
<!ELEMENT %n.writing;   - -  (%paraContent;)                    >
<!ATTLIST %n.writing;        %a.global;
          who                IDREF               %INHERITED
          type               CDATA               #IMPLIED
          script             IDREF               #IMPLIED
          gradual            (y | n | u)         #IMPLIED
          TEIform            CDATA               'writing'      >
]]>

<!ENTITY % shift 'INCLUDE' >
<![ %shift; [
<!ELEMENT %n.shift;     - O  EMPTY                              >
<!ATTLIST %n.shift;          %a.global;
          who                IDREF               #IMPLIED
          feature            (tempo | loud | pitch | tension |
                             rhythm | voice)     #REQUIRED
          new                CDATA               normal
          TEIform            CDATA               'shift'        >
]]>

<!-- (end of 11.2.7)                                          -->
<!-- The base tag set for transcriptions of speech uses the   -->
<!-- standard default text-structure elements, which are      -->
<!-- embedded here:                                           -->
<![ %TEI.singleBase [
<!ENTITY % TEI.structure.dtd system 'teistr2.dtd'               >
%TEI.structure.dtd;
]]>
<!-- (end of 11)                                              -->
```

C A few relevant web links

WP4 pages at Lancaster:
http://www.ling.lancs.ac.uk/eagles/

EAGLES SLWG telecooperation facilities:
http://coral.lili.uni-bielefeld.de/EAGLES/SLWG/

VERBMOBIL:
http://www.phonetik.uni-muenchen.de/
http://www.phonetik.uni-muenchen.de/VMtrlex2d.html

MAPTASK:
http://www.cogsci.ed.ac.uk/hcrc/wgs/dialogue/dialog/maptask.html

NFS Interactive Systems Grantees' Workshop:
http://www.cse.ogi.edu/CSLU/isgw97/reports.html

CHRISTINE project:
http://www.cogs.susx.ac.uk/users/geoffs/RChristine.html

Discourse Resource Initiative:
http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html

A corpus of Swedish dialogues:
http://www.ida.liu.se/~nlplab/dialogues/corpora.html

TRAINS:
http://www.cs.rochester.edu/research/speech/dialogues.html

Spoken language project at Gothenburg:
http://www.ling.gu.se/~sylvana/SLSA/