

## DELIVERABLE IDENTIFICATION

Identification number	LRE-4001 - SD1.3.2
Type	Technical Report
Title	Specification of orthographic transcription and lexicon conventions
Status	Final
Deliverable	SD1.3.2
Work Package	WP 1
Task	Task 1.3
Period covered	T01 - T06
Date	16 January 1997
Version	2.4
Number of pages	22
Author(s)	Francesco Senia, CSELT Jeroen G. van Velden, Philips
Workpackage (WP)/ Task (T) responsible	WP 1.4 - Richard Winski, Vocalis / Task 1.3 - Kamran Kordi, GEC
Project contact point:	Harald Hoege Siemens AG, ZFE T SN 5, D-81730 München phone: + 49 89 636 3374 fax: + 49 89 636 49802 e-mail: hoege@habicht.zfe.siemens.de
CEC project officer	Mr. J. Soler
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	

Key words	Speech, database, transcription, orthographic, lexicon, recognition.
Abstract	<p>This document provides a specification of the transcription and lexicon conventions used for the telephone speech databases to be collected over the telephone network in the SpeechDat project. It should be read in conjunction with other accompanying deliverables in WP1 specifying recording conditions, speaker attribute coverage, transcription and validation.</p> <p>These conventions are based on the conventions used by LDC/ARPA in producing the ATIS CD-ROMs. The transcription is intended to be a quick and broad ORTHOGRAPHIC transcription. A few details are included that represent audible acoustic events present in the waveform files.</p> <p>The overall aim is to keep as much speech in the corpus as possible and to avoid the need for deleting recordings from the corpus due to some extra noises, disfluencies, etc.</p> <p>In summary, the principles used in these conventions are “Keep it simple” and “Document everything adequately”.</p>
Status of the abstract	public

Received on	
Recipient's catalogue number	

## DOCUMENT EVOLUTION

Version	Date	Status	Notes
1.0	26/07/96	first draft	discussed by WP1 partners
2.0	08/10/96	second draft	to be discussed in Aalborg
2.1	29/10/96	third draft	Aalborg agreements included
2.2	20/11/96	Pre final	Minor changes
2.3	16/01/97	Final draft	Labels for noise markers changed
2.4	16/01/97	Final	Approved final version. Minor changes made.

## Contents

<b>INTRODUCTION</b>	<b>5</b>
<b>1 POINTS OF DEPARTURE</b>	<b>5</b>
<b>2 GENERAL TRANSCRIPTION CONVENTIONS</b>	<b>6</b>
2.1 Case	6
2.2 Spelling	6
2.3 Number sequences	7
2.4 Letter sequences	7
2.5 Punctuation	8
2.6 Mispronunciations	8
2.7 Unintelligible words	8
2.8 Word Fragments	9
2.9 Verbal Deletions	9
2.10 Non-Speech Acoustic Events	9
Since 50% of the calls of the Speaker Verification Databases is also from the mobile network, the same conventions as for the MDB databases apply.	11
2.11 Some language specific options	11
<b>3 PROSODIC ANNOTATION (OPTIONAL)</b>	<b>11</b>
<b>4 TRUNCATED WAVEFORMS</b>	<b>11</b>
<b>5 RECORDING QUALITY ASSESSMENT (OPTIONAL)</b>	<b>12</b>
<b>6 TRANSCRIPTION PROCEDURE</b>	<b>13</b>
<b>7 PRONUNCIATION LEXICON FILE</b>	<b>14</b>
7.1 Objective for the pronunciation lexicon	14
7.2 Handling of orthographic and pronunciation variants	15
7.3 Structure and conventions	16

<b>7.4 Apostrophes and hyphens</b>	<b>17</b>
<b>7.5 Other optional information</b>	<b>17</b>
<b>7.6 Language-specific issues</b>	<b>19</b>
<b>APPENDIX A SAMPA-COMPUTER READABLE PHONETIC ALPHABET</b>	<b>20</b>
<b>BIBLIOGRAPHY</b>	<b>23</b>

## **Introduction**

This document contains the transcription conventions for the SpeechDat corpora. The starting point of these conventions are the conventions used by LDC/ARPA in producing the ATIS CD-ROMs [1]. The project has decided to simplify the transcription task to enable it to be performed quickly and to represent the most important acoustic events adequately for training and testing of automatic speech recognisers.

The documentation accompanying each language database will describe fully all optional conventions and transcriptions used.

### **1 Points of departure**

The transcription is intended to be an ORTHOGRAPHIC, lexical transcription with a few details included that represent audible acoustic events (speech and non speech) present in the corresponding waveform files. The extra marks contained in the transcription aid in interpreting the text form of the utterance.

The transcription is intended to be a quick and broad transcription. Transcribers should not have to agonise over decisions, but rather realise that their transcription is intended to be a rough guide that others may examine further for details.

Transcriptions should be made in two passes: one pass in which WORDS are transcribed, and a second pass in which the additional details are added. Background noises and sounds like “uh” are easy to miss unless specifically attended to. It is recommended that transcribers have some background in phonetics and/or linguistics, or that their training and preparation for the transcription task cover some basics in acoustic phonetics and dialect and style variations.

The overall aim is to keep as much speech in the corpus as possible and to avoid the need for deleting recordings from the corpus due to some extra noises, disfluencies, etc.

The conventions comprise both mandatory and optional transcriptions. All transcriptions should precisely follow the mandatory guidelines. The optional transcriptions are marked OPTIONAL in this document, and if provided should be documented and should follow these guidelines precisely. This is to regulate the task of external validation. Markings which are optional have been chosen to be easily removed or translated by automatic means to yield the base transcription form.

The documentation provided with the database transcriptions should accurately provide details of which optional transcriptions were performed, and all relevant additional information, such as standard dictionary, preferred spelling variants, etc.

In summary, the principles are “Keep it simple” and “Document everything adequately”.

## 2 General transcription conventions

### 2.1 Case

Transcriptions are CASE SENSITIVE, unless specified otherwise in the documentation of the database. The use of small letters and capital letters may be advantageous in two ways:

- (1) It allows to keep the names together in the lexicon, if each name starts with a capital letter<sup>1</sup>.
- (2) It allows to distinguish the pronunciation of single letter words and of spelled letters in the lexicon. For example if a language has the word 'v' (pronounced as /v/), and the spelled letter 'v' (pronounced as /vi:/), then the spelled letter could be written (case-sensitive) with a capital. This would avoid confusion in the lexicon<sup>2</sup>.

If these two arguments are not relevant for a language then a case-insensitive approach or a single-case-only approach can be adopted.

The character set to be used for the transcriptions is ISO-8859-1 (ISO-8859-2 for Slovenian and ISO-8859-7 for Greek). The table used must be printed onto the CDs in postscript. The directory to be used is <database>\DOC; the filename to be used is ISO88591.PS (or ISO88592.PS / ISO88597.PS).

### 2.2 Spelling

Normal lexical items will be represented by their spellings in the normal way. It is advised to stick to the normal spelling as much as possible. This means i.e. that hyphens are used in the normal way. One dictionary or word list should be chosen (e.g. Duden for German, Larousse for French, Van Dale for Dutch). Each site/language maintains a lexicon of spellings of words used in the SpeechDat corpus.

In many languages there are words or expressions which can be spelled in two or more different ways. To maintain consistency, each site/language must compile a list of

---

<sup>1</sup> Per language decisions must be made concerning the treatment of compound proper nouns like “*Stephenson Way*”, “*mister Giscard d'Estaing*”, etc. These should be documented in each case.

<sup>2</sup> This differs from the WSJ convention which did not distinguish English indefinite article "a" from spelled letter "a".

such items, with the normalised spelling. For instance, in American English the spelling forms “*all right*” and “*alright*” co-exist; one of these forms must be established as the standard. Only this standard form will be used for annotation. The list of items with alternative spellings will be included on the CD-ROMs (<database>\DOC\SPELLALT.DOC).

It is probably profitable to always select the form yielding least “words”, because that should yield the most powerful language model.

Abbreviations should be represented by their full orthographic forms, unless they are spoken in their abbreviated form. Exceptions are normally occurring abbreviations such as Mr, Mrs, Messrs, some of which do not have non-abbreviated forms.

To support homogeneity in spelling conventions used, it is strongly recommended to employ an electronic spelling checker. If so, the make and type of the checker should be reported.

Pronunciation variations should not be indicated by different spellings in the transcription, but by different phonemic transcriptions in the lexicon. The most common pronunciation (if known) should be given first. It is acknowledged that the link of a specific pronunciation to individual appearances of a word in the databases is lost in this way.

### **2.3 Number sequences**

Number sequences (flight numbers, times, dates, aircraft types, money amounts, etc.) will be spelled out to reflect what was said (“*flight six one three*”; “*seven thirty*”; “*august twenty first*”; “*seven forty seven*”; “*four hundred and ten dollars*”). If digits have alternate pronunciation forms (e.g. “*zero*” or “*oh*” or “*naught*” in English), the transcription should accurately reflect the form actually pronounced.

### **2.4 Letter sequences**

Letter sequences occur in spelled words, ZIP-codes, acronyms and abbreviations (“*D F W*”; “*A P slash eighty*”; “*P M*”; “*C O*”; etc.) Letters should be in upper case, separated by a space. If letters have several names, like Y in Dutch, the actual name used must be transliterated instead of ‘Y’. The AM and PM of times (e.g., “*five thirty P M*”) will be treated as examples of letter sequences, i.e., upper case and separated by a space, with no periods.

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.

If a speaker realises letters, especially consonants, by producing their phonetic form, letters within solidus lines are used e.g.: /B/ /A/ /L/ spelling the Dutch word 'bal'. Entries between such solid lines should also appear as such in the lexicon.

As there can be different pronunciation schemes for letter spellings which it would be beneficial to identify, it is suggested that A B C ... Z be used for the most common spelling form, and variations be marked by unique letter sequences which are not confusable with words e.g. ZEE for US Z. It is recommended to document the normal spelling form (e.g. B : /b i:/, C : /s i:/, etc.). In any case these will be in the lexicon.

## 2.5 Punctuation

No punctuation will be provided in the transcription other than those symbols used for special transcription purposes. However the label files should retain all punctuation provided to the speaker in the prompting text, including mistakes if these occurred.

## 2.6 Mispronunciations

Mispronounced words that are nevertheless intelligible will be marked with one star<sup>3</sup> attached to the left of the word which is mispronounced e.g.

*\*transportation*

instead of the mispronounced

*transportetation*

Words preceded by a star include mispronunciations such as words with extra or omitted syllables, but a star should not be used to indicate pronunciations of words that represent normal dialectal (e.g., "warshed" for "washed" or "cah" for "car") or stylistic variation (e.g., "bout" for "about" or "wanna" for "want a" or for "want to"). If the speaker would not consider the pronunciation an error, the star notation should not be used. Obviously, there may be some clear and some unclear cases; transcribers should use their best judgement.

In stretches of speech that are mispronounced, each mispronounced word is marked individually.

## 2.7 Unintelligible words

---

<sup>3</sup> Note: This differs from the WSJ convention which used two stars, and also permitted the mispronunciation to be represented by modifying the orthography.

Words or stretches of speech that are completely unintelligible are denoted by a sequence of two asterisks: "\*\*\*"<sup>4</sup>. The "\*\*\*" marker is separated from neighbouring words with spaces.

## 2.8 Word Fragments

Word fragments, i.e. instances in which the speaker did not complete a word, will be considered a mispronunciation. It is accordingly marked with a star attached to the left of the intended word. The full word should appear, not a text fragment, as this can complicate the lexicon and create confusion if fragments are textually the same as valid words<sup>5</sup>.

## 2.9 Verbal Deletions

Verbal deletion means words spoken by the user but which, in the opinion of the transcriber, are superseded by subsequent speech explicitly (e.g., "*show flights i mean fares*") or implicitly (e.g., "*show me the fares flights to Boston*"). Verbal deletions are not indicated as such in the SpeechDat corpora. Transcribers should simply indicate all the words they hear. Verbal deletions occur any time there is a repetition or restart. In repetitions, one or more words are repeated, and there may or may not be extra material inserted into the repetition, for example:

*show me the flights the flights to boston*

*show me the flights the nonstop flights to boston*

In restarts, words are not repeated, but the speaker changes direction, as in:

*show me the how many flights go to boston*

## 2.10 Non-Speech Acoustic Events

4 categories of non-speech acoustic events must be transcribed. Events will only be transcribed if they are clearly distinguishable. Very low-level, non-intrusive events will be ignored. The event will be transcribed at the place of occurrence<sup>6</sup>, using the defined symbols in square brackets. For noise events that occur over a span of one or more words, the transcription should indicate the beginning of the noise, just before the first word it affects<sup>7</sup>.

---

<sup>4</sup> The corresponding WSJ marking was: [unintelligible]

<sup>5</sup> This is different from the SpeechDat(M) specification which used text fragments

<sup>6</sup> It is often difficult to localise these events; transcribing the utterance first, and listening for these events in a second pass is the correct procedure.

<sup>7</sup> There is no notation for spans of noise events as in the WSJ conventions.

The first two categories of acoustic events originate from the speaker, and the other two categories originate from another source. Sounds originating from the speaker usually do not overlap with the target speech, sounds originating from other sources can of course occur simultaneously with the speech.

The 4 categories are:

- [fil]: Filled pause. These sounds can well be modelled in a filled pause model in speech recognisers. Examples of filled pauses: uh, um, er, ah, mm.
- [spk]: Speaker noise. All kinds of sounds and noises made by the calling speaker that are not part of the prompted text, e.g. lip smack, cough, grunt, throat clear, tongue click, loud breath, laugh, loud sigh.
- [sta]: Stationary noise. This category contains background noise that is not intermittent and has a more or less stable amplitude spectrum. Examples: car noise, road noise, channel noise, GSM noise, voice babble (cocktail-party noise), public place background noise, street noise.
- [int]: Intermittent noise. This category contains noises of an intermittent nature. These noises typically occur only once (like a door slam), or have pauses between them (like phone ringing), or change their colour over time (like music). Examples: music, background speech, baby crying, phone ringing, door slam, door bell, paper rustle, cross talk.

The distinction between stationary and intermittent noise is considered a relevant one. Utterances containing stationary noise can be used to train robust models with noise background, especially if the marker is before the first word of the sentence. In contrast, utterances containing intermittent noise will generally be discarded for training.

Only the 4 defined categories should appear in the final transcriptions, without any further sub-categorisation. (If other symbols are used, for instance because it is easier for the transcribers, a post-hoc conversion should ensure that only the 4 defined categories show up in the final transcriptions). The symbols should be separated from surrounding words by spaces.

For MDB databases some additional considerations must be made. Annotation of non-speech acoustic events are meant to indicate unexpected events in the recorded speech. For the recordings of the fixed database, having in principal a good communication channel, this means that nearly all noise events must be transcribed. For GSM however the channel itself is usually noisy, and it makes no sense to mark this common characteristic. Additionally four acoustic conditions (Office, public place, road-side, moving vehicle) are defined to make the mobile recordings. Again transcribing noises that are common in such environment, is very redundant. Therefore only the non-speech acoustic events that are not common for the channel and the recording environment, are to be annotated. To ensure familiarity of the transcriber

with the common characteristics of the recording condition, he should listen to a number of recordings made in this condition, before starting the transcription. At each item, the transcriber is prompted from which environment the call originated. Preferably in one session only the recording from one environment should be transcribed.

Additionally a number of typical events can occur in mobile communication channels. They are fading, losing time frames, heavy distortion. Every word affected by these typical mobile distortions should be marked by a & sign, attached to the left of the word without intervening spaces.

Since 50% of the calls of the Speaker Verification Databases is also from the mobile network, the same conventions as for the MDB databases apply.

## **2.11 Some language specific options**

Frequently in French a schwa may occur at the end of words ending in a consonant. For extra syllable at the end of the word { } can be used, e.g. neuf{e}. So, the curly brackets can be used for extra syllables, provided that there is regular variation in pronunciation.

For French liaison a special indication can be used: no marking if liaison has been applied. If no liaison is applied, the + is put after the consonant, e.g. petit enfant (pronounced with t petiT enfant), petit+ enfant (pronounced without T: peti enfant). The + can be used for regular variation in segmental pronunciation. If words both with and without the + appear in the transcriptions, they should also appear separately in the lexicon, with the corresponding pronunciations.

## **3 Prosodic Annotation (optional)**

No prosodic annotation will be normally used.

Optionally lengthening of sounds may be indicated by a colon, e.g.

*show me the f:lights to Boston*

Silent pauses may be marked with a period (“.”). The use of the period indicates a significant silence, i.e., one that is clearly noticeable by listening, and which is significantly longer than a silence associated with a stop consonant closure for the rate of speech used by the speaker. Example:

*show me the . flights to boston*

## **4 Truncated Waveforms**

If a speech signal file is truncated due to a recording error, the following notation is to be used:

- Beginning of utterance truncation:        ~*transcription*
- End of utterance truncation:               *transcription*~
- Beginning and end of utterance truncation:   ~*transcription*~

There is a difference between an utterance which is truncated and is now incomplete, but which has not damaged the initial or final words, and an utterance when word(s) have been damaged. The ~ indicates truncation of the word it is attached to. Otherwise truncated but good utterances will not be marked in any way. As with word fragments the full word should appear in the transcription, not a text fragment<sup>8</sup>. Please note that the tilde symbols '~' must NOT be separated from the truncated word by a white space<sup>9</sup>.

## 5 Recording quality assessment (optional)

Following the Dutch Polyphone [2] experience, the quality of each utterance can be assessed, independently of (and after completion of) the transcription. This information can be stored in the label file by using the mnemonic ASS, but it may also be stored in a separate log file. Four different assessment types will be used:

### GARBAGE

File to be marked as "Garbage" if

- it contains only background noise
- it contains noises produced by a non-co-operative subject

Garbage files are not counted as contribution to the database.

### NOISE

Files are marked "Noise" if they contain clearly audible background noise in addition to the speech. A criterion triggering this rating can be the failure of the recording platform to stop recording after the speaker completed the utterance (the platform can be set to consider two seconds of 'silence' as end\_of\_utterance).

### OTHER

Files are marked "OTHER" if they contain

---

<sup>8</sup> This is different from the SpeechDat(M) specification which used text fragments

<sup>9</sup> This is different from the ATIS convention.

- one or more disfluencies, hesitations or stuttering
- an exceptional pronunciation of a word
- speech that is (partly) unintelligible

## OK

Files were rated OK in all other conditions. Note that OK does not mean that the subject adhered exactly to the prompting text in read items; if (s)he did not hesitate in speaking something else and there is no high level background noise, the item is rated OK. Also, utterances trimmed at the beginning or end are rated OK, provided that the first or last word present in the file are in no way damaged.

If a file can be rated both as "NOISE" and "OTHER", it must be rated "OTHER". Each file must be given exactly one rating.

Note that an automatic program could probably perform this classification based on the transcription, in which case it avoids the need for the transcriber to make this judgement. This simplifies the task and can provide more consistency to these categories. But it does mean that the transcription and non-speech events etc. have to be sufficient and reliable for this purpose.

It is useful to provide an opportunity to make a comment about speaker characteristics (which could be stored once for all calls by that speaker) which are helpful for later analysis and selection of utterances, e.g. foreign or non-native speaker accent; very unclear, quiet or loud speakers; and especially stuttering or other significantly serious production characteristics, significantly poor voice quality, or uncooperative speakers whose data is not useful for training or testing. This could be optionally marked and later included in the label file for the transcriptions or in the speaker table.

## **6 Transcription procedure**

The following recommendations are made for the transcription procedure.

1. Use headsets for transcribers. Each transcriber should have the same type of headsets. Transcriptions should be made in a quiet environment.
2. Compose a fixed set of training sentences for the transcribers. Submit this same set to the validation centre, to be used for instructing the transcription validator.
3. Show waveform of the utterance as well. In this way non speech acoustic events can be located more easily.
4. Use a speech recogniser as a pre processor to find digits and keywords.

5. Hire your transcribers on the basis of half day employment. If they transcribe the whole day long, this will not be beneficial for the quality of the database.

In Dutch Polyphone it proved advantageous to present the transcribers with the prompting text, whenever that was unambiguous. In those items the task of the transcriber was to decide whether the subject had adhered to the prompt without making additional noises. If the speaker deviated from the prompting text, or if there were extra noises, the default transcription had to be edited accordingly. If there were no discrepancies between prompt and speech, the default transcription could be added to the transcription file by hitting <CR>.

## **7 Pronunciation lexicon file**

This specification is based on the EAGLES SLWG Handbook chapter on Lexica. Although the EAGLES specification was meant to be reasonably international it may be that other languages may encounter some issues not well supported by these specifications. These issues should be raised as soon as possible, and be documented for the appropriate language. The requirements for this lexicon are rather simple, and transformation into other forms should readily be accomplished by automatic means if later required.

### **7.1 Objective for the pronunciation lexicon**

As Spoken Language lexica can have different purposes and corresponding structure and contents, it is important to keep the main purpose of these for the SpeechDat project clearly in mind. Quite simply the speech databases are a resource to aid training and testing speech recognisers. The main function of the lexica is to provide a pronunciation dictionary for the corresponding language corpus. The technology developers can associate with each utterance the most probable phonetic transcription in the absence of a detailed, manually provided phonetic transcription of each utterance.

The project is therefore primarily concerned with providing a corpus word list with associated pronunciation transcriptions, including common variants. This is considered to be a minimum resource that can still enable each database to be used to bootstrap a basic speech recognition system using automatic segmentation methods. This will particularly apply to simple isolated word, keyword and short phrase recognition. Fluently spoken, continuous speech utterances introduce many more complex features of coarticulation which require specific, detailed phonological knowledge for each language.

There is no requirement to provide more detailed information, such as grapheme to phoneme rules, coarticulation rules etc. There is also no requirement to supply anything other than a broad class phonetic transcription, based on SAMPA. however it is evident that phonological variations will be encountered for many canonical

transcriptions as a result of regional accent, coarticulation, different speech rates, different productions (over-deliberate/careful/casual, isolated/fluent, read/spontaneous) etc. The basic aim is to provide the most frequent pronunciation and, optionally, commonly encountered variants, for the standard variety of the language considered.

While syntactic and semantic information could be of great benefit in identifying the correct words and corresponding pronunciations for each word occurring in the database, the project has initially decided to take the very simple approach of listing all word forms which occur in the corpus with their corresponding canonical phonemic transcription. The intention is therefore to list all fully inflected forms encountered, and not the stem and morphological components. It will therefore be up to the user to ensure that the correct word forms are identified and applied in using the material.

## 7.2 Handling of orthographic and pronunciation variants

There are 5 special cases that need to be considered when preparing the lexicon:

homonyms	2 words with the same orthography and phonological form but different syntax and/or meaning e.g. “ <i>mate</i> ” (a friend, a chess position, to join together)
homographs	2 different words with the same orthography but different phonology e.g. “ <i>read</i> ” /ri:d/ or /red/
homophones	2 different words with different orthography but identical phonology e.g. “ <i>bred</i> ” and “ <i>bread</i> ”
heterophones	2 or more phonological forms for the same word (multiple pronunciations) e.g. “ <i>either</i> ”- /aɪD@/ or /i:D@/
heterographs	2 or more orthographic forms for the same word e.g. “ <i>recognise</i> ” and “ <i>recognize</i> ”

As we are primarily concerned with providing the full set of possible phonetic transcriptions for each word form, and are not encoding syntax or semantic information, we will minimally supply entries for all word forms encountered, and treat each of the cases above as follows:

- homonyms will not be distinguished by us (indeed they cannot be in our tables);
- homophones are easily distinguished and should be listed if present in the corpus;
- homographs should be listed if present in the corpus;
- heterophones should ideally be listed wherever they are known;
- heterographs should not occur as we have agreed to use standard dictionaries and spellings.

With the exception of homographs and heterophones, there is a one-to-one mapping between each word and its canonical phonemic transcription. However homographs and heterophones cannot be identified simply by processing the utterance

orthographic transcriptions and careful checking will be required to identify these, unless a partner has access to a large lexicon which could be used to support this task.

Homographs can be considered very similar phenomena to heterophones for pronunciation purposes, and could conceivably be merged in the lexicon as we cannot easily distinguish them using the transcriptions. (Strictly speaking they should really have separate table entries.) The proposed schema is the only one for which word lists and frequency counts can be automatically produced (and verified) by simple processing of the orthographic transcriptions. For the purposes of the project this is considered to be sufficient if automatic segmentation is used to match the phonetic transcription with the orthography.

### 7.3 Structure and conventions

The lexicon is an alphabetically ordered list of distinct lexical items (essentially words in our case) which occur in the corpus with the corresponding pronunciation information. The lexicon entries must be identical to the transcription words. Each distinct word should have a separate entry. The pronunciation information should be in the appropriate SAMPA format (See Appendix A). The lexicon is stored in the file <database>\TABLE\LEXICON.TBL. The SAMPA table used must be put onto the CD-ROMS in postscript format, in the file <database>\DOC\SAMPALEX.PS

It is useful to include a frequency count for each entry in the lexicon e.g. to help indicate rare words whose transcriptions are perhaps less important or reliable. As the transcription activity is not phonetically based, we cannot accurately identify occurrences of homographs or heterophones, and so cannot provide frequency counts for these items; in fact the frequency count is orthographic based and not phonemic based. In the example supplied below, we know there are 673 occurrences of the word “*read*” but we don’t know how many “*r i: d*” and “*r e d*” there are. It is recommended that pronunciation transcriptions be listed in decreasing order of frequency of occurrence for the corpus.

It has been decided to supply the lexicon file in ASCII delimited format with a TAB character (ASCII 9) as field delimiter. Further information on the format are in SD 1.3.1. An example of a small English lexicon file is in the table below (“\t” represents the TAB character):

WORD	\t	FREQUENCY	\t	PHONEMIC TRANSCRIPTIONS			
chin	\t	342	\t	t S I n			
cut	\t	122	\t	k V t			
either	\t	234	\t	aI D @	\t	i: D @	
heard	\t	463	\t	h e r d	\t	h 3: d	\t 3: d
mock	\t	123	\t	m Q k			
pin	\t	678	\t	p I n			
read	\t	673	\t	r i: d	\t	r e d	

red	\t	123	\t	r e d				
thin	\t	354	\t	T I n				

**Figure 1 - Example of English lexicon file**

#### 7.4 Apostrophes and hyphens

Words in the lexicon are only split at spaces, not at hyphens, apostrophes etc. This will enlarge the lexicon considerably (especially for French), but it avoids problems in consistency that arise if words are split at other positions as well. It is (probably) more client-friendly, since most speech recognisers will split at spaces. And finally it makes the total of the lexicons of SpeechDat databases more uniform.

#### 7.5 Other optional information

Stress information and word/syllabic/morphological boundary information are each individually optional but, if included, should follow the EAGLES conventions. The convention has been designed to permit the removal of information which is not required, or the selection of useful subsets of the table using simple UNIX tool commands. The use of apostrophe for primary and secondary stress permits simple generalisation over both.

Word in compounds	#
word in phrases	##
morpheme	+
syllable	. (period)
primary stress	' (quote)
secondary stress	" (two single quotes)

**Table 1 - Conventions for marking boundaries and stress**

- The boundaries # and ## each imply coextensive + and . boundaries.
- Where + and . boundaries are coextensive, . is written before +
- The stress marks ' and " are written immediately before the vowel, not before the syllable

The following table reproduces an example containing all of these for a German lexicon example.

Word	Freq.	Transcription
Angst	23	? ' a N s t
Annahme	25	? ' a n # n ' ' a : . m + @
Apparat	64	? a . p a . r ' a : t
April	68	? a . p r ' I l
Aprilwoche	23	? a . p r ' I l # v ' ' O . x + @

Arzttermin	78	? ' a 6 t s t # t E 6 . m ' ' I : n
Aschermittwoch	45	? ' ' a . S 6 # m ' I t # v ' ' O x
August	45	? a U . g ' U s t
Augustwoche	23	? a U . g ' U s t # v ' ' O . x + @
Ausweichmöglichkeit	9	? ' a U s # v ' ' A i C # m ' ' 2 : k . + I I C . + k a I t

The following table contains just the plain SAMPA representation for the above.

<b>Word orthography</b>	<b>Freq.</b>	<b>Transcription</b>
Angst	23	? a N s t
Annahme	25	? a n n a : m @
Apparat	64	? a p a r a : t
April	68	? a p r I I
Aprilwoche	23	? a p r I I v O x @
Arzttermin	78	? a 6 t s t t E 6 m I : n
Aschermittwoch	45	? a S 6 m I t v O x
August	45	? a U g U s t
Augustwoche	23	? a U g U s t v O x @
Ausweichmöglichkeit	9	? a U s v a I C m 2 : k I I C k a I t

## **7.6 Language-specific issues**

Conventions required to deal with language specific issues must be clearly documented. However no deviations from the common conventions are allowed, unless agreed with all partners.

## **Appendix A SAMPA-computer readable phonetic alphabet**

*(text copied from <http://www.phon.ucl.ac.uk/home/sampa/home.htm>)*

SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987-89 by an international group of phoneticians, and was applied in the first instance to the European Communities languages Danish, Dutch, English, French, German, and Italian (by 1989); later to Norwegian and Swedish (by 1992); and subsequently to Greek, Portuguese, and Spanish (1993). Under the BABEL project, proposals are being actively considered for extending it to Bulgarian, Estonian, Hungarian, Polish, and Romanian. Under the aegis of COCOSDA it is hoped to extend it to cover many other languages (and in principle all languages).

SAMPA basically consists of a mapping of symbols of the International Phonetic Alphabet onto ASCII codes in the range 33..127, the 7-bit printable ASCII characters. Associated with the coding (mapping) are guidelines for the transcription of the languages to which SAMPA has been applied. Unlike other proposals for mapping the IPA onto ASCII, SAMPA is not one single author's scheme, but represents the outcome of collaboration and consultation among speech researchers in many different countries. The SAMPA transcription symbols have been developed by or in consultation with native speakers of every language to which they have been applied, but are standardised internationally.

A SAMPA transcription is designed to be uniquely parsable. As with the ordinary IPA, a string of SAMPA symbols does not require spaces between successive symbols.

SAMPA has been applied not only by the SAM partners collaborating on EUROM 1, but also in other speech research projects (e.g. BABEL, Onomastica) and by Oxford University Press.

In its basic form SAMPA was seen as catering essentially for segmental transcription, particularly of a traditional phonemic or near-phonemic kind. Prosodic notation was not adequately developed. This shortcoming has now been remedied by a proposed parallel system of prosodic notation, SAMPROSA. It is important that prosodic and segmental transcriptions be kept distinct from one another, on separate representational tiers (because certain symbols have different meanings in SAMPROSA from their meaning in SAMPA: e.g. H denotes a labial-palatal semivowel in SAMPA, but High tone in SAMPROSA).

A recent proposal for an extended version of the segmental alphabet, X-SAMPA, would extend the presently agreed conventions so as to make provision for every symbol on the Chart of the International Phonetic Association, including all diacritics.

In principle this would make it possible to produce a machine-readable phonetic transcription for every known human language.

The present SAMPA recommendations (as devised for the basic six languages) are set out in the following table. All IPA symbols that coincide with lower-case letters of the Latin alphabet remain the same; all other symbols are recoded within the ASCII range 37..126. In this current WWW document the IPA symbols cannot be shown, but the columns indicate respectively a SAMPA symbol, its ASCII/ANSI number, the shape of the corresponding IPA symbol, and the symbol's meaning or use.

### **Vowels**

A	65	script a	open back unrounded, Cardinal 5, Eng. <i>start</i>
{	123	æ ligature	near-open front unrounded, Eng. <i>Trap</i>
6	54	turned a	open schwa, Ger. <i>Besser</i>
Q	81	turned script a	open back rounded, Eng. <i>Lot</i>
E	69	epsilon	open-mid front unrounded, C3, Fr. <i>Même</i>
@	64	turned e	schwa, Eng. <i>Banana</i>
3	51	rev. epsilon	long mid central, Eng. <i>Nurse</i>
I	73	small cap I	lax close front unrounded, Eng. <i>Kit</i>
O	79	turned c	open-mid back rounded, Eng. <i>Thought</i>
2	50	ø	close-mid front rounded, Fr. <i>Deux</i>
9	57	oe ligature	open-mid front rounded, Fr. <i>Neuf</i>
&	38	s.c. OE lig	open front rounded
U	85	upsilon	lax close back rounded, Eng. <i>Foot</i>
}	125	barred u	close central vowel, Swedish <i>sju</i>
V	86	turned v	open-mid back unrounded, Eng. <i>Strut</i>
Y	89	small cap Y	lax [y], Ger. <i>Hübsch</i>

### **Consonants**

B	66	beta	voiced bilabial fricative, Sp. <i>Cabo</i>
C	67	ç	voiceless palatal fricative, Ger. <i>Ich</i>
D	68	–	voiced dental fricative, Eng. <i>Then</i>
G	71	gamma	voiced velar fricative, Sp. <i>Fuego</i>
L	76	turned y	palatal lateral, It. <i>Famiglia</i>
J	74	left-tail n	palatal nasal, Sp. <i>Año</i>
N	78	eng	velar nasal, Eng. <i>Thing</i>
R	82	inv. s.c. R	vd. Uvular fric. or trill, Fr. <i>Roi</i>
S	83	esh	voiceless palatoalveolar fricative, Eng. <i>ship</i>
T	84	theta	voiceless dental fricative, Eng. <i>Thin</i>
H	72	turned h	labial-palatal semivowel, Fr. <i>Huit</i>
Z	90	ezh (yogh)	vd. Palatoalveolar fric., Eng. <i>Measure</i>
?	63	dotless ?	glottal stop, Ger. <i>Verein</i> , also Danish <i>stød</i>

### **Length, stress and tone marks**

:	58	colon	length mark
"	34	vertical stroke	primary stress
%	37	low vert. str.	secondary stress
`	96	(see note)	falling tone
'	39	(see note)	rising tone

*Note:* The SAMPA tone mark recommendations were based on the IPA as it was up to 1989-90. Since then, however, the IPA has changed its symbols for falling and rising tones. These SAMPA tone marks may now be considered obsolete, having in practice been superseded by the SAMPROSA proposals.

### **Diacritics**

(shown with another symbol as an example)

=n	60	inferior stroke	syllabic consonant, Eng. <i>Garden</i>
O~	126	superior tilde	nasalization, Fr. <i>Bon</i>

Some other pages provide a brief outline of the phonemic distinctions in various languages: Danish, Dutch, English, French, German, Greek, Italian, Norwegian, Portuguese, Spanish, and Swedish.

For queries please contact John Wells at:

Department of Phonetics and Linguistics,  
University College London,  
Gower Street,  
London WC1E 6BT.

Tel: +44 171 380 7175  
e-mail: [j.wells@ucl.ac.uk](mailto:j.wells@ucl.ac.uk)

Last revised 1 November 1995  
<http://www.phon.ucl.ac.uk/home/sampa/home.htm>

## **BIBLIOGRAPHY**

- [1]    Macrophone spec and transcription guidelines
- [2]    Dutch Polyphone spec and transcription guidelines