EAGLES

Expert Advisory Group on
Language Engineering Standards

EAGLES

Preliminary recommendations on Spoken Texts

EAGLES Document EAG–TCWG–SPT/P

Version of May, 1996

Contents

1 Author

J. Llisterri
Universitat Autonoma de Barcelona
Filologia Espanyola
Bellaterra, Barcelona
Spain
08193
Tel: +34.3.581.12.16
Fax: +34.3.581.16.86
E-mail: joaquim.llisterri@cc.uab.es

2 Introduction

2.1 Scope of the guidelines

The Guidelines presented in this document are concerned with the symbolic representation of speech in written form. The expression *spoken texts* refers in this context to a symbolic representation of speech using conventional spelling enriched with a set of conventions to represent different kinds of information which are present in the spoken form but cannot be conveyed by means of the normal spelling conventions. It is assumed that speech transcribed thus has been previously recorded and that the sound wave is available and related in some way to the symbolic representation.

The transcription of spoken language requires the representation of a series of events that take place simultaneously with the production of speech, especially when speech is unprepared; these events are transcribed by means of different conventions and different tags are used to describe them. The discussion in this chapter will concentrate on the events themselves and will not deal with the tags that can be used for descriptive purposes during the process of annotation.

The present Guidelines are a preliminary answer to the questions put forward by Sinclair (1993:64–65):

- *What features of the sound wave apart from the alphabetic codes should be recommended for documents which are destined to be included in a corpus?*
- *What are the best conventions for representing these issues?*
- *What features of a speech event other than the sound wave is it necessary to encode?*

2.2 Transcription and representation needs in different research communities

Symbolic representations of speech are needed by at least two different scientific communities: on the one hand, the corpus linguistics community, whose main aim is the description of the spoken language from a language-oriented point of view — typical domains of use of spoken corpora within this community are discourse analysis, conversation analysis, sociolinguistics, dialectology, psycholinguistics, child language acquisition, speech pathology, second language acquisition, descriptive linguistics based on large amounts of data or corpus-based lexicography; on the other hand, the speech community comprising those who are interested in the basic processes underlying speech production and perception from a phonetically-oriented point of view and those who are more concerned with the application of this knowledge to speech technology.

Although these two communities have traditionally used different material and have viewed spoken corpora as very different objects due to the different aims of their research, a gradual convergence is taking place such that the same body of data can be fruitfully used in corpus linguistics and in speech work. However, this process is largely subject to the provision that data are collected, transcribed, encoded and annotated taking into account a minimal set of standard requirements. Guidelines concerned with these common standards are then needed, and they should also provide guidance about how to fulfill the more specific requirements of a particular area of research.

2.2.1 The corpus linguistics community

The traditional work in corpus linguistics, when spoken language is addressed, starts with deriving an orthographic transcription from a recording of large stretches of speech. This transcription is afterwards enriched using different annotation systems aiming at reflecting all the important events that take place in the process of speech production — especially when speech is spontaneously produced or an interaction takes place between two or more speakers — and that are not adequately captured by conventional spelling. Furthermore, grammatical information such as parts of speech (tagging) and syntactic structure (parsing) can be added to carry out linguistic descriptive work.

The main aim is to acquire large amounts of data reflecting the natural use of language, therefore emphasis is usually put on the naturalness and spontaneity of the recording, avoiding experimentally controlled situations where the speaker is constrained to utter a number of previously prepared short sequences. Also for this reason, words are transcribed as lexical units and the phonetic details of their realization are not usually taken into account. In certain studies, prosodic information is added in symbolic form, but the systematic use of a phonetic transcription system such as the IPA (International Phonetic Alphabet) is not common in this kind of studies. This also implies that the recorded speech signal is only accessed during the transcription phase and that subsequent work takes place at the level of the symbolic representation.

Corpora collected for the purposes described above and containing orthographic or phonetic / phonemic transcriptions are sometimes called *spoken corpora* (Sinclair, 1994, 1996). A useful definition summarizing its main features is provided by Sinclair:

> *A spoken language corpus is a corpus consisting of recordings of speech which are accessible in computer readable form, and which are transcribed orthographically, or into a recognised phonetic or phonemic notation*
>
> (Sinclair, 1996:28)

2.2.2 The speech community

Within the speech community, the emphasis so far has been on speech databases (Carré, 1992) rather than on spoken corpora in the sense used in the previous section. This is due to the need to obtain controlled speech data for basic research aimed at modelling and describing the articulatory and acoustic properties of speech or, in the field of speech technology, to derive data for speech synthesis or to build up material for training and testing speech recognition, speaker recognition/verification or spoken language dialogue systems (see the chapter on corpus design in the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995) for a review of the applications of spoken language corpora and Lamel - Cole (1996) for a survey of recent activities in the area of speech corpora).

Moore (1991) offers a typology of the types of recorded speech usually encountered in speech research:

- Analytic-diagnostic material designed to get basic information on the articulatory and acoustic features of speech (e.g. lists of consonant-vowel combinations);
- General purpose material for speech technology applications (e.g. vocabularies); and
- Task-specific materials pertaining to different discourse domains and oriented towards the needs of applications in man-machine communication (e.g. train timetable inquiries).

A more detailed account of the linguistic content of these type of corpora is provided in the chapter on corpus design in the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995); the following types of material are distinguished:

- Read aloud isolated items: phonemes, words, sentences or text fragments;
- Semi-spontaneous speech;
- Spontaneous speech about a predetermined subject;
- Simulated person-machine interactions;
- Spontaneous speech.

The central issue here is the speech signal itself, and its symbolic representation is usually made by means of a phonetic alphabet — the IPA or a computer-readable equivalent being the commonly agreed international system (see 5.1.1) — allowing the phonetic modifications of words when they are spoken in context to be represented. The speech wave is first segmented into units that can be related to phonetic symbols and labelled to temporally synchronize a symbol representing a set of phonetic categories with a given part of the signal — a process known as *alignment*; the phonetic representation can be also related with the orthographic representation and thus aligned with the speech signal. Although this process used to be done manually by expert phoneticians, it can be now performed (semi-)automatically, depending on the type of speech; however, manual verification is still needed to achieve the required accuracy of the result.

Corpora with the characteristics described in this section are sometimes called *speech corpora* (Sinclair, 1994, 1996).

2.2.3 Differences

The main differences in the approach to corpora containing spoken materials between the corpus linguistics community and the speech community that we have reviewed so far can be summarized in the following table:

	Corpus linguistics	**Speech research**
Materials	Unprepared, unelicited speech	Controlled, elicited speech
Scope	Discourse, dialogue	Utterance
Recordings	Natural environment	Controlled environment
Transcription	Orthographic enriched (transcription)	Phonetic and orthographic aligned with the speech signal (labelling)
Oriented towards	Symbolic, categorical representation	Speech signal, temporal representation

A discussion of other differences between collections of written and spoken data can be found in the chapter devoted to corpus design in the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995). Seven main differences are outlined there, having to do with the following aspects:

- The durable character of text as opposed to the transient nature of speech, which requires to be recorded in some form in order to be studied.
- The different production times involved in writing and speaking.
- The different nature of the error correction processes in writing and in speaking; while in written texts collections the editing process is not reflected, transcriptions of unprepared speech reflect interruptions, hesitations, repetitions and self-repairs made by the speaker.

- The variations in the spoken versions of orthographically identical word forms as opposed to the invariant nature of their written representation.
- The discrete nature of written text and the continuous character of speech which requires the development of segmentation tools for the later.
- The size and storage requirements for written and spoken corpora
- The categorical information present in the written text and the lack of categorial information in the speech signal.

Biber (1988) and Halliday (1989) contain a more in-depth discussion of differences between speaking and writing from a linguistic perspective.

As discussed in the next section, there has been in very recent times a tendency towards integrating the needs of both communities, especially because the notion of speech database used in speech research has been gradually enlarged to encompass large collections of more natural data that are characteristic of work in corpus linguistics. However, one should not forget the differences due to the historical development of both fields that have led to emphasis on elicited spoken language in the speech research community and to emphasis on unelicited speech in corpus linguistics (Sinclair, 1993:68, 1994, 1996).

2.2.4 Towards convergence

A relatively recent development in speech research has been the study of speech obtained in a more natural situation — although a controlled recording environment has to be kept for the purposes of a detailed acoustic analysis. The first type of speech has been labeled as *spontaneous speech* to differentiate it from the so-called *laboratory speech* — that is, speech material consisting of short read sentences prepared in advanced by the researcher and recorded in laboratory conditions (Lindblom, 1987). To put it in Teubert's words:

> *The speech community has commenced to express their interest in large spoken language corpora. Even general purpose corpora of impromptu, unrehearsed, unscripted, non elicited informal conversations now seem to arouse some interest in speech research as they can be used as test-beds for speech recognition systems.*
>
> (Teubert, 1993:4)

This has triggered research in the field known as *speaking styles* (Eskénazi, 1993) which in some aspects is closely linked to research in pragmatics and in sociolinguistics as well as to typologies of texts developed in corpus linguistics (see Sinclair & Ball, 1995 for a discussion of style and text typology). Some of the speaking styles that are of current interest to speech research are identified by Moore (1991): read speech; spontaneous speech arising from directed monologue; spontaneous speech arising from a dialogue between human interlocutors; spontaneous speech arising from simulated human-computer interaction; spontaneous speech arising from real human-computer interaction; material that reflects the influence of physiological or environmental factors on the voice of the talker; speech collected from talkers representing a large range of age and accent groups; and speech collected from different microphones and microphone arrangements. Although some of the styles mentioned by Moore are still specific to the interests of speech technology — e.g. samples of human-computer interaction — dialogues and monologues have been collected in linguistic work with different aims. It seems, then, that both the materials and the scope of spoken corpora in speech research are being enlarged and tend to be closer to the interests of corpus linguistics.

As far as representation is concerned, Moore (1991:3) points out that:

> *For many purposes (especially in speech technology) it has become clear that speech data can be very useful if accompanied by machine-readable annotations consisting, at the very least, of an orthographic transcription with paragraph or phrase level pointers into the acoustic data.*

The interest in orthographic transcription can be explained by the existence of previously mentioned methods that allow the (semi-)automatic segmentation and labelling of the speech wave and the temporal alignment of the signal with the phonetic and the orthographic representation; thus, the signal in a large corpus can be conveniently accessed through the orthographic transcription. Moreover, speech recognition systems need language models to train the grammars included in the system; since large corpora of orthographically transcribed speech are required to obtain these models, this is another reason for the speech community to be interested in the orthographic representation (Moore, 1991:3; Atwell, 1996).

On the other hand, the value for the corpus linguistics community of the technical advances in the field of digital speech processing and in (semi-)automatic segmentation and labelling of the speech wave, together with the possibility to align it with the orthographic representation has recently been acknowledged. It is important to mention here that a fundamental recommendation issuing from the Network of European Reference Corpora (NERC) is that the digitized speech signal should be included as a component of a corpus (Sinclair, 1993:65–70).

In summary, it can be noticed that at the same time that speech databases are becoming larger and are including more natural data together with their orthographic representation linked to the speech signal, corpus linguistics can take advantage of the technology developed in speech research to automate the process of easily storing and accessing large quantities of spoken data and to obtain a categorical representation in terms of a phonetic transcription or an orthographic representation. There is then a potential for sharing data between both communities that is

> *entirely dependent on the agreement of satisfactory data format, transcription and annotation interchange standards.*
>
> (Moore, 1991:3)

.

2.3 Existing transcription and representation practices for spoken texts

Transcribing and annotating spoken texts is an activity that has been carried out for a long time not only within the corpus linguistics community but also by linguists interested in pragmatics, discourse and conversation. In these Guidelines, two major initiatives will be surveyed: the criteria developed by the Network of European Reference Corpora (NERC) and the Guidelines issued by the Text Encoding Initiative (TEI). Before discussing these, a brief summary of the types of events that are usually represented when a spoken text is transcribed is presented. This is based on examination of transcription conventions found mainly in discourse and conversation analysis (Llisterri, 1994a).

2.3.1 Events represented in the transcription of spoken texts

The list presented here is a first attempt to put together the events that are symbolized in the transcription of spoken language. This is the result of a partial survey of different transcription systems used in conversation and discourse analysis (see Edwards, 1992, 1993, 1995; Edwards & Lampert (Eds.), 1993; Ochs, 1979). The aim of the list is to provide a preliminary set of items in order to start a discussion of the type of events and labels that should be recommended for use in the symbolic transcription of spoken language. However, it should be noted that there is a certain degree of overlap between categories and that this section is to be regarded as a compilation of current practices organized according to the classical levels used in linguistic analysis. The interested reader can find more details in some of the sources for this compilation (Atkinson - Heritage (Eds.), 1984; Blanche-Benveniste et al., 1991; Coulthard & Montgomery (Eds.), 1981; Crowdy, 1994, 1995; Du Bois, 1991; Du Bois et al., 1993; Gumperz - Berenz, 1993; MacWhinney, 1991; Nelson, 1995; Payne,1995; Peppé, 1995; Stenström, 1994; Tusón,1995).

Segmental level: Lengthening produced by the speaker, phonetic quality (when different from the standard one in the language or variant transcribed) and stress are usually marked. Also, segments that do not appear in the speaker's production but would appear in the standard form of the language are sometimes registered.

Syllabic level: Syllable boundaries and syllable lengthening are marked in detailed transcriptions of spoken language.

Word level: Word boundaries, truncated words, non-standard forms, unfamiliar words, onomatopoeic forms, spelt-out words, acronyms and abbreviations are usually marked up in transcription.

Prosodic aspects of the word such as changes in intonation from the beginning to the end of the word and word stress are transcribed in some systems. Auditorily perceived pauses within a word and between words can be also transcribed.

Utterance level: Utterance boundaries and utterance modality can be coded. Also, breaks and interruptions in the utterance — with or without pauses — are signalled in transcription systems. Utterance boundaries can be also marked.

Suprasegmental level: Intonational units: Intonation unit and embedded sub-unit boundaries, incomplete or truncated tone units, resets, junctures or break indices can be marked.

Final contours of terminated and non-terminated tone units are transcribed according to basic patterns (falling, rising, level) and combinations of these.

Pitch: Changes in pitch over the course or part of an utterance, pitch level, range, register and pitch on word and on phrase together with transitional continuities can be part of a prosodic markup

Stress: Stress in any part of an utterance is usually signalled; syllable weight, pitch accent and different levels of stress can also be introduced in the transcription.

Prominence, emphasis or contrastively stressed syllables can be also indicated.

In some transcription systems, indications about the tension and the rhythmic qualities of the utterance are introduced.

Intensity or loudness: Relative or absolute intensity of parts of the utterance can be marked in subjective terms.

Speech rate or tempo: In spoken language transcriptions, speech rate can be timed or untimed. In this latter case, absolute or relative subjective assessments can be found. Irregularities in rhythm are also registered.

Pauses: Both silent and vocalized pauses are usually marked in spoken language transcription. Audible breathings can also be indicated in the transcription. Pauses can be timed or untimed, depending on the accuracy and aims of the transcription.

Paralinguistic events or vocalized semi-lexical or non-lexical phenomena: Semi-lexical – such as *aha*, *erm*, *mm* – and non-lexical vocalized events are represented in transcription, together with their position relative to the lexical stretch. Voice quality and other vocal events – such as shouting or singing – can be also coded in the transcription.

Speaker turns: Speaker turns are signalled in conversation analysis. Speaker identity, the nature of the transition between utterances, its sequential relation, the type of overlapping and latching of speaker's utterances are usually marked.

Contextual comments on the transcription: There should be means to include in the transcription any additionalinformation provided by the researcher: non-vocalized non-communicative phenomena, non-vocalized communicative phenomena (*kinesic* information), information about the type of text being transcribed or background noise.

Transcription difficulties: Researchers have devised ways of codifying and noting the difficulties found in the process of transcribing recorded spoken language. These difficulties may be related to the performance of the speaker or to technical problems in the recordings. Ways to assess the degree of accuracy of the transcription have also been developed.

2.3.2 Transcription conventions adopted by the Network of European Reference Corpora (NERC)

The NERC project has aimed at the definition of a minimal level of textual representation for European corpora, both written and spoken. In his final report on phonetic and prosodic annotation Teubert (1993:2) concludes that:

> *After careful analysis, the NERC consortium has decided to recommend the Transcription Conventions as developed by J.P. French, and in particular the level two transcription rules, for orthographic transcripts.*

French's system (French, 1991, 1992) was mainly used for the transcription of the spoken corpus developed within the COBUILD project, a joint venture between the University of Birmingham and Collins publishers established in 1980 (Sinclair (Ed.), 1987). The transcription system involves four levels that will be described in more detail in 2.5.2. The recommended Level Two is an enhanced orthographic representation that contains basic information about the speaker, turn-taking and non-verbal elements — speaker identity, speaker change, overlaps, laughs, etc. According to French, this is suitable for linguistic studies that do not require intonational information.

Illustrations of the system can be found in French (1991) and Payne (1995) for English, Pisa (1992) for Italian, De Jong (1992) for Dutch, Scheiter (1992a, b) for German, or Villena-Ponsoda (1992, 1994) for Spanish. For a specific treatment of conversational exchanges see Psathas & Anderson (1992).

2.3.3 The Text Encoding Initiative (TEI) recommendations

A chapter of the TEI Guidelines is devoted to the transcription of spoken texts (Sperberg-McQueen & Burnard (Eds.), 1994). It describes the basic structure of the TEI representation of a spoken text — header, text and divisions — and defines ways to signal basic structural elements: contextual information, temporal information, utterances, pauses, semi-lexical and non-lexical vocalized elements, kinesic events, other types of communicative events and text presented in written form to the speaker. Guidelines on segmentation and alignment are also provided, together with recommendations for the transcription of speaker overlaps, word forms, prosody and paralinguistic features — tempo, loudness, pitch range, tension, rhythm and voice quality — and disfluencies. For the representation of phonetic information, use of the International Phonetic Alphabet (IPA) is recommended.

Johansson (1995a, b) provides a clear overview and discussion of the TEI conventions for the encoding of spoken texts.

More information on the TEI can be found at URL http://etext.virginia.edu/TEI.html;http://www-tei.uic.edu/orgs/tei ; http://info.ox.ac.uk/archive/teilite.

An example of a TEI-conformant transcription of a Spanish spoken corpus can be found in Marcos Marín et al. (1993) included on the CD-ROM that has been produced by the ECI (*European Corpus Initiative*) (see more information on this initiative at URL http://www.cogsci.ed.ac.uk/elsnet/eci.html). The British National Corpus (Crowdy, 1995) is a major initiative using TEI-conformant transcriptions for spoken language. More information on this corpus is found at URL http://sable.ox.ac.uk/bnc/index.html or at URL http://info.ox.ac.uk/bnc/.

2.4 Evaluation of the Text Encoding Initiative (TEI) recommendations on symbolic transcription of spoken language

The compatibility between TEI conventions and NERC proposals has been assessed within the framework of the NERC project. Sinclair (1993) mentions "reasonable compatibility" between both proposals and Teubert (1993) reports that French's conventions are "on the whole compatible" with TEI guidelines, while stressing the important fact that NERC's system is more easily interpreted by readers. However, the major effort in comparing both systems has been made in a detailed report written by Payne (1992) to which the reader is referred for a complete assessment of the compatibility between both systems. A summary of Payne's evaluation is presented in Johansson (1995a)

Payne (1992) mentions several general shortcomings of the TEI Guidelines:

- Lack of an explicit analysis of different requirements for different levels of transcription.

  This leads to difficulties in deciding what should and what should not be encoded. NERC proposes different levels that can be used according to the needs of the transcriber, in order to avoid the costs involved in encoding more information than is necessary for a given purpose.

- Lack of balance between the requirements of different types of transcriptions, giving too detailed recommendations in certain areas and showing an under-provision of recommendations in other areas.

  According to Payne (1992) this is due to the fact that *the TEI Guidelines have attempted to foresee what will be required in any given circumstance, and suggested an appropriate specialized tag, attribute or value.*

- Lack of time to develop and modify the guidelines according to experience, contrary to French's proposal.

However, after careful analysis Payne (1992:60) concludes that:

> *The TEI proposals are broadly compatible with current practice in the user community, as represented by the J.P. French conventions. Furthermore, in the majority of cases it will be a straightforward matter to link the machine-friendly TEI codes to more user-friendly encoding systems such as the J.P. French conventions by means of a simple conversion program.*

The idea of having an automatic link between a representation that facilitates the work of the transcriber and increases readability and TEI conventions is also favoured in Sinclair's preface to Payne (1992). However, it is clear from NERC documents that TEI conventions should be followed wherever possible and that recommendations should be made for areas in which TEI Guidelines are not fully adequate.

The analysis of the TEI Guidelines performed by Payne follows the same structure of the TEI document and the results of his work will be briefly summarized here in the same order.

2.4.1 Elements in a text

Header: While, in the TEI Guidelines, contextual information about a text is provided in the *header*, in French's system this kind of information is found in the *text ID* which contains contextual details such as the date, location and manner of the recording as well as information about the participants and the topic

Since the TEI header is more detailed, Payne (1992:12) suggests that: *This is an area in which it may be necessary in the short term to regard the TEI conventions as a target [...] Users will need to exercise their discretion within the context of the long-term aim of full conformity.*

Divisions: Units intermediate between the text and the utterance can be marked as *divisions* in the TEI Guidelines. No explicit provision for this subdivision is found in French's conventions.

Basic structural elements: TEI distinguishes seven *structural elements* in spoken texts: utterances, pauses, vocal, kinesic, events, writing and shifts. They are defined as follows (Sperberg-McQueen - Burnard, Eds., 1994):

Utterance: a stretch of speech usually preceded and followed by a pause or by a change of speaker.

Pause: a perceived pause within or between utterances.

Vocal: a vocalized but not necessarily lexical phenomenon (e.g. voiced pauses).

Kinesic: any communicative phenomenon, not necessarily vocalized (e.g. a gesture).

Event: any phenomenon or occurrence, not necessarily vocalized or communicative (e.g. incidental noises).

Writing: a passage of written text revealed to participants in the course of a spoken text.

Shift: marks the point at which there is a change in some paralinguistic feature.

Each basic structural element is identified by a tag that has various attributes:

Utterances: TEI *utterances* tend to correspond with speaker turns, while French defines *functional sentences*, according to grammatical, semantic and pragmatic criteria. Straightforward procedures for conversion between both representations are suggested by Payne (1992:16ff).

Segments: French's functional sentences can be related to TEI *segments*.

Tone units: According to Payne (1992:28–29), TEI conventions should be enlarged to allow the simultaneous transcription of syntactic and intonation units. NERC conventions for the delimitation of *tone units* can be translated to TEI standards.

Shifts: Changes in paralinguistic features — voice quality, loudness, pitch range and speech rate — can be specified in the TEI conventions by a specific tag for *shifts* with associated feature values. In French's system this will be signalled by means of the *transcriber comments*.

However, it is worth mentioning that in Payne's opinion, the detailed transcription of changes in paralinguistic features might not be useful for a large variety of users: instead, he favours *a way of linking the orthographic transcription to a recording of the particular spoken text, with a more accessible lexicon of descriptive texts* (Payne, 1992:30).

Timing: TEI provides guidelines to define the *temporal relationship* of points in an utterance using different strategies, while French's system uses the tape counter. Conversion between both systems is possible. Payne (1992:32ff) notes that when accurate timing is necessary the alignment should be made between the digitized speech signal and units on the CD-ROM on which speech is stored.

Simultaneous events: In the TEI Guidelines, *simultaneous events* are represented referring to an external alignment map, while in French's conventions only *speaker overlaps* are coded at a detailed level of transcription. Suggestions for improvements in the proposed system can be found in Payne (1992:36–7).

Pause: Mechanisms for conversion between TEI and NERC systems for annotating *pause* can be easily implemented; some suggestions are presented by Payne (1992:17ff)

Vocal: In the TEI Guidelines, *vocal* refers to vocalized non-lexical — burp, click, cough, giggle, laugh, sneeze, sob, yawn — and semi-lexical events; in French's system these would appear as *non-verbal*. A quite straightforward conversion is possible when vocals are produced by the same speaker of the utterance.

Kinesic: *Kinesic* refers to non-vocalized communicative events such as gestures or facial expressions. Since the distinction between vocals and kinesic is not made in French's system, automatic conversion from NERC's adopted system to TEI standards would require a previous classification of French's *non-verbal* elements.

Event: In a TEI transcription, a non-vocalized non-communicative *event* that affects communication — e.g. a sudden noise — can be described with a specific tag. These events would appear in the *transcriber comments* in French's system.

Writing: *Writing* revealed to participants during the course of a spoken text can be marked with a TEI tag, while this fact would be included in the *transcriber comments* in French's system. Payne (1992:26) suggests the need for more flexibility in the TEI system of representing the disclosure of written text.

2.4.2 Transcription practices

Payne's opinion regarding the detail of development of guidelines for transcription practices in the systems under comparison is quite explicit:

> *The TEI Guidelines have relatively little to say about practical transcription questions, while much of French (1992) deals specifically with such problem. It is therefore likely that the J.P. French conventions have most to contribute to the future development of the Guidelines in this area.*
>
> (Payne, 1992:38)

We now wow enter into detail into these issues

Speaker overlap: The representation of speaker overlaps is dealt with in detail by the TEI and by French's conventions. Conversion between both systems can be made in most cases, although Payne (1992:43) remarks on the difficulties due to the need for an external alignment map in TEI transcriptions and suggests the TEI Guidelines could be improved in this area.

Word form and punctuation: According to Payne (1992:44) the section on this topic in the TEI Guidelines *cites a number of problem areas, but makes few definitive proposals, while French (1992) is much more specific.*

Variations in word form: Both the TEI Guidelines and French's system suggest the use of a *list of variants* to ensure consistency in the transcription.

Semi-lexicals: TEI accepts the possibility of dealing with semi-lexical phenomena as *words*, while French provides a *list of acceptable words* for English that could be easily converted to TEI usage.

Spelling conventions: According to Payne, sets of *spelling conventions* for various languages similar to those developed by French for English could be incorporated into the TEI Guidelines.

Non-conventional spellings: French's proposals for anglicized versions of foreign words could be incorporated into the TEI Guidelines and developed for other languages.

Punctuation: In Payne's view, the TEI Guidelines do not contain *a fully developed practice in this area* (Payne, 1992:47), in contrast with French's clear guidance in using punctuation. Payne sees advantages in retaining the normal punctuation conventions with some adaptations and suggests the need for automatic procedures for converting these conventions into a TEI format.

Prosody: Payne (1992:51ff) mentions the lack of development of guidelines for encoding *prosody* in the TEI scheme and discusses some inconsistencies of the statements about prosody in the TEI Guidelines. The favoured solution would be to incorporate basic prosodic information in the orthographic transcription and to use a *fundamental frequency tracing* aligned with the text in cases where a detailed prosodic analysis is needed.

Tone units: Although an easy conversion can be made between French's *boundary markers* and TEI tags delimiting *tone units*, Payne (1992) notes the difficulties of transcribing melodic contours with TEI conventions.

Tonic syllables: TEI Guidelines do not provide an indication of *tonic syllables* as straightforwardly as in French's system. As Payne (1992:55) points out *if the tonic syllable is going to be marked, it should be marked in the orthographic transcription, and the TEI Guidelines should be extended to provide a way of doing this in a straightforward manner.*

Tones: Payne (1992:56) suggests the extension of TEI Guidelines to allow distinguishing *tones* as in French's conventions; such an extension could be based in different specifications for the tag `<syllable>`.

Prominent non-tonic syllables: *Prominent non-tonic syllables* are marked in French's system, but no provision for such feature is found in the TEI Guidelines.

Speech management: TEI has no specific guidelines for the transcription of *disfluency phenomena*, recommending transcription using IPA or other systems of phonemic transcription. On the other hand, French's conventions, adopted by NERC, are much more specific and deal with different phenomena not covered by TEI, such as *guessed* or *unintelligible fragments.*

2.5 Levels of transcription and encoding

The discussion of different needs in the transcription of spoken corpora outlined in 2.2 aimed at emphasizing the point that, although a certain degree of convergence can be achieved, different communities may require different levels of transcription and encoding. Approaches to transcription developed within the speech and the corpus linguistics community will be surveyed, and a proposal for relating these respective systems will be presented in this part of the report.

2.5.1 Levels of transcription in speech-oriented research

For the speech community, transcription of a spoken corpus is linked to the notion of labelling. According to Barry & Fourcin (1992:2):

> *The 'labeling' of a recorded utterance involves the temporal definition and naming of its parts with reference to the acoustic signal. These 'parts' may be temporarily discrete or over-lapping, and may be defined in acoustic, physiological, phonetic or higher level linguistic terms.*

It is clear from this definition that, apart from the orthographic representation, other levels are necessary in this particular domain.

Various levels of labelling have been defined and used in different projects. Two of these will be reviewed here, since they are the basis of the recommendations presented in this document.

Barry & Fourcin (1992) offer a comprehensive system consisting of five levels:

Physical level in which *labels are defined solely with reference to the physically defined events in an utterance* (Barry & Fourcin, 1992:2);

Acoustic-phonetic level based on labels representing well known phonetic categories;

Narrow-phonetic level using phonetic transcription symbols such as the IPA;

Phonemic level in which only distinctive sounds (phonemes) in a given language are represented; and

Broad phonetic level: this constitutes an intermediate level in which only symbols for phonemes are used although these may transcribe non-phonemic continuous speech phenomena.

Moreover, prosodic labels should be added. As Barry & Fourcin (1992:11) point out *prosodic labels can be defined as a separate tier at each level* since different categories of prosodic events can be found at different levels.

Similar levels are found in other proposals, such as the levels of segmentation and labelling discussed by Tillmann & Pompino-Marschall (1993) and used in the German PHONDAT project. Again, five levels of representation are defined: orthographic, canonical word forms, actual word realizations, sound segments and sub-segmental acoustic-phonetic events. The conventions developed within the German VERBMOBIL project (Kohler et. al, 1994; Hess at al., 1995; more information on the project is found at URL http://www.dfki.uni-sb.de/verbmobil/overview-us.html and at URL http://www.ims.uni-stuttgart.de/projekte/verbmobil/index-en.html), and those proposed by Autesserre et al.(1989), together with the work carried out under the SAM project (chapter 5 of Fourcin et al. (Eds.), 1989) should also be mentioned in this context.

The *EAGLES Hanbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995) provides a careful discussion of transcription and labelling levels in the chapter devoted to corpus representation. The following types of transcriptions are described:

Orthographic transcription or transliteration ,in which *the standard spelling of a given language is used to indicate the spoken words.*

Phonemic/phonotypical transcription ,*based on the phonemes of a given language.*

Allophonic transcription ,in which *different symbols are used for a single phoneme when this phoneme occurs in different contexts.*

Phonetic transcription representing *the pronunciation of words of individual speakers.*

Prosodic transcription

Labelling levels are also proposed by the EAGLES Spoken Language Working Group (1995) and are defined as follows:

Orthographic level ,in which *the standard spelling of a given language is used to indicate the spoken words.*

Citation-phonemic level containing *the output phoneme string derived from the orthographic form (by lexical access, by letter-to-sound rules, or both.*

Broad-phonetic or phonotypical level , which *uses only symbols that have the status of phonemes, marking the output of connected speech processes that insert or delete phonemes, or transform one phoneme into another.*

Narrow-phonetic level , which *attempts to represent what the speaker actually said at the time of recording.*

Acoustic-phonetic level , which *distinguishes every portion of speech that is recognisably a separate segment of the acoustic waveform or spectrogram.*

Physical level representing acoustic parameters or articulatory data.

Non-linguistic phenomena including speaker noises, extraneous noises and paralinguistic information.

2.5.2 Levels of transcription in corpus linguistics-oriented research: the NERC proposal

Transcription of a spoken text for corpus-linguistic research can also be made at different levels. We will present here the proposal developed by French (1992) adopted by the NERC consortium. This system consists of four levels, numbered from one to four. Each successive level introduces more detail in the transcription, allowing several levels of detail according to different needs in particular types of research.

Level I consists in the orthographic representation with minimal punctuation and without interactional information, so that change of speakers is not marked. The description of Level I includes conventions for orthographic representations and for punctuation.

Level II is an enhanced orthographic representation with basic information about speaker identity, turn-taking, and non-verbal elements.

Level III contains all the information included in Level II plus extra intonational and interactional information. Tone unit boundaries and tonic syllables are marked, precise identification of overlap onset and resolution are included. According to French (1992), transcription at this level needs to be done by trained phoneticians and a recording of substantial quality is necessary.

Level IV is the most detailed level of transcription. It includes all information present in Level III plus additional intonational codings and acoustic and phonetic information. Tones, head syllables and a phonemic transcription are aligned with a digital representation of the wave form accompanied by a fundamental frequency tracing and a spectrogram of the utterances. A further possibility for this level would include tagging.

2.5.3 Relationship between the two types of transcription: a proposal

It is clear that the categorical levels of transcription presented so far as used by the speech and the corpus linguistic communities are different, except that both systems include an orthographic representation of the spoken text. Labelling of speech is a process that starts with the low-level units and ends at the highest ones, while the transcription of spoken corpora goes in the opposite direction. The suggestion of the Spoken Texts subgroup is that both systems can be related at the lexical level. Since French's Level II recommended by NERC contains words transcribed in orthographic form and the proposed level S2 (see below) in speech transcription consists of a phonemic representation of words in their citation form, it should not be too difficult to relate both types of representations. The level at which words are phonemically transcribed in their citation form can become, as Barry & Fourcin (1992:8) point out, *the 'mediator' between the signal and the lexicon.* The role of lexical databases in the automatic transcription of speech corpora has been explored, among others, by the research group at IRIT in Toulouse (see, for example, de Ginestel et al., 1993), and more information on spoken lexica can be found in the corresponding chapter of the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995).

For the purpose of a symbolic transcription of spoken language, three levels of representation and labelling have been identified within the EAGLES Spoken Texts subgroup:

S1 — Orthographic representation of the text

- This level corresponds to the orthographic or transliteration level also defined by the EAGLES Spoken Language Working Group (1995). Recommendations are presented in 4.1

S2 — Phonemic representation of words in citation form: that is, the forms in which words are pronounced in isolation.

- This is equivalent to the phonemic level of Barry & Fourcin (1992) and corresponds to the citation-phonemic level defined by the EAGLES Spoken Language Working Group (1995). This level could be, as noted before, automatically derived from the orthographic representation if a pronunciation lexicon or a set of letter-to-sound conversion rules is available. A phonotypical or broad phonetic representation reflecting systematic phenomena known to occur in connected speech can be also derived by rule from the citation-phonemic representation. Symbolic transcription systems adequate to this level are discussed in 5

S3 — Phonetic transcription reflecting a discrete symbolic representation of the perceived actual realization of the utterance.

- This corresponds to the narrow phonetic level of Barry & Fourcin (1992) and of the EAGLES Spoken Language Working Group (1995). Symbolic systems for representing this level are also discussed in 5.

Moreover, it should not be forgotten that it is a *fundamental recommendation* (Sinclair, 1993:70) of NERC, also adopted in this document is that a digitized version of every sample of recorded speech be included as a component of a corpus.

2.6 Interface between the transcription and the speech signal

As has been seen in the previous sections, NERC Level IV includes a digitized speech wave together with the symbolic annotation. This will make it possible to use, in corpus linguistics work, the (semi)automatic procedures for aligning orthographic transcriptions with the speech signal, developed in the field of speech technology (see, for example, Andersson & Broman, 1993; Blomberg & Carlson, 1993). However, it must be clear that, at the present stage, automatic alignment systems present certain limitations when applied to spontaneous or conversational speech.

In line with the NERC recommendation, it was suggested during the Madrid workshop 'Issues in Corpus Work' organized by the EAGLES Text Corpora Working Group in January 1996 that an alignment between the speech signal and word end-points would be desirable both for speech and for spoken corpora. Of course, this would require the development of adequate TEI mechanisms.

The work carried out in MARSEC (Roach & Arnfield,1995; Knowles,1995; more information is found at URL http://midwich.reading.ac.uk/research/speechlab/marsec/marsec.html) to link the transcriptions ofthe Spoken English Corpus with the acoustic waveform is a recent example of the conversion of a spoken corpus into a segmented and labelled database.

3 Recommendations for data acquisition

As has been discussed in the introduction to this chapter, data acquisition procedures are essentially different in speech and in corpus linguistics research, due to the different aims of both communities. However, Sinclair (1993:67) points out that:

> *For any level of transcription, a high quality recording improves the efficiency of the transcription process: for anything beyond Level Two the quality must be well above domestic.*

In some cases, it would be practical for corpus linguistics work to follow some of the data acquisition techniques traditionally used by speech scientists. Although sometimes this might be unpractical — field recordings may not allow the use of the standard SAM workstation with its associated software EUROPEC (SAM, 1992; Fourcin et al. (Eds.), 1989) and the environmental conditions required — some benefits might be obtained from the experience acquired in speech research.

The chapter on corpus collection in the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995) contains recomendations concerning procedures for the acquisition of spoken data. A discussion of microphone types and recording techniques and devices leds to the following recommendations that can be also of importance for the collection of spoken corpora to be used in corpus linguistics and are thus summarised here. More details are found in the chapter mentioned above.

- *If acceptable in the recording environment, and for optimal acoustical quality, use headset microphones*
  - The use of headset microphones is recommended in order to avoid problems found with other types of microphones. Close-up microphones attached to the speaker clothes can record noises like the frothing of clothes; table-top microphones, on the other hand, are sensitive to echoes in the room, to eventual tapping on table, movement of papers, and to overlaps in the recordings when more than one speaker is present and the microphones are not properly spaced; finally, room microphones suffer from the interference of surronding noises. However, it has to be pointed out that some speakers might be unconfortable with a head set and other alternatives can be considered if care is taken not to introduce extraneous noises in the recording (see also Sinclair, 1996:29).
  - It is recommended to place the microphone slightly to the left or the right of the mouth and a bit below the lower lip to avoid breathing noises. Cables should not touch the microphone arm, and the speaker should be confortable with the headset
- *Use digital recording devices*
  - This recommendation is based in the fact that analogue speech recordings suffer from a degradation in quality after repeated copies, offer a poor quality in terms of signal-to-noise ratio and are not easy to access when they need to be studied; the recording equipment is also subject to mechanical problems. DAT (Digital Audio Tape) is recommended then as a medium of recording. In a laboratory environment, the use of a computer to make direct recordings on a hard disk is also strongly recommended, although this might be not always feasible in all corpus collection situations; when this is the case, the use of DAT is to be favoured (see also Sinclair, 1996:29)

It is worth reminding that the documentation of the corpus should contain information concerning the recording session – date and time, recording environment –, the microphone – make, type, position –, and the recording equipment used.

Legal issues in data acquisition are not discussed here, and the reader is referred to the chapter on corpus collection of the *EAGLES Handbook on Spoken Systems* (EAGLES Spoken Language Working Group, 1995) for further details. A more extensive presentation of this topic can be found in a booklet edited by the American Dialect Society (1992).

4 Recommendation for a minimal set of encoding for spoken texts

In section 2.3 transcription and representation practices for spoken texts are reviewed, paying special attention to the NERC and TEI proposals. A survey of events represented and encoded in spoken texts (2.3.1) shows that an important number of phenomena can be of interest to different types of research. However, it seems necessary to consider a minimal set of events to be encoded according to the TEI-compliant Corpus Encoding Standard (CES) proposed for EAGLES (Ide, 1996). The present document will only be concerned with the events themselves, and the encoding of the International Phonetic Alphabet, of the transcription, and of the linguistic annotation of speech will be presented as part of CES. Proposals for the encoding of spoken texts within the TEI initiative can also be found in Johansson (1995a, b).

As a starting point, it should be noted that there are important differences between the transcription of read text - when the original written source is available - and the transcription of spontaneous speech. These differences are reviewed in detail in the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995) and can be summarized in the following points:

- The planning process of spontaneous speech is reflected in several types of disfluencies which do not normally occur in read speech, increasing the difficulty of the transcription process and the complexity of the representation. In section 2.3 most of the usually transcribed events related to this fact are presented.
- The criteria to define utterances are not clear in spontaneous speech, neither in monologues nor in conversations.
- In the case of dialogues, interruptions and overlappings still add more complexity to the representation.

Similar problems in the transcription of speech are mentioned by Johansson (1995b), who still adds one more dimension, *i.e.*, the fact that since speech is generally addressed to a limited audience in a private setting, an adequate knowledge of the context and the situation is needed for a correct understanding.

Despite the difficulties involved in the transcription of unprepared speech, it should be possible to define a minimal common set of events to be encoded in the transcription of different types of spoken texts.

In section 2.4 the structural elements considered in the TEI Guidelines have been defined; they are listed again here for the reader's convenience:

- Utterance
- Pause
- Vocal
  - Semi-lexical
  - Non-lexical

- Kinesic
- Event (non-vocalised, non-communicative)
- Writing
- Shift

The *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995) considers a set of non-linguistic phenomena that should be annotated when transcribing a speech corpus:

- Omissions in read text
- Verbal deletions and corrections
- Word fragments
- Unintelligible words
- Hesitations and filled pauses
- Non-speech acoustic events
  - Produced by the speaker
  - Produced by other speakers or environmental noises
- Simultaneous speech
- Speaking turns

A comparison between these recommendations shows that there are elements which are common to both proposals, and therefore, they could possibly be part of the minimal set of elements to be encoded. These elements are the following:

Vocal semi-lexical events

- Included in this category are filled or voiced pauses and hesitations. As will be proposed in the next section, it is convenient to keep a list of standardized spellings for these phenomena, using, when possible, the conventional orthographic forms which appear in reference dictionaries for a given language.

Vocal non-lexical events

- This category includes burps, clicks, smacks, coughs, giggles, laughs, sneezes, sobs, yawns, heavy breathing and all the non-speech acoustic events produced by the speaker. The number of these events can be variable, and a description of the event is used in the annotation.

Non-vocalised non-communicative events

- This includes all the extraneous noises produced by other speakers or those wich result from the recording environment such as doors slaming, telephone ringing, etc. The annotation is, as in the previous category, a written description of the event.

Note that the first two categories correspond to those subsumed under the tag `<vocal>` in the TEI, while the third corresponds to `<event>`.

The transcription of spoken interactions where more than one speaker is involved also requires the consideration of the following elements:

Speaker identity

- In the TEI encoding this information is indicated in the header within the 'profile description' `<profileDesc>` element, which has a 'participant' `<partics>` sub-element containing a series of elements 'person' `<person>`. Among the attributes of `<person>` there is one – named 'id' – coding the identity of the speaker. Within the text, each utterance can have an attribute 'who' with the value corresponding to the identity of the speaker coded in the 'id' attribute (Sperberg-McQueen - Burnard (Eds.), 1994; Johansson, 1995a). Other simplified forms of encoding can be found, but, in any case, this is a necessary element in the transcription of spoken interactions.

Speaking turns, indicating a change of speaker

- Changes of speaker can be coded in the TEI by means of changes in the value of the 'who' attribute, and appear to be the basis for the definition of utterances. Independently of the mechanisms that can be used, this is an essential information in the transcription of conversations.

Simultaneous speech or overlapping

- Proposals for marking this phenomenon are found in the TEI (see 2.4) as part of the strategies for encoding simultaneous events. Although other ways of representing speech overlapping can be found, again this is an important element in the transcriptions of the type of spoken material discussed here.

A third group of elements to be transcribed is related to the performance of the speaker. The convenience to include them in transcriptions is discussed in the *EAGLES Handbook on Spoken Language Systems*, where three different types of phenomena are identified:

Omissions in read text

- Where a written script exists, it might be recommended that the words or segments omitted by the reader should be marked in the transcription as such.

Self-repairs

- In spontaneous speech, the planning process is sometimes evidenced by the presence of self-repair phenomena used by the speaker to correct speech production errors 'on-line' (see Cutler (Ed.), 1982 and Fromkin (Ed.), 1973, 1980 for a psycholinguistic approach to the topic). They might be explicitly indicated by the speaker (using, for example, forms such 'I mean') or they might be implicit; in other cases they might involve restarts or repetitions. Also in read speech it is possible to find corrections of errors detected by the reader himself in the course of the reading. Such phenomena should not be omitted in a transcription.

Word fragments

- Word fragments are one or more sounds belonging to a word which is not fully pronounced by the speaker at a first attempt and are then repeated when the speaker succeeds in producing the complete word. In some systems they are marked by a hyphen (e.g. 'fli-flights), while in others a star is used (e.g. 'fli* flights). It seems also adequate to indicate these hesitations in the transcription.

Moreover, the encoding of spoken texts should contain a documentation of the difficulties encountered during the transcription process. The NERC proposals mention 'guessed' and 'unintellible fragments', while the SpeechDat conventions include a notational device for partially or totally unintelligible words. It seems also adequate to provide means for the notation of the uncertainties of the transcriber:

Unintelligible fragments

- Fragments, words or part of words which are not intelligible to the transcriber should be indicated. A distinction between 'guessed' or 'uncertain' and 'unintelligible' can be made if necessary.

Finally, the encoding of *utterances* – defined as a strecht of speech usually preceded and followed by a pause or by a change of speaker – should be considered. We have already recommended the marking of changes of the speaker, and in section 5.2.2 devoted to prosody it is also proposed that *pauses* should be part of the elements to be encoded. This implies that utterances are necessarily encoded, since they are related to these elements.

An important point which has to be considered is the usability of the TEI recommendations from the point of view of the transcriber. Sinclair (1995) and Chafe (1995) discuss this issue, which is also mentioned by the EAGLES Spoken Language Working Group. As a general rule, a balance between the advantages offered by the TEI, the aims of the corpus and the demands imposed on the transcriber should be sought. The distinction put forward by Sinclair (1995:107) between *conformity* and *compatibility* with TEI is useful in clarifying the debate. In fact, the need to develop conversion software between a user-friendly system of transcription and the TEI encoding scheme was one of the recommendations arising from the EAGLES Workshop on 'Issues in Corpus Work' organized by the Text Corpora Working Group in Madrid in January 1996.

4.1 Recommendations for the orthographic representation of spoken texts

As defined in 2.5.1the orthographic representation of the text corresponds to a representation of the speakers utterances using the standard spelling of a given language (*i.e.*, a transliteration). This level of representation is thus common to spoken and written corpora, and consequently conventions for orthographic representation have been developed both in corpus linguistics and in speech research.

Three representative proposals will be reviewed here and will form the basis of a set of recommendations: the NERC conventions, the SpeechDat guidelines and the EAGLES Spoken Language Working Group recommendations.

Within the tradition of corpus linguistics, the NERC initiative has adopted the conventions for orthographic transcription proposed by French (1992:3ff). They are mainly intended for the transcription of the spoken materials present in the type of reference corpora considered within the project. These recommendations can be summarized for English as follows:

- The words spoken are represented in accordance with standard orthographic conventions;
- The only contractions used are those accepted as standard in the *Oxford English Dictionary*;
- Sentence boundaries are marked by a full stop and capital letter;
- Commas are not used within sentences;
- Direct quoted speech or quotations from written texts are placed in single quotation marks;

- Apostrophes are used in accordance with standard conventions in possessives and in contractions.

These conventions can be compared with the ones developed by Boves & den Os (1995) and adopted for the transcription of the SpeechDat spoken corpora in different languages (more information on SpeechDat can be found at URL http://www.icp.grenet.fr/SpeechDat/home.html). They are based on the ones used by the LDC/ARPA (Linguistic Data Consortium/Advanced Research Projects Agency) for the production of the ATIS (Air Travel Information System) corpus,and are specially conceived for the transcription of a corpus aimed at training and assessing speech recognition systems over the telephone. Other proposals also oriented towards the transcription of speech corpora for phonetic research and speech technology have been developed, for example, within the German VERBMOBIL project (Kohler*et al.*1994; Hess et al., 1995; more information on the project is found at URL http://www.dfki.uni-sb.de/verbmobil/overview-us.html and at URL http://www.ims.uni-stuttgart.de/projekte/verbmobil/index-en.html), for the transcription of the HCRC Map Task corpus (Anderson et al.,1991; and more information at URL http://www.cogsci.ed.ac.uk/elsnet/Resources/Map-Task/mt_corpus.html) or for the transcription of spontaneous spoken dialogues (Fink et al., 1995).

The most relevant SpeechDat conventions for the purpose of the present recommendations are summarised below (Boves & den Os, 1995):

- *Normal lexical items will be represented by their spellings in the normal way.*
  - It is recommended to chose a standard dictionary for each language and to use the spelling forms which appear there. It is also recommended to maintain a lexicon of the spelling forms used in the transcription. This lexicon also contains the forms chosen as the standard for words or expressions which can be spelt in more than one way.
- *It is possible to include, a very restricted number of markings for regular variations in pronunciation*, provided that they are documented and no more than two or three regular variations are indicated.
- *Abbreviations should be represented by their full orthographic forms, unless they are spoken in their abbreviated form.*
  - Exceptions are abbreviations which do not have non-abbreviated forms.
- *Number sequences (flight numbers, times, dates, aircraft types, money amounts, etc.) will be spelled out to reflect what was said*
  - *If digits have alternate pronunciation forms the transcription should accurately reflect the form actually pronounced.*
- *If a speaker pronounces letters, acronyms or abbreviations as a word, for example "British Rail" for BR, then these should be spelled out as words.*
- *No punctuation will be provided in the transcription other than those symbols used for special transcription purposes*

Recommendations for the orthographic representation are also provided in the chapter devoted to corpus representation of the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995). The following conventions are discussed:

- Reduced word forms

  - It is recommended to use reduced word forms as they appear in a standard dictionary.
  - If necessary, other reduced forms not existing in the dictionary can be used
  - The use of reduced forms is recommended if they occur frequently and if they involve syllable deletion

- Dialect forms
  - Dialect forms have to be marked in the transcription
- Numbers
  - Numbers are transliterated as words
- Abbreviations and spelled words
  - Full forms of abbreviations are used in orthographic transcriptions.
  - Abbreviations spoken as words are also transliterated as words
  - Spelling has to be indicated in transcriptions
- Interjectives
  - They should be indicated according to the standard spelling found in the dictionary

The general philosophy behind the proposals put forward by the Spoken Language Working Group is that standard spelling should be used as much as possible and that all non-standard forms used in the transcription should be clearly documented. It is also proposed to generate a list of words and word forms, so that

> *the graphemic forms of the words can be converted to phonemes by means of computerised grapheme-to-phoneme conversion. The result of this is a list of citation forms, also called canonical forms. This forms indicate the pronunciation of words when spoken in isolation.*
>
> (EAGLES Spoken Language Working Group, 1995)

The consistent use of standard spelling forms ensures then the possibility of linking levels S1 and S2 previously described in 2.5.1.

Taking into account this three proposals a set of general recommendations for the orthographic transcription of spoken materials - either read or spontaneous - can be proposed:

- Use conventional spelling forms as they appear in a standard dictionary. This also applies to contractions, reduced word forms, apostrophes, dialect forms, interjections and vocalised semi-lexical events (see 2.3.1)
  - This implies selecting a standard dictionary for each language; in some languages there are dictionaries produced by the relevant normative body (for example the *Diccionario de la Lengua Española* from the *Real Academia Española*), while in others there are dictionaries which are traditionally considered as reference works, such as the *Oxford Dictionary* for English or the *Robert* for French.
- If more than one orthographic form is possible or if non-standard spellings or spelling variations are necessary, maintain a lexicon of the spelling forms used in the transcription

  - The purpose behind this recommendation is to help transcribers to maintain consistency and to provide an accurate documentation. Moreover, if a full list of the spelled forms in created, it is possible to automatically generate the phonemic citation forms of level S2 (see 2.5.1). The creation of a list of the spelling forms used in the transcription in the case of variations in word form, spelling variants and semi-lexical phenomena is also part of the TEI recommendations for transcription practices.

- Represent numbers, abbreviations, acronyms and spelled words in full orthographic form as pronounced by the speaker

  - The aim of this recommendation is to accurately reflect in the transcription the actual utterances of the speaker. Numbers are always transliterated as words, as well as abbreviations and acronyms; however, if one of these later forms is spelled by the speaker, it should then be transcribed as such.

These recommendations are of a very general nature and constitute basic principles to be applied to the transcription of spoken materials. One aspect which would need a more in-depth discussion is punctuation. The NERC proposal suggests to mark sentence boundaries with a full stop and a capital letter and avoids using commas within sentences, while the SpeechDat recommendations suggest not to use punctuation at all. One should be aware that in spontaneous speech the delimitation of units such as sentences is not a trivial matter, since a combination of syntactic, semantic, pragmatic and prosodic criteria is required (see, for example Schuetze-Coburn, 1991), and for this reason introducing punctuation in an orthographic transcription can be sometimes a difficult and controversial activity.

5 Symbolic transcription system

5.1 Segmental level

5.1.1 Transcription systems

Phonetic and phonemic transcription is usually represented by means of the symbols of the International Phonetic Alphabet (IPA). The IPA was revised at the Kiel Convention in 1989 and the most recently revised version has appeared in 1993 (IPA, 1993) (also available at URL http://www.arts.gla.ac. uk/IPA/ipachart.html). The principles on which the IPA is based can be found in the report on the Kiel Convention published in the Journal of the International Phonetic Association (IPA, 1989). Illustrations of the application of the systems with sample transcriptions for several languages regularly appear in the same Journal, and recordings showing the pronunciation of each of the sounds are also available (Wells - House, 1995). The International Phonetic Association has its home page at URL http://www.arts.gla.ac.uk/ IPA/ipa.html.

The International Phonetic Alphabet is not only the most common transcription system used in linguistic research, but is also the standard for representing phonemic and phonetic information recommended by NERC and by the TEI.

However, different traditions have developed different phonetic alphabets more adapted to their respective needs, such as the current American system arising from work in the transcription of American-Indian languages, the system used by European romance philologists engaged in diachronic and dialectological research or the conventions used for scholars working with African, Slavonic or Indian languages. Specific phonetic alphabets are sometimes linked to national traditions, arising from the needs to have an accurate narrow transcription of a specific language. A useful guide to phonetic symbols has been published by Pullum & Ladusaw (1986), where different symbols and usages are explained.

Extensions of the IPA have been also developed for special purposes, in particular for the transcription of disordered speech. The conventions proposed by Ball and his collaborators (Duckworth et al., 1993; Ball et al., 1994; Ball et al., 1996) have been adopted by *Clinical Linguistics & Phonetics*, the official journal of ICPLA, the International Clinical Phonetics and Linguistic Association (the association has its home page at URL http://tpowel.comdis.lsumc.edu/icpla/icpla.htm).

5.1.2 Computer phonetic alphabets

The increasing need for electronic exchange of texts containing phonetic transcriptions has led to the computer coding of the International Phonetic Alphabet (Esling, 1988, 1990; Esling & Gaylord, 1993; IPA, 1989). A numerical equivalent for each of the IPA symbols — IPA number — has been defined and translation tables can be developed to relate ASCII codings to IPA numbers. These mappings are part of the CRIL (Computer Representation of Individual Languages) conventions, also discussed in the chapter on corpus representation of the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995). Worldbet, developed by Hieronymus (1994) is another proposal for the ASCII coding of phonetic symbols and has been used in the 22 Language Telephone Speech Corpus developed by the Oregon Graduate Institute.

Other systems have been developed for specific goals. For example, CHILDES (Child Language Data Exchange System), a project aimed at collecting samples of children's language, makes use of PHONASCII (Allen, 1988), a coding system including a phonemic — UNIBET — and a phonetic alphabet, allowing narrow and broad transcriptions (see more information on CHILDES at URL http://poppy.psy.cmu.edu/childes/).

Within the ESPRIT project *Linguistic analysis of European languages* a Computer Phonetic Alphabet (CPA) was developed for seven European languages, based on the IPA (Kluger-Kruse, 1987).

However, the main effort in the provision of a computer-readable transcription system that covers the phonemic inventories of most European languages has been made within the ESPRIT SAM *Speech Assessment Methodology* projects. The SAM Phonetic Alphabet (SAMPA) defines a set of ASCII codings corresponding to the IPA symbols necessary for the phonemic transcription of all major European languages included in the EUROM corpus (Chan et al., 1995; more information is available at URL http: //www.phon.ucl.ac.uk/resource/eurom.html) and is being successfully used in other European and national projects. SAMPA is described in Wells (1987, 1989), Wells et al.(1992), and is also fully discussed in an appendix to the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group 1995); a presentation of the system and the SAMPA adaptations to Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish and Swedish with ASCII and IPA equivalents can be equally found at URL http://www.phon.ucl.ac.uk/home/sampa/home.htm.

SAMPA, like the IPA, is in principle based on a phonemic principle, representing only sounds which serve to distinguish word meanings in a given language; this is also in accordance with the principle of phonotypical transcriptions discussed in 2.5.1. However, phonetic notation of certain allophones is also allowed with the current set of symbols although it is not encouraged for methodological reasons.

Wells (1995) has recently proposed an extension of SAMPA known as X-SAMPA (described at URL http://www.phon.ucl.ac.uk/home/sampa/home.htm/x-sampa.htm). It consists in a keyboard-compatible coding for the entire set of IPA symbols, including diacritics and tone marks. The system is specially intended for the electronic transmission of materials transcribed using the International Phonetic Alphabet.

In the context of phonetic transcription systems applied to speech technology it is worth mentioning the standards adopted within the ONOMASTICA project (Schmidt et al., 1993) for the transcription of proper names.

5.1.3 Proposals for the transcription of the segmental level

According to Johansson (1995a),

> *The degree of phonetic detail given in speech transcription varies from none to a very precise phonetic or phonemic transcription [...] Where there is a great deal of phonetic or phonemic detail, it will be more convenient to design a specialized writing system.*

The *specialized writing system* recommended by the Text Encoding Initiative is the International Phonetic Alphabet. For the electronic exchange of texts, a machine-readable alphabet has to be used, and in this respect *SAMPA is considered to be a computer version of part of the IPA system* (EAGLES Spoken Language Working Group, 1995). The X-SAMPA extension proposed by Wells provide the parts which were missing from SAMPA, and should then be considered as a system fitted for the purposes of segmental transcription of spoken corpora.

However, possible problems of compatibility between SAMPA and Unicode (more information on Unicode can be found at URL http://www.stonehand.com/unicode/standard.html have been detected in the discussions that took place during the EAGLES Workshop on 'Issues in Corpus Work' (Madrid, January 1996), and this issue should be carefully explored.

5.2 Suprasegmental level

5.2.1 Transcription systems

The process of prosodic encoding can be defined as the symbolization of the linguistically relevant variations that occur in the domains of time, frequency and intensity in the sound wave corresponding to a speaker's utterance. The process of encoding implies deciding which variations in the physical parameters of the speech wave carry out linguistic information and finding a way to describe them by means of a symbolic system. Since physical parameters such as frequency and intensity are continuously varying over time, a symbolic coding implies also converting continuous information to a set of discrete units. Thus, symbolic coding of prosody involves at least two different levels of abstraction: a linguistic interpretation of changes in physical properties of the speech wave, and a classification of these changes into discrete categories. Finally, a notational system has to be designed in order to represent these categories. The review of events transcribed in the tradition of pragmatics, discourse and conversation analysis 2.3.1 has shown that there is a clear need for such a notational symbolic system in these areas.

For detailed surveys of prosodic transcription and encoding systems the reader is referred to Llisterri (1994b) (available at URL http://www.lpl.univ-aix.fr/projects/multext/CES/CES2.html) – from most of the material in this section is taken –, Grnnum Thorsen (1987), Léon & Martin (1970) – which contains a chapter devoted to classical approaches to prosodic transcription – and to Gibbon (1989), reviewing most of the work in this area carried out within the SAM ({em Speech Assessment Methodologies) project. A discussion of this topic is also found in the text representation chapter of the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995).

A great diversity of proposals exist in the field of pragmatics, discourse and conversation analysis, as has been mentioned before. Examples of notational conventions can be found in the literature reviewed in section 2.3.1. All those systems share the fact that the transcription is based in conventional spelling, enriched with some conventions to represent information that is present in the spoken discourse but can not be conveyed by means of normal spelling conventions. Symbols representing intonation unit boundaries, terminal pitch direction, accent, accent unit boundaries, pitch movements and pauses are then used in those systems.

Within the corpus linguistics tradition, Leech (1991) reports that notable exceptions to the lack of prosodic coding in spoken corpora are the London-Lund Corpus (LLC) - - described in Svartvick

(Ed.)(1990) – and the Lancaster/IBM Spoken English Corpus (SEC) – described, for example, in Knowles & Lawrence (1987) – An example of the kind of work carried out in prosodic coding in corpus linguistics is found in the papers published by Knowles (1991) and by Wichmann (1991) using the SEC. As mentioned before, SEC has recently been converted to MARSEC, and part of the project has consisted in the alignment of the prosodic annotation (Knowles, 1995; more information is found at URL http://midwich.reading.ac.uk/research/speechlab/marsec/marsec.html). The annotation is based on a tonetic stress mark system, within which types of accent, tone-unit boundaries and nuclear and non-nuclear syllables are distinguised.

The Text Encoding Initiative (Sperberg-McQueen & Burnard, (Eds.) 1994) considers the transcription of prosodic phenomena including pauses, tone units or intonational phrases and shifts, defined as the point at which some paralinguistic feature – - tempo, pitch range, tension, rhythm, and voice quality – of a series of utterances by any one speaker changes. The TEI also provides an example of a set of prosodic features for the representation of stress and pitch patterns that can be defined and documented by the transcriber.

French (1992) proposals adopted by the Network of European Reference Corpora (NERC) include prosodic information in Level Three and Level Four (see 2.3.2). In Level Three tone boundaries and tonic syllables are identified, while in Level Four Level Four head syllables and tone are transcribed. There is also provisions for an orthographic and a phonemic transcription aligned with a spectrogram and a fundamental frequency (Fo) contour.

The IPA (International Phonetic Alphabet) has a set of symbols for the representation of suprasegmental elements. On the occasion of the Kiel convention in 1989 a working group on Suprasegmental Categories coordinated by Bruce was set up (Bruce, 1988,1989). It was concluded that additions were needed to represent suprasegmentals within the IPA framework. As far as intonation was concerned, it was noted that there are no specific symbols for the notation of intonation - except for tones - in the IPA. Bruce's conclusions are that

> *There exists an apparent need for a direct way of symbolizing intonation in a phonetic transcription. However, the opinions diverge regarding the exact way of transcribing intonation. For a phonological transcription of intonation the symbolization is very much dependent on the language and the analysis.*
>
> (Bruce, 1989: 36-37)

The full set of symbols used for the transcription of suprasegmental elements in the IPA can be found in IPA (1993) (also available at URL http://www.arts.gla.ac. uk/IPA/ipachart.html).

In the domain of prosodic transcription systems to be used in speech research and in speech technology, ToBI (Tone and Break Index Tear) was developed to fulfill the need of a prosodic notation system providing a common core to which different researchers can add additional detail within the format of the system; it focuses on the structure of American English, but transcribes word grouping and prominences, two aspects which are considered to be rather universal (Price, 1992).

As described by Silverman et al. (1992) the system shows the following features: (1) it captures categories of prosodic phenomena; (2) it allows transcribers to represent some uncertainties in the transcription; (3) it can be adapted to different transcription requirements by using subsets or supersets of the notation system; (4) it has demonstrated high inter-transcriber agreement; (5) it defines ASCII formats for machine-readable representations of the transcription; and (6) it is equipped with software to support transcription using Waves and UNIX programmes.

A ToBI transcription for an utterance consists of symbolic labels for events on four parallel tiers: (1) orthographic tier, (2) break-index tier, (3) tone tier and (4) miscellaneous tier. Each tier consists of symbols representing prosodic events, associated to the time in which they occur in the utterance. The conventions for annotation according to TOBI are defined for text-based transcriptions and for computer-based labeling systems such as Waves.

Although primarily intended for English, work using the ToBI system is being carried out in other language such as Italian (Grice & Savino, 1995), German (Grice & Benzmueller, 1995) or Hungarian (Grice et al., 1995).

The system is also discussed in the chapter devoted to corpus representation of the *EAGLES Handbook on Spoken Language Systems* (EAGLES Spoken Language Working Group, 1995), and in Roach & Arnfield (1995).

The *Handbook* also discusses the analysis of intonation developed at the IPO (Eindhoven, The Netherlands). Rather than a transcription system, the IPO group has developed a full hierarchical theory of intonation based on the modelling of intonation contours – pitch or Fo contours – as stylized representations which are linguistically equivalent to the original contour (see 'T Hart et al., 1990 for a complete presentation of the model). A set of basic language-dependent "pitch movements" are proposed, and they are grouped in sequences of "pitch configurations"; "pitch contours" are build up from these configurations, and "intonation patterns" are defined on the basis of grouping similar pitch contours.

As explained in section 5.1.2 SAMPA (SAM Phonetic Alphabet) was developed to cater for the needs of speech technology applications. SAMPA offers symbols for the transcription of prosodic features such as lenght, word accent, stress, tonal movements, pauses and prosodic boundaries. In a review of the system Gibbon (1989) criticizes the theory-oriented character of the system and its inseparability from the tonetic theory of stress marking. Further work on prosodic transcription within the SAM *Speech Assessment Methodologies* project has lead to the development of other systems such as PROSPA, SAMSINT and SAMPROSA, all of them discussed in Gibbon (1989) and briefly summarised below.

PROSPA was developed by Selting and Gibbon (Selting, 1987, 1988) specially to meet the needs of discourse and conversation analysis but was also discussed within the Prosody Group in the SAM project (Wells at al., 1992). PROSPA is aimed at the high-level broad transcription which is needed for discourse analysis, and therefore, the categories used in the transcription are based on auditive criteria.

SAMSINT *SAM System for Intonation Transcription* has been proposed by the SAM Prosody Working Group, and was intended to be a computer-readable system for the transcription of intonation contours within defined intonation units. The system is based on INTSINT (see below), incorporating additional facilities and simplifications (Wells et al., 1992).

SAMPROSA *SAM Prosodic Alphabet* has been initially proposed by Gibbon (1889), incorporating results from discussions within the SAM Prosody Working Group. The system is intended both for prosodic transcription for linguistic purposes, and for prosodic labelling in speech technology and experimental phonetic research. The system allows the transcription of global, local, terminal and nuclear tones, length, stress, pauses and prosodic boundaries. It is documented in Wells et al. (1992) and the relevant information can be found at URL http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm.

Finally, INTSINT *International Transcription System for Intonation* aims at providing a system for cross-linguistic comparison of prosodic systems It has been developed by Hirst (1991,1994; Hirst & Di Cristo, forthcoming), based on a stylization procedure of the fundamental frequency – or pitch – contour (Fo) build up from interpolation between target points in which significant changes occur. It is then a system which is closely linked to the the phonetic realization of the intonation contour, but at the same time is able to symbolize this contour in terms of a phonological representation. INTSINT aims therefore at the symbolization of pitch levels or prosodic target points, each characterising a point in the fundamental frequency curve.

The Fo modelling is carried out automatically by a program called MOMEL (Hirst & Espesser, 1991) that, after Fo detection, provides 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.

The symbolization of prosodic target points is made by means of arrow symbols corresponding to different pitch levels, either relative or absolute.

The system has already been applied to several languages (see, for example, Hirst et al., 1993) and is being used in MULTEXT *Multilingual Text Tools and Corpora* project (Hirst et al., 1994; more information on the project is available at URL http://www.lpl.univ-aix.fr/projects/multext/index.html) for the encoding of intonation in the paragraphs contained in the EUROM.1 corpus.

This review shows that the systems developed so far have been designed with different purposes in mind and within different traditions. Nevertheless, it is possible to find some parameters that may help in comparing the external and internal features of each transcription system in order to assess its possible use in corpus linguistics and in speech work. The following dichotomies are suggested (Llisterri, 1994b) (available at URL http://www.lpl.univ-aix.fr/projects/multext/CES/CES2.html:

- Multi-tiered *vs.* one-tiered systems
  - One-tiered systems include the symbols for the representation of prosodic events within the segmental – orthographic or phonetic/phonemic – transcription, while in multi-tiered systems it is possible to distinguish different layers or levels, separating the segmental transcription from the suprasegmental coding. Examples of one-tiered systems can be found in the domain of discourse and conversation analysis or in the conventions adopted by the TEI and NERC; IPA, SAMPA and its derivations can be also classified within this category. TOBI and INTSINT are very clear examples of multi-tiered systems allowing the separation of different types of events from the segmental transcription. As far as the labelling of speech databases is concerned, the later systems seem to offer clear advantages.
- Machine readable symbols *vs.* non-machine readable symbols
  - Some of the transcription systems reviewed include a mapping between ASCII numbers and transcription symbols (e.g. SAMPA, SAMSINT or SAMPROSA). Other systems such as those used in discourse analysis and in corpus linguistics make use of characters which are usually available in computer keyboards; TOBI is another example of this category. It seems that a prosodic coding system aimed at facilitating exchange of labelled databases should ideally make use of machine-readable symbols.
- Systems that can be applied automatically *vs.* systems that rely on the transcriptor's judgment
  - The great majority of the systems described depend on the transcriptor's judgment, in the sense that the transcriber himself decides, after an auditory or acoustic analysis of the utterance, which is the symbol that more adequately reflects a given prosodic phenomenon. Only INTSINT can be automatically applied, taking the speech wave as a starting point and producing an abstract representation in a completely automatic way. Of course, this is an advantage when labelling of large speech databases has to be undertaken, since it ensures at least homogeneity of criteria.
- Multilingual *vs.* non-multilingual systems
  - Systems such as TOBI or PROSPA have been developed having one language in mind. Others such as SAMPA or SAMSINT address European languages, and IPA and INTSINT have been designed to cover a wider range of languages - actually both of them contain the term "international" in their denomination -. For the purposes of a multilingual project, it is essential that the coding system should be able to convey prosodic contrasts in a number of languages, and it seems logical to use a system conceived with that purpose.

- Theory-driven systems *vs.* data-driven systems
  - Some authors explicitly claim that their system is not model-dependent; this is the case of SAMPA. On the contrary, other authors provide the theoretical background in which their coding system is based; examples are SAMPROSA, TOBI or INTSINT. In both cases the assumptions behind the system are of phonological nature, or are based on the author's conception of the phonetics-phonology interface. On the other hand, the theory behind systems used in discourse and conversation analysis is defined by the needs, the practices and the models used in the field, since the events which are coded are those which are known to be relevant in order to explain the discursive or the interactional behaviour of the speakers.

5.2.2 Proposals for the transcription of the suprasegmental level

In a classical book on English intonation, Crystal (1969) discusses the principles that should guide a prosodic transcription system. According to him these are the following:

- – accuracy;
- – consistency;
- – be as automatically applicable as possible;
- – use the minimum of symbols;
- – establish degrees of complexity of symbols to reflect the different significance attached to the data; and
- – be broad, covering only those aspects which are linguistically significant.

On the overall, most of the systems examined in the previous section fulfill the conditions proposed by Crystal. However, different aims of the transcription may require different systems, and for this reason specific recommendations might be necessary according to the type of research for which prosodic transcription is needed.

The first issue to be discussed in this section is the prosodic events to be encoded in a spoken text. The following elements seem to be common to many prosodic transcription systems:

- Prosodic boundaries and prosodic units
- Tone or pitch level, terminal and non-terminal
- Pitch movements, pitch direction or pitch contour, both local and global
- Accent, at word or phrase level
- Lengthening
- Pauses

The Text Encoding Initiative (see 2.3.3) proposes the encoding of three elements which can be related to prosody:

- *Utterance*, defined as a strecht of speech usually preceded and followed by a pause or by a change of speaker

- *Pause*
- *Shift*, that might be used to signal changes in paralinguistic features – voice quality, loudness, pitch range and speech rate.

We have seen, moreover, that stress and pitch patterns can be represented. The symbols used for these purpose are punctuation marks in the example provided by Sperberg-McQueen & Burnard (Eds.) (1994): ¡.¿ for a low-fall intonation, ¡,¿ for a fall-rise, ¡?¿ for a low rise, ¡!¿ for a rise fall, and ¡:¿ for a lengthened syllable.

The proposal adopted by the Network of European Reference Corpora (NERC) (see 2.5.2) includes prosodic information in levels III and IV:

- Tone unit boundaries (Level III)
- Tonic syllables (Level III)
- Tones (Level IV)
- Head syllables (Level IV)

Taken together, TEI and NERC allow for the representation of global prosodic phenomena. When a more detailed representation is needed the NERC report suggests the use of SAMPROSA (*SAM Prosodic Alphabet*) (Teubert, 1993; Sinclair, 1993).

In terms of prosody encoding, a proposed recommendations would be to represent, at least, the two TEI elements *Utterance* and *Pause* (see 4). The representation of other prosodic phenomena such as those mentioned in NERC levels III and IV seems to be more adequately cared for by a transcription system such as SAMPROSA. Among these phenomena, at least tone unit or tone group boundaries and stress (or tonic syllables) could be included in a transcription containing basic prosodic information.

As far as the transcription system to be used is concerned, it is worth quoting the opinion of the EAGLES Spoken Language Working Group:

> *It is reasonable to assume nowadays that a prosodic transcriber will have access to at least the waveform and the Fo curve for the speech to be transcribed. In that case, the recommendation is to use either the ToBI or the IPO system (and the MARSEC system if a purely auditory transcription is being carried out. If the language to be transcribed is not English, and specially if the projected application of the prosodic transcription is in the field of speech technology, then it is probably best to use the IPO system if possible (i.e., if the basic "grammar" of contours has already been researched for that language). However, these can only be provisional recommendations, as little work has been carried out in prosodic labelling. In this situation, it may be that a different system entirely will prove more appropriate to the given language, and it is not possible to make absolute recommendations.*

More important than the choice of a particular system is the acknowledgement of the difficulties in providing recommendations in this area given the present state-of-the-art. Although ToBI is rapidly becoming a standard despite its orientation towards the transcription of English and the theoretical phonological assumptions underlying it, SAMPROSA offer the advantages of being accepted by NERC and of having been developed with both linguistics and speech technology needs in mind. However, the diversity of current proposals can be overcomed by developing mapping between systems in order to allow for conversions between them. This was one of the recommendations issued from the Madrid workshop 'Issues in Corpus Work' organised by the Text Corpus Working Group in January 1996.

In terms of the dichotomies presented in the previous sections, it would be advisable to choose a multi-tiered, machine-readable and multilingual prosodic transcription system. If it can be applied automatically instead of relying on the judgement of the transcriber, this would be an important advantage in the labelling of large corpora.

6 Summary of proposals and recommendations

We have tried to explore in this document some ways to achieve compatibility between NERC and TEI proposals, developed within the corpus linguistics community, and the practices usually followed by the speech community as documented in the *EAGLES Handbook on Spoken Language Systems* and in other sources reviewed.

Recommendations suggested here are based on surveys of current practice and tend to be based in common elements found in different traditions. For this reason, they are of a very general nature and have to be further developed to cover more specific needs.

For the encoding of spoken texts, the following set of elements to be encoded is suggested (see 2.5.3):

Vocal semi-lexical events

Vocal non-lexical events

Non-vocalised non-communicative events

Speaker identity

Speaking turns, indicating a change of speaker

Simultaneous speech or overlapping

Omissions in read text

Self-repairs

Word fragments

Unintelligible fragments

The need to develop conversion software between a user-friendly system of transcription and the TEI encoding scheme is also acknowledged.

A proposal for transcription and labelling has been put forward, consisting in three levels (see 2.5.3):

S1 — Orthographic representation of the text.

S2 — Phonemic representation of words in citation form: that is, the forms in which words are pronounced in isolation.

S3 — Phonetic transcription reflecting a discrete symbolic representation of the perceived actual realization of the utterance.

Of course, all these levels have to be linked to the speech signal itself, and the use of automatic alignment techniques to do so is encouraged.

As far as the orthographic representation is concerned, the following recommendations can be suggested (see 4.1):

Use conventional spelling forms as they appear in a standard dictionary. This also applies to contractions, reduced word forms, apostrophes, dialect forms, interjections and vocalised semi-lexical events.

If more than one orthographic form is possible or if non-standard spellings or spelling variations are necessary, maintain a lexicon of the spelling forms used in the transcription

Represent numbers, abbreviations, acronyms and spelled words in full orthographic form as pronounced by the speaker

It has to be noted that punctuation is still one aspect which would need a more in-depth discussion.

The rationale behind these recommendations is the possibility to create an automatic link between the orthographic transcription and the phonemic representation in level S2.

Concerning the choice of a segmental transcription system (see 5.1.3), the IPA (International Phonetic Alphabet) is to be recommended. Whenever a machine-readable equivalent is necessary, SAMPA (SAM Phonetic Alphabet) is recommended for phonemic transcriptions such as those proposed at level S2, and the X-SAMPA extension is to be considered for a phonetic transcription such as the one proposed at level S3.

The prosodic elements to be encoded are discussed in 5.2.2, where it is suggested to represent, at least, the two TEI elements *Utterance* and *Pause*.

The choice of a prosodic transcription system is also discussed in 5.2.2. ToBI (Tone and Break Indices) and SAMPROSA (SAM Prosodic Alphabet) – complemented by the X-SAMPA extension – are considered standard machine-readable systems, and the need to develop mappings between different systems is acknowledged. In general, the use of a multi-tiered, machine-readable and multilingual prosodic transcription system is recommended.

Some recommendations for data acquisition are also provided in 3, and can be summarised as follows:

If acceptable in the recording environment, and for optimal acoustical quality, use headset microphones.

Use digital recording devices. If direct recording into a computer is not possible, DAT (Digital Audio Tape) is recommended.

It is clear that these recommendations can only be provisional in the sense that they have to be validated and refined by applying them to different types of spoken materials, although most of them are based on current practice in different scientific communities. However, they are intended to be a first step towards a common set of working conventions which could improve the reusability of speech and spoken language resources.

7 References

ALLEN, G.D. (1988) “The PHONASCII System”, Journal of the International Phonetic Association 18,1: 9-25.

American Dialect Society (1992) Legal and Ethical Issues in Surreptitious Recording. Tuscaloosa & London: The University of Alabama Press (Publication of the American Dialect Society, 76).

ANDERSON, A.H. - BADGER, M.- BARD, E.G.- BOYLE, E.- DOHERTY, G.- GARROD, S.- ISARD, S.- KOWTKO, J.- McALLISTER, J.- MILLER, J.- SOTILLO, C.- THOMPSON, H.S.- WEINERT, R. (1991) “The HCRC Map Task corpus”, Language and Speech 34,4: 351-366

ANDERSSON, A.- BROMAN, H. (1993) “Towards automatic text-to-speech alignment” in Eurospeech’93. 3rd European Conference on Speech Communication and Technology. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 301-304

ATKINSON, J.M. - HERITAGE, J. (Eds.) (1984) Structures of social action. Studies in conversation analysis. Cambridge / Paris: Cambridge University Press / Editions de la Maison des Sciences de l’Homme

ATWELL, E. (1996) “Machine learning from corpus resources for speech and handwriting recognition”, in THOMAS, J.- SHORT, M. (Eds.) Using Corpora for Language Research. Studies in Honour of Geoffrey Leech. London: Longman. pp. 151-166

AUTESSERRE, D.- PÉRENNOU, G.- ROSSI, M. (1989) “Methodology for the transcription and labeling of a speech corpus”, Journal of the International Phonetic Association 19,1: 2-15

BALL, M.J.- CODE, C.- RAHINY, J.- HAZLETT, D. (1994) “Non segmental aspects of disordered speech: developments in transcription”, Clinical Linguistics & Phonetics 8,1: 67-88

BALL, M.J.- RAHILLY, J.- TENCH, P. (1996) The Phonetic Transcription of Disordered Speech. San Diego - London: Singular Publishing Group Inc.

BARRY, W.J.- FOURCIN, A.J. (1992) “Levels of Labelling”, Computer Speech and Language 6: 1-14

BIBER, D. (1988) Variation across Speech and Writing. Cambridge: Cambridge University Press.

BLANCHE-BENVENISTE, C.- BILGER, M.- ROUGET, Ch.- van den EYNDE, K. (1991) Le franais parlé. Études grammaticales. Paris: Editions du Centre National de la Recherche Scientifique (Sciences du Langage)

BLOMBERG, M.- CARLSON, R. (1993) “Labelling of speech given its text representation” in Eurospeech’93. 3rd European Conference on Speech Communication and Technology. Berlin, Germany, 21-23 September 1993. Vol. 3 pp. 1775-1778

BOVES, L.- den OS, E. (1995) Proposal for Transcription and Documentation Conventions for SpeechDat. SpeechDat deliverable, June 1995.

BRUCE, G. (1988) “2.3. Suprasegmental categories and 2.4. The symbolization of temporal events”, Journal of the International Phonetic Association 18,2: 75-76

BRUCE, G. (1989) “Report from the IPA working group on suprasegmental categories”, Lund University Department of Linguistics and Phonetics, Working Papers 35: 25-40

CARRÉ, R. (1992) “Speech Databases” in AINSWORTH, W.A. (Ed.) Advances in Speech, Hearing and Language Processing. A Research Annual. Volume 2. London: Jai Press. pp. 199-216.

CHAFE, W. (1995) “Adequacy, user-friendliness, and practicality in transcribing”, in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 54-61

CHAN, D.- FOURCIN, A.- GIBBON, D.- GRANSTRM, B.- HUCVALE, M.- KOKKINAKIS, G.- KVALE, K.- LAMEL, L.- LINDBERG, B.- MORENO, A.- MOUROPOULOS, J.- SENIA, F.-

TRANCOSO, I.- IN'T VELD, C.- ZEILIGER, J. (1995) "EUROM- A Spoken Language Resource for the EU", in Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology. Madrid, Spain, 18-21 September, 1995. Vol 1, pp. 867-870

COULTHARD, M.- MONTGOMERY, M (Eds.) (1981) Studies in Discourse Analysis. London: Routledge and Keagan Paul.

CROWDY, S. (1994) "Spoken corpus transcription", Literary & Linguistic Computing, 10: 25-28.

CROWDY, S. (1995) "The BNC spoken corpus", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 224-234

CRYSTAL, D. (1969) Prosodic Systems and Intonation in English. Cambridge: Cambridge University Press (Cambridge Studies in Linguistics, 1)

CUTLER, A. (Ed.) (1982) Slips of the Tongue and Language Production. Berlin: Mouton.

de GINESTEL-MAILLAND, A.- DE CALMÉS, M.- PÉRENNOU, G. (1993) "Multi-Level Transcription of Speech Corpora from Orthographic Forms" in Eurospeech'93. 3rd European Conference on Speech Communication and Technology. Berlin, Germany, 21-23 September 1993. Vol. 2 pp. 1441-1444

de JONG, E.D. (1992) Transcription and normalization method. Dutch spoken language. Utrecht, Working Paper, NERC-159

du BOIS, J.W. (1991) "Transcription design principles for spoken discourse research", Pragmatics 1: 71-106

du BOIS, J.W.- SCHUETZE-COBURN, S.-CUMMING, S.- PAOLINO, D. (1993) "Outline of discourse transcription", in EDWARDS, J.A.- LAMPERT, M.D. (Eds.) Talking Data: Transcription and Coding in Discourse Research. Hillsdale, N.J.: Lawrence Erlbaum Associates. pp. 45-90

DUCKWORTH, M.- ALLEN, G.- HARDCASTLE, W.- BALL, M. (1993) "Extensions to the International Phonetic Alphabet for the transcription of atypical speech", Clinical Linguistics & Phonetics, 7: 273-280

EAGLES SPOKEN LANGUAGE WORKING GROUP (1995) EAGLES Handbook on Spoken Language Systems.

EDWARDS, J.A. (1992) "Design principles in the transcription of spoken discourse" in SVARTVIK, J. (Ed.) Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Stockholm, 4-8 August, 1991. Berlin: Mouton de Gruyter. pp. 129-147

EDWARDS, J.A. (1993) "Principles and Contrasting Systems of Discourse Transcription", in EDWARDS, J.A.- LAMPERT, M.D. (Eds.) Talking Data: Transcription and Coding in Discourse Research. Hillsdale, N.J.: Lawrence Erlbaum Associates. pp. 3-32

EDWARDS, J.A. (1995) "Principles and alternative systems in the transcription, coding and markup of spoken discourse", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 19-34

EDWARDS, J.A.- LAMPERT, M.D. (Eds.) (1993) Talking Data: Transcription and Coding in Discourse Research. Hillsdale, N.J.: Lawrence Erlbaum Associates.

ESKÉNAZI, M. (1993) "Trends in Speaking Styles Research" in Eurospeech'93. 3rd European Conference on Speech Communication and Technology. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 501-512

ESLING, J.H. (1988) "Computer coding of IPA symbols and detailed phonetic representations of computer databases", Journal of the International Phonetic Association 18,2: 99-106

ESLING, J.H. (1990) "Computer Coding of the IPA: Supplementary Report ", Journal of the International Phonetic Association 20,1: 22-26

ESLING, J.H.- GAYLORD, H. (1993) “Computer Codes for Phonetic Symbols”, Journal of the International Phonetic Association 23,2: 77-82

FINK, G.A.- JOHANNTOKRAX, M.- SCHAFFRANIETZ, B. (1995) “A flexible formal language for the orthographic transcription of spontaneous spoken dialogues , in Eurospeech’95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology. Madrid, Spain, 18-21 September, 1995. Vol 1, pp. 871-874

FOURCIN, A.- HARLAND, G.- BARRY, W. - HAZAN, V (Eds.) (1989) Speech Input and Output Assessment. Multilingual Methods and Standards. Chichester: Ellis Horwood Ltd.

FRENCH, J. P. (1992) “Transcription proposals: multi-level system”, Working Paper, University of Birmingham, October 1992. NERC-WP 4-50

FRENCH, J.P. (1991) “Updated notes for soundprint transcribers + one page sample text from COBUILD corpus”, JP French Associated, York and COBUILD, Birmingham, October 1991, NERC-WP4-47

FROMKIN, V. (Ed.) (1973) Speech Errors as Linguistic Evidence. The Hague: Mouton.

FROMKIN, V. (Ed.) (1980), Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand. New York: Academic Press.

GIBBON, D. (1989) Survey of Prosodic Labelling for EC Languages. SAM-UBI-1/90, 12 February 1989; Report e.6, in ESPRIT 2589 (SAM) Interim Report, Year 1. Ref. SAM-UCL G002. University College London, February 1990.

GRICE, M.- BENZMUELLER, R. (1995) “Transcription of German intonation using ToBI tones - The Saarbruecken system”, Phonus 1, University of the Saarland, pp. 33-51

GRICE, M.- SAVINO, M. (1995) “Low tone versus ‘sag’ in Bari Italian intonation; a perceptual experiment”, Proc. XIII International Congress of Phonetic Sciences, Stockholm

GRNNUM THORSEN, N. (1987) “Suprasegmental transcription”, ARIPUC, Annual Report of the Institute of Phonetics, University of Copenhagen 21: 1-28

GUMPERZ, J.J.- BERENZ, N. (1993) “Transcribing Conversational Exchanges”, in EDWARDS, J.A.- LAMPERT, M.D. (Eds.) Talking Data: Transcription and Coding in Discourse Research. Hillsdale, N.J.: Lawrence Erlbaum Associates. pp. 91-122

HALLIDAY, M.A.K. (1989) Spoken and written language. Oxford: Oxford University Press (Language Education Series)

HESS, W.- KOHLER, K.- TILLMANN, H.G. (1995) “The PhonDat-Verbmobil Speech Corpus” , in Eurospeech’95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology. Madrid, Spain, 18-21 September, 1995. Vol 1, pp. 863-866

HIERONYMUS, J.L. (1994) ASCII phonetic symbols for the world’s languages: Worldbet. AT&T Bell Laboratories, Technical Memo.

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. (1994) “The symbolic coding of fundamental frequency curves: from acoustics to phonology”, in FUJISAKI, H. (Ed.) Proceedings of International Symposium on Prosody, Satellite Workshop of ICSLP 94, Yokohama, September 1994.

HIRST, D.J. - IDE, N.- VRONIS, J. (1994) “Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project”, Proceedings of the ESCA/IEEE Workshop on Speech Synthesis, New York, September 1994.

HIRST, D.J. - DI CRISTO, A.- LE BESNERAIS, M.- NAJIM, Z.- NICOLAS, P.- ROMÉAS, P. (1993) "Multilingual modelling of intonation patterns", in HOUSE, D.- 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. - 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.- VRONIS, J. (1994) "Coding fundamental frequency patterns for multilingual synthesis with INTSINT in the MULTEXT project", Proceedings of the ESCA/IEEE Workshop on Speech Synthesis, New York, September 1994.IDE, N. (Coord.) (1996) Corpus Encoding Standard. April,1996

HIRST, D.J. - DI CRISTO, A. (Eds.) (forthcoming) Intonation Systems. A Survey of 20 Languages. Cambridge: Cambridge University Press.

IPA (1989) "The IPA 1989 Kiel Convention Workgroup 9 report: Computer Coding of IPA symbols and Computer Representation of Individual Languages", Journal of the International Phonetic Association 19,2: 81-92

IPA (1993) "IPA Chart, revised to 1993", Journal of the International Phonetic Association 23,1.

JOHANSSON, S. (1995a) "The approach of the Text Encoding Initiative to the encoding of spoken discourse", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 82-98

JOHANSSON, S. (1995b) "The Encoding of Spoken Texts", Computers and the Humanities 29,1: 149-158; in IDE, N.- VÉRONIS, J. (Eds.) (1995) The Text Encoding Initiative: Background and Context. Dordrecht: Kluwer Academic Publishers.pp. 149-158

KUGLER-KRUSE, M. (1987) Computer Phonetic Alphabet. ESPRIT Linguistic Analysis of the European Languages. Report BU-CPA0267, July, 1987.

KNOWLES, G. (1991) "Prosodic labelling: the problem of tone group boundaries", in JOHANSSON, S.- STENSTRM, A. (Eds.) English Computer Corpora. Selected Papers and Research Guide. Berlin: Mouton de Gruyter. pp. 149-163

KNOWLES, G. (1995) "Converting a corpus into a relational database: SEC becomes MARSEC", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 208-219

KNOWLES, G.- LAWRENCE, L. (1987) "Automatic intonation assignment" in GARSIDE, R.- LEECH, G.- SAMPSON, G. (Eds.) The Computational Analysis of English: A Corpus-based Approach. London: Longman. pp. 139-148

KOHLER, K.- LEX, G.- PÄTZOLD, M.- SCHEFFERS, M.- SIMPSON, AP..- THON, W., in collaboration with DRAXLER, C.- JOHNE, B.- SCHIEL, F.- FAUST, L. (1994) Handbuch zur Datenaufnahme und Transliteration, in TP14 from VERBMOBIL - 3.0 Verbmobil Technisches Dokument 11, Kiel: IPDS, March 1994.

LAMEL, L.- COLE, R. (1996) "Spoken Language Corpora", in COLE, R.A.- MARIANI, J.- USZKOREIT, H.- ZAENEN, A.- ZUE, V. (Eds.) Survey of the State of the Art in Human Language Technology.(http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html)

LEECH, G. (1991) "The State of the Art in Corpus Linguistics" in AIJMER, K.- ALTENBERG, B. (Eds.) English Corpus Linguistics. Studies in Honour of Jan Svartvik. London: Longman. pp. 8-29.

LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) (1995) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman.

LÉON, P.- MARTIN, P. (1970) Prolegomènes à létude des structures intonatives. Montréal: Didier (Studia Phonetica 2).

LINDBLOM, B. (1987) "Adaptive Variability and Absolute Constancy in Speech Signals" in Proceedings XIth ICPhS. The Eleventh International Congress of Phonetic Sciences. August 1-7, 1987, Tallinn, Estonia. USSR. Vol. 3. pp. 9-18.

LLISTERRI, J. (1994a) Events symbolized and labels used in the transcription of spoken language. Draft Report, October 1994. EAGLES Spoken Texts Cross-Group.

LLISTERRI, J. (1994b) Prosody Encoding Survey. WP 1 Specifications and Standards. T1.5. Markup Specifications. Deliverable 1.5.3. Final version, 15 September 1994. LRE project 62-050 MULTEXT. (http://www.lpl.univ-aix.fr/projects/multext/CES/CES2.html)

MacWHINNEY, B. (1991) The Childes Project: Tools for Analyzing Talk. Hillsdale, N.J.: Lawrence Erlbaum.

MARCOS MARÍN, F.- BALLESTER, A.- SANTAMARÍA, C. (1993) "Transcription Conventions used for the Corpus of Spoken Contemporary Spanish", Literary and Linguistic Computing 8, 4: 283-292

MOORE, R.K. (1991) "User Needs in Speech Research", Proceedings of the Workshop on European Textual Corpora, Pisa, Italy, 1991.

NELSON, G. (1995) "The International Corpus of English: mark-up for spoken language", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 220-223

OCHS, E. (1979) "Transcription as Theory" in OCHS, E.- SCHIEFFELIN, B.B. (1979) Developmental Pragmatics. New York: Academic Press. pp. 43-72

PAYNE, J. (1992) "Report on the compatibility of J P French's spoken corpus transcription conventions with the TEI guidelines for transcription of spoken texts", Working Paper, COBUILD Birmingham and IDS Mannheim, December 1992, NERC - WP8/WP4 -122

PAYNE, J. (1995) "The COBUILD spoken corpus: transcription conventions", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 203-207

PEPPÉ, S. (1995) "The Survey of English Usage and the London-Lund Corpus: computerizing manual prosodic transcription", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp.187-202

POLS, L. C. W. (1987) "Speech Technology and Corpus Linguistics", en MEIJS, W. (Ed.), Corpus Linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi.

PRICE, P. (1992) Summary of the Second Prosodic Transcription Workshop: the TOBI (TOnes and Break Indices) Labeling System. Nynex Science and Technology, Inc. 5-6 April, 1992. Linguist List vol. 3-761, 9 October 1992.

PSATHAS, G.- ANDERSON, T. (1992) The 'practice' of transcription in conversation analysis. Working paper, INL Leiden, June 1992. NERC-WP4-163

PULLUM, G.K.- LADUSAW, W.A. (1986) Phonetic Symbol Guide. Chicago: The University of Chicago Press.

ROACH, P.- ARNFIELD, S. (1995) "Linking prosodic transcription to the time dimension", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 149-160

SAM (1992) "Europec software V.4.1 User's Guide (SAM-ICP-045)" in SAM User Guide to ETR Tools. ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation. Ref., SAM-UCL-G007.

SELTING, M. (1987) "Descriptive categories for the auditive analysis of intonation in conversation", Journal of Pragmatics 11: 777-791

SELTING, M. (1988) "The role of intonation in the organization of repair and problem handling sequences in conversation", Journal of Pragmatics 12: 293-322.

SCHEITER, S. (1992a) German spoken language corpora and their text representation schemes - an overview. Working Paper, IDS Mannheim, August 1992, NERC-WP4-543

SCHEITER, S. (1992b) Text Representation and Annotation Schemes in German Language Corpora. Technical Report, IDS, Mannheim, NERC-135

SCHMIDT, M.S. -SCOTT, C.- JACK, M.A. (1993) "Phonetic transcription standards for European names (ONOMASTICA)" in Eurospeech'93. 3rd European Conference on Speech Communication and Technology. Berlin, Germany, 21-23 September 1993. Vol. 1 pp. 279-282

SCHUETZE-COBURN, S. (1991) "Units of intonation in discourse: a comparison of acoustic and auditory analysis", Language and Speech 34,3: 207-234

SILVERMAN, K.- BECKMAN, M.- PITRELLI, J.- OSTENDORF, M.- WIGHTMAN, C.- PRICE, P.- PIERREHUMBERT, J.- HIRSCHBERG, J. (1992) "TOBI: A standard for labelling English prosody", Proceedings of the Second International Conference on Spoken Language Processing, ICSLP-92. Banff, October 1992. pp. 867-870

SINCLAIR, J. (Ed.) (1987) Looking Up, An Account of the COBUILD Project. London: Collins

SINCLAIR, J. (1992) "NERC WP 4 Spoken Language Encoding. Evaluation for English" NERC

SINCLAIR, J. (1993) "Text Representation: Written Language / Spoken Language" (Draft version) Chapter 3 NERC Report. Final version in NERC (1994) NERC-1. Network of European Reference Corpora. Final Report. Pisa. ("Spoken Language", "Phonetic – Phonemic and Prosodic Annotation"); to be published as CALZOLARI, N.- BAKER, M.- KRUYT, P.G. (Eds.) Towards a Network of European Reference Corpora. Pisa: Giardini.

SINCLAIR, J. (1994) EAGLES Corpus Typology. Draft Work in Progress. EAGLES Document EAG-CSG/IR-T1.1, October, 1994

SINCLAIR, J. (1995) "From theory to practice", in LEECH, G.- MYERS, G.- THOMAS, J. (Eds.) Spoken English on Computer: Transcription, Markup and Applications. Harlow: Longman. pp. 99-112

SINCLAIR, J. - BALL, J. (1996) EAGLES Text Typology. July, 1995.

SPERBERG-McQUEEN, C.M.- BURNARD, L. (Eds.) (1994) Guidelines for Electronic Text Encoding and Interchange. TEI P3. Chapter 11: Transcriptions of Speech. Association for Computational Linguistics – Association for Computers and the Humanities – Association for Literary and Linguistic Computing: Chicago and Oxford.

STENSTRÖM, A.-B. (1994) An Introduction to Spoken Interaction. London - New York: Longman (Learning about Language).

SVARTVIK, J. (Ed.) (1990) The London-Lund Corpus of Spoken English: Description and Research. Lund: Lund University Press.

TEUBERT, W. (1993) "Phonetic/Phonemic and Prosodic Annotation" Final Report, IDS Mannheim, February 1993, NERC-WP8-171

' T HART, J.- COLLIER, R.- COHEN, A. (1990) A Perceptual Study of Intonation. An Experimental-Phonetic Approach to Intonation. Cambridge: Cambridge University Press. (Cambridge Studies in Speech Science and Communication)

TILLMANN, H.G.- POMPINO-MARSCHALL, B. (1993) "Theoretical Principles Concerning Segmentation, Labelling Strategies and Levels of Categorical Annotation for Spoken Language Database Systems" in Eurospeech'93. 3rd European Conference on Speech Communication and Technology. Berlin, Germany, 21-23 September 1993. Vol. 3 pp. 1691-1694

TUSÓN VALLS, A. (1995) Anàlisi de la conversa. Barcelona: Empúries (Biblioteca Universal Empúries, 73)

VILLENA PONSODA, J.A. (1992) Representation procedures and schemes for Spanish oral corpus of University of Málaga. Working Paper, University of Málaga, December 1992, NERC-WP4-141

VILLENA PONSODA, J.A. (1994) "Pautas y procedimientos de representación del corpus oral de la Universidad de Málaga. Informe preliminar", in ALVAR EZQUERRA, M.- VILLENA PONSODA, J.A. (Coord) Estudios para un corpus del español. Málaga: Universidad de Málaga. pp. 73-102

WELLS, J.C. (1987) "Computer Coded Phonetic Transcription" ,Journal of the International Phonetic Association 17,2: 94-114.

WELLS, J.C. (1989) "Computer-coded phonemic notation of individual languages of the European Community ", Journal of the International Phonetic Association 19,1: 31-54

WELLS, J.C. (1995) "Computer-coding the IPA: a proposed extension of SAMPA". Draft version. (Available as a postscript file at ftp: pitch.phon.ucl.ac.uk, internet address 128.40.52.11, (username: ftp, password: ftp) directory: /pub/sam, file name: ipasam-x.ps.

WELLS, J.C.- BARRY, W.- GRICE, M.- FOURCIN, A.- GIBBON, D. (1992) Standard Computer-Compatible Transcription. SAM Stage Report Sen.3 SAM UCL-037, 28 February 1992. In SAM (1992) ESPRIT PROJECT 2589 (SAM) Multilingual Speech Input/Output Assessment, Methodology and Standardisation. Final Report. Year Three: 1.III.91-28.II.1992. London: University College London.

WELLS, J.C.- HOUSE, J. (1995) The Sounds of the International Phonetic Alphabet. London: Department of Phonetics and Linguistics, University College London. Booklet + Tape.

WICHMANN, A. (1991) "A study of up-arrows in the Lancaster /IBM Spoken English Corpus", in JOHANSSON, S.- STENSTRÖM, A. (Eds.) English Computer Corpora. Selected Papers and Research Guide. Berlin: Mouton de Gruyter. pp. 165-178

WINSKI, R. - MOORE, R.- GIBBON, D. (1995) "EAGLES Spoken Language Working Group: Overview and Results", in Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology. Madrid, Spain, 18-21 September, 1995. Vol 1, pp. 841-844.