The CSLU Labeling Guide

T. Lander
Center for Spoken Language Understanding
Oregon Graduate Institute

May 15, 1997

Acknowledgements

The Authors of this Book would like to thank the following people without whose support this work could not have been completed: Dr. Ron Cole, Director of CSLU, for his vision of phonetically labeled speech and the support which enabled the research; Dr. Beatrice Oshika, for linguistic expertise in the development of initial label sets, for editorial comments on this document, and for ongoing support and training of labelers; Dr. Etienne Barnard, for input as head of the Language Identification Project; Dr. Jim Hieronymus, for the multi-language transcription set, Worldbet; Vince Weatherill, for input as Center Manager and initial documentation of conventions; Dr. Mark Fanty, John Pochmara, and Johan Schalkwyk, for the development of the OGI Speech Tools; Mike Noel for input as Corpus Development Manager; and the labelers including: Takayuki Arai (Japanese), Li Jiang (Mandarin), Troy Bailey (Spanish, English), Anne Johansen (Spanish, German), Kay Berkling (German), Dana Mitchell (English), Jim Brennan (English), Victoria Noel (English), Marlyse Cathery (French, English), Katsutoshi Ohtsuki (Japanese), David Cole (English), Kal Shobaki (English), Terri Durham (English), Angie Fujioka (English), Vince Weatherill (English), Alexandra Guerra (Spanish), Amie Wilson (English), Zhihong Hu (Mandarin), Yonghong Yan (Mandarin), and Neena Jain (Hindi).

Preface

The CSLU labeling document was created as a reference manual for the corpus development staff at the Center. As such, it is a "living document" that continues to evolve as languages are added to the database and new problems (and solutions) are encountered.

There are several goals of corpus devlopment at CSLU: to provide training data for speech recognition, to supply training data for automatic language identification, and to offer a body of data to the research community to enable analysis of language at all levels. As our speech data and transcriptions are released to the research community, we have documented our transcription conventions to make transcriptions useful to others.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Overview

1.1 Purpose

The CSLU Labeling Guide is intended to accompany the data distributed by CSLU. It describes the conventions used to transcribe those data.

1.2 Overview

Speech data at CSLU are transcribed at two levels: orthographic and broad phonetic. We produce non-time aligned orthographic transcriptions to provide quick access to the content of an utterance. Some orthographic transcriptions contain markers for word boundaries, to support access and retrieval at the lexical level. Time aligned phonetic transcriptions give a more detailed representation of the utterance to enable phonetic and phonemic analysis.

1.3 Levels of Transcription

Following is a description of the three types of transcriptions distributed by the CSLU.

1.3.1 Non-Time Aligned Orthographic

Non-time aligned word level transcription is designed to indicate the content of an utterance, without reference to time. It is represented in a standard orthography or romanization and distinguishes between speech and non-speech information. Conventions for non-time aligned transcriptions are described in chapter 2. CSLU Corpora that have utilized non-time aligned labeling in part or all of the corpus are: *Spelled and Spoken Names*, *Stories*, *Words, Numbers and Phrases*, *Names*, *Numbers*, *22-Language*, *Cellular Speech* and *Alphadigit*.

Non-time aligned word transcriptions are generally created in a text editor, so the text files appear much the same as the text of this document, but without punctuation, capitalization or indentation.

A non-time aligned transcription of the phrase "Oh, my dream house..." might look like the following:

<pau> ohh my dream house

The non-speech label indicating a pause is abbreviated and appears in pointy brackets. The exclamation "oh" is transcribed *ohh* to disambiguate between the letter *o* and the number *oh*.

1.3.2 Time Aligned Orthographic

An example of a time aligned word transcription appears in Figure 1.1.

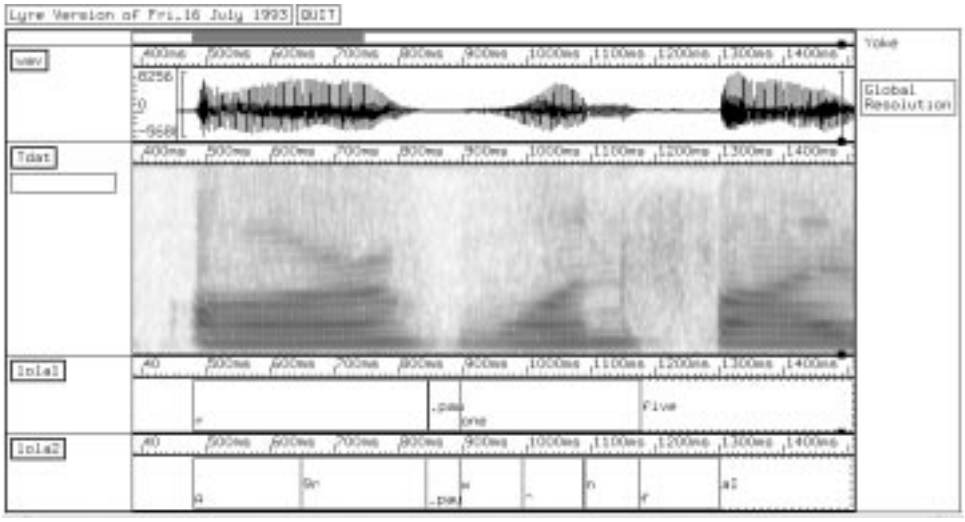


Figure 1.1: Orthographic and phonetic time alignment of the alpha-digit sequence, "r one five"

The text file corresponding to the time aligned word level transcription in Figure 1.1 appears below. The first two lines of the file contain the header, and the remaining lines contain the words spoken preceded by the start and end times in milliseconds.

```
MillisecondsPerFrame: 1.000000
END OF HEADER
481 846 r
846 894 .pau
894 1176 one
1176 1507 five
```

Time aligned word transcriptions are also represented in a standard orthography or romanization. Speech and non-speech phenomena are distinguished. The transcriptions are aligned to a waveform, by placing boundaries to mark the beginning and ending of words. In addition to the specification of boundaries, this level of transcription includes additional commentary on salient speech and non-speech characteristics, such as glottalization, inhalation, and exhalation. Conventions used to produce time aligned orthographies are in chapter 2. CSLU Corpora that have utilized time aligned word transcriptions in part or in all of corpus are: *English Census Corpus* and *Stories Corpus*.

1.3.3 Phonetic

An example of a time aligned phonetic transcription appears in Figure 1.1.

Following is the text file containing the time aligned phonetic transcription. The header occupies the first two lines, and the following lines contain the start and end times in milliseconds and the phonetic labels corresponding to the marked segment.

```
MillisecondsPerFrame: 1.000000
END OF HEADER
480 649 A
649 844 9r
844 895 .pau
895 994 w
994 1089 ^
1089 1174 n
1174 1300 f
1300 1506 aI
```

Our level of labeling

The aim of phonetic/phonemic transcriptions is to represent the phonetic content of an utterance at a given level of detail. The phonetic transcription conventions used to label at the CSLU are found in Chapter 5 and following chapters. We label with Worldbet [**?**], and the result is a transcription containing mostly phonemic base labels and limited phonetic detail. The phonetic detail is made explicit by use of diacritics. Phonetic phenomena that we transcribe include excessive nasalization, glottalization, frication on a stop, centralization, lateralization, rounding and palatalization. See chapter **??** for a description of all salient phonetic features captured in our transcriptions.

Other levels of labeling

Phonemic transcriptions capture only those distinctives that are contrastive in a certain language, while phonetic transcriptions explicitly mark allophonic variation. Consider the three contrastive plosives occurring in English. For a purely phonemic transcription one would posit only the three voiceless stops p, t, and k, and would not differentiate the phonetic variation existing in certain contexts (like /k/ in "kite" as opposed to /k/ in "queue.") As many phonetic realizations are predictable by phonological rule, they do not always need to be transcribed explicitly.

The TIMIT label set, which is used to transcribe English, is an example of a phonemic transcription set that captures mainly language specific distinctives. (TIMIT has a few labels that are not phonemic, such as the reduced vowel label **ix**.) OGIbet is based on TIMIT. It was formerly used at the CSLU for broad-phonetic transcriptions. OGIbet is mainly phonemic, but by use of diacritics and some phonetic base symbols, the transcriber is able to capture more phonetic features of the speech signal.

If one wishes to capture fine phonetic detail, the IPA conventions are a candidate. The IPA has long been recognized as the standard for phonetic transcriptions. "The IPA is intended to be a set of symbols for representing all the possible sounds of the world's languages." The choice of symbols is "usually guided by the principles of phonological contrast." [**?**]

If one is involved in producing multi-language transcriptions, it is not sufficient to have symbol sets that capture only distinctives existing within a given language. Worldbet [**?**], developed by Dr. James Hieronymus, was adopted at CSLU because of the need to capture sounds in many different languages. Worldbet contains labels which are similarly defined across all languages. Like the IPA, Worldbet is robust enough to capture all contrastive sounds in the worlds languages, as well as salient phonetic distinctions.

A correspondence can be made among a standard phonological transcription such as TIMIT, a detailed phonetic transcription such as the IPA, and the multi-language motivated Worldbet.

CSLU Corpora that utilize time-aligned phonetic transcriptions in part or all of the corpus are: *Spelled and Spoken Names Corpus*, *OGI Multi-Language*, *Stories*, *Names* and *Numbers*.

1.3.4 IPA versus Worldbet

Both the IPA and Worldbet are functional and useful for multi-language transcriptions. The IPA is the commonly recognized standard for phonetic transcriptions. Worldbet is relatively new and less well known. It is designed upon different principles than the IPA. Many of the differences between Worldbet and the IPA are purely academic and do not affect the training of spoken language systems at all. Still, I will discuss the major differences here for those who may be interested.

- The IPA uses special (non-ASCII) symbols. Diacritics appear as subscripts or superscripts. Special fonts must be installed in order to produce and view symbols. IPA fonts are widely available and there are nice fonts available free of charge through the SIL (Summer Institute of Linguistics). See their web page:
  http://www.sil.org/computing/sil_computing.html#silsoftware.

- Worldbet uses an ASCII based character set, and can be typed on a standard keyboard. There is no need for special tools to view or produce symbols. Diacritics are separated from the base label by use of an underscore.

- IPA symbol choice is based mainly on phonological contrast. For example, the dental place of articulation on the Spanish stop could either be transcribed with a diacritic, or not transcribed at all. It is not important to mark the dental place if one considers phonological contrast alone, because there is no language for which dental and alveolar stops are contrastive; the label **t** suffices to mark both the dental and alveolar /t/ without conflict.

- Worldbet symbol choice is based not only on phonological contrast, but on descriptive principles. For example, a Worldbet base label explicitly contains information about aspiration or place of articulation in situations where the IPA might not transcribe this explicitly. For example, Worlbet uses **th** in English (aspirated alveolar plosive), **t[** in Spanish (dental unaspirated plosive) and **t** in French (unaspirated alveolar plosive), where the IPA would generally use **t** for all three cases.

- One can transcribe either phonemically or phonetically with the IPA, but there is no mechanism for transcribing both levels at the same time.

- Worldbet attempts to transcribe both phonetic and phonemic levels of labeling in a single tier. Base labels are usually phonemes, with diacritics showing phonetic detail. If a phonemically voiceless alveolar fricative /s/ becomes completely voiced during articulation, the

segment would be transcribed **s_v** in Worldbet, but **z** in the IPA. The Worldbet scheme has the advantage of retaining any length distinctions left on the voiced /s/, but the disadvantage lies in lack of consistency. (Some base labels are allophones, not phonemes, so a simple stripping off of diacritics does not necessarily leave one with a purely phonemic transcription.) There are ways to get around the inconsistencies, and a smart diacritic stripper is the first step. Such a tool is under development, and is anticipated to be a part of the OGI Speech Tools in the near future.

Chapter 2

Word Level Conventions

2.1 Purpose

Word level transcriptions provide quick, lexical access to an utterance as well as limited paraspeech data. The purpose of word level transcriptions is to transcribe the words spoken as they would appear in a standard dictionary. For exceptions see Section 2.4.1.

Little information regarding dialect or ideolect is encoded in the transcription. Words that are mis-pronounced can be tagged with <**pron**>. Entire files can be flagged with **(pron)** if a strong accent is detected. Apart from this we have no convention for specifying the actual pronunications of words. If more detailed information is needed one could consider creating phonetic transcriptions.

In addition to transcription of words, word level transcriptions record non-speech sounds such as background noise (.bn), line noise (.ln), and breath noise (.br). The non- speech labels are discussed in Chapter 3.

2.2 Overview

Chapter 2 describes conventions used to transcribe speech at the time aligned and non-time aligned word levels. Time aligned and non-time aligned transcriptions differ in two ways: First, time aligned transcriptions contain boundary markers to show where segments begin and end, and non-time aligned transcriptions contain no reference to time. Second, time aligned non-speech labels are preceded by dots (**.ns**) whereas they appear in pointy brackets in non-time aligned transcriptions (<**bn**>).

2.3 Non-time Aligned Conventions

This section covers conventions for non-time aligned word level transcription. For time aligned word level conventions see section 2.5.

2.3.1 Enclosing brackets: <>, (), []

We want our transcriptions to be computer parsable. This means the structure needs to be predictable. There are two possible structures in word level transcriptions.

1. ***[ ]*word*[ ]*< >**

2. **(tag)**

This syntax will become clearer as you read.

Pointy brackets

Non-speech labels are usually enclosed in pointy brackets. See Chapter 3, Table 3.1. Nothing should appear between pointy brackets except for these non-speech tags.

Parentheses

Enclose a non-speech tag in parentheses when the non-speech event happens during the entire file. Put tags in parentheses at the beginning of the file.

Note the following example, appropriate for an Austrailian speaking in a crowded room:

(pron) (bs) i like vegimite

The **(pron)** tag indicates non standard pronunciation, in this case the Australian dialect, and **(bs)** indicates background speech that can be heard during the entire file.

Square brackets

Square brackets are used to contain the transcriber's best guess when speech has been cut off or when there has been a false start. If a text appears in square brackets, there should be no acoustic evidence for that utterance. When transcribing text in square brackets you are using your knowledge of the language to supply information that is not actually in the signal.

Note the following example:

the colors of the flag are red white and b[lue]*

Here, although the speaker was cut off while uttering the word blue, as speakers of the language we can confidently supply the information missing from the signal. If a confident guess cannot be made about the cut off speech, no guess should be made. See 2.3.6.

my name is sp*

2.3.2 Non-Speech Labels: Quick Summary

1. Any word or string of characters not appearing within pointy brackets or parenthesis is assumed to be a word spoken by the caller. This will be referred to as a "foreground event."

2. Words spoken by the caller should never appear in pointy brackets or parenthesis. Inferred speech, speech that was intended, but not actually uttered, may appear in square brackets. See Section 3.

3. Putting a non-speech tag in parenthesis means that the non-speech event happens during the entire file. The parenthesized tag(s) must appear at the beginning of the transcription.

4. Connecting non-speech labels to words signifies that the non-speech event happens at the same time the word was spoken. The tags should always be attached to the end of the word they modify; order of tags is not important. See section **??**.

5. Non-speech labels should not be connected to one another unless they are also connected to a word. Non-speech labels in parentheses should never be connected to anything.

2.3.3 Ordering of symbols

There is actually a great deal of information captured in the word transcriptions at the non-time aligned level. In the absence of time markers, we show simultaneous events by connecting non-speech labels to the word spoken in the foreground; there are over 20 different types of non-speech events. With all this information, transcriptions must be structured in order to be parsable. With this need for structure in mind, lets revisit the ***[ ]*word*[ ]*< > format.

"Word" is the only required element. Neither the asterisks nor the []'s are required, but if the [] are present, the asterisk is required. Non-speech information appearing in pointy brackets *always* appears last in the word. The attachment of a non-speech label means nothing as far as where in the word the event happened, whether closer to the beginning or end, rather the attachment only indicates simultaneity. This enables the non-speech label to always appear at the end of the word.

The following are all valid transcriptions:

english (basic word)
***english** (part of sound "e" cut off)
english* (part of sound "sh" cut off)
***[eng]lish** ("eng" cut off)
eng[lish]* ("lish" cut off)
english<ln> (simultaneous line noise)
eng[lish]*<ln> (ln simultaneous with "eng", "lish" cut off)
***[eng]lish<ln>** (ln simultaneous with "lish", "eng" cut off)
english<ln><bn> (two simultaneous events)

2.3.4 Simultaneous Sounds

It is difficult to show overlapping phenomena in transcriptions. Following is a description of the convention used to distinuguish between sequential and simultaneous sounds at the non-time aligned level. See section 2.5.2 for conventions at the time aligned level.

1. Put a space before and after pointy brackets of non-speech labels when the sounds do not occur simultaneously. For example, if you hear a noise between the words "to" and "fly" you would transcribe it:

   i like to <bn> fly kites

2. If noise is heard while the talker is saying a word, connect the non-speech label to the word that was affected.

i like to fly<bn> kites

Here background noise could be heard while the speaker was saying the word "fly."

3. If background noise or other non-speech phenomena can be heard throughout the entire file transcribers can illustrate this by typing the appropriate non-speech label at the beginning and at the end of the file.

   Consider the following example, in which this sentence is assumed to be the entire file:

   <**bs**> i was born in san paulo
   and i wanted to learn to ski from the
   time i was old enough to speak <**bs**>

   This convention aleviates the need to put a non-speech label in every pause or silence, although it introduces ambiguity in the case where noise only occurred at the beginning and end of the file.

2.3.5 A Note About Connecting Non-speech Labels

The non-time aligned non-speech labels can be divided into three categories:

- Those that must be connected to a word: <**long**>, <**sp**>, <**asp**>, <**alt**>, <**pron**>.
- Those that must not be connected to a word (cannot indicate simutineity): <**br**>, <**burp**>, <**cough**>, <**ct**>, <**ls**>, <**pau**>, <**sneeze**>, <**sniff**>, <**tc**>.
- And those than can be used either way: <**beep**>, <**blip**>, <**bn**>, <**bs**>, <**fp**>, <**laugh**>, <**ln**>, <**nitl**>, <**ns**>, <**uu**>, <**yawn**>, <**vs**>.

2.3.6 Cut Off Speech

Cut offs: Non-Time Aligned
The asterisk is to be used in the word file at the time aligned and non-time aligned levels when speech is cut off. By "cut off" we mean the person was either already speaking when the system began recording, or that she/he was still speaking as the system stopped recording.

Note: There must be evidence in the waveform that a cut off has occurred. Do not use the asterisk when you think the person wanted to say something more, but did not atually say anything.

When a person is cut off in the middle of a word, the waveform will only contain the part of the word being uttered when recording either stopped or started. Often the transcriber can supply the whole word from contextual clues, but other times the word cannot be ascertained. Transcriptions will be different depending on the transcribers' ability to supply the missing part of the word. If part of the word can be ascertained, it should appear in square brackets [ ] to disambiguate it from the non-speech labels which appear in pointy brackets <>. Consider the following valid transcriptions:

1. ***[m]y name is jerry**
2. **jerry is my na[me]***
3. ***my name is jerry**
4. ***isideck<sp> is my mother's name**

Notice in example (1) that the asterisk is placed at the beginning of the first word in the transcription to indicate that the word at the beginning of the file was cut off. In example (2) the asterisk has been placed at the end of the last word in the file to indicate that the speech has been cut off at the end of the word. Example (3) is like example (1) in that the beginning of the file has been cut off.

Note also the bracketed information in examples (1) and (2). In example (1), **[m]** signifies that the "m" sound was cut off, but the entire word could be supplied by the transcriber. Note also that even though the sound the transcriber heard was "ay," (or in Worldbet, aI) the trancription is ***[m]y**, which reflects the standard orthography. If the transcriber had thought instead that the word was "buy" the transcription would have been ***[b]uy**, again reflecting the standard orthography characteristic of word level transcriptions.

In transcription (2) the last sound in the word "name" was cut off. Again, this was a familiar word to the transcriber, so she/he was able to supply the final two letters of the word "name."

In the examples (3) and (4) there is no information in brackets, but for different reasons. Let's take example (3) first.

Transcription (3) signifies that only part of the "m" was cut off, but part of "m" could still be heard by the transcriber. In this respect, (3) is most like (1); the only difference is that in (1) the sound m could not be heard at all, whereas in (3) part of the m could be heard.

In example (4), nothing appears in brackets. This is clearly a case of a cut off word, as the asterisk is the first character in the file. In this transcription, since the utterance was of an unfamiliar name, the trancriber was unable to supply the rest of the word. The label <sp> was added to show that it was an unfamiliar word whose spelling could not be ascertained.

Placement of Asterisk

For placement of asterisk, see 2.3.3
In summary:

- When a word is cut off but the transcriber understands what the speaker was intending to say, the part that was actually uttered is transcribed (not in brackets), and the part that the transcriber supplies appears in square brackets. (See Examples 1 and 2 above)
- If only part of a single sound is cut off, then nothing should appear in brackets, but the asterisk will appear at the beginning (or end if it is cut off at the end) of a word—with no intervening space–to indicate that part of the initial (or final) sound in the word was cut off. (Example 3 above).
- If instead only part of an *unfamiliar* word can be heard due to a cut off, the part that can be heard is "sounded out" so to speak, and the asterisk is added to the word.

2.3.7 False Starts

The asterisk is also used to label false starts. A false start is a certain type of disfluency in which the speaker begins to say something and then stops in the middle of the utterance. Consder the following examples:

1. **i'd like to thi[nk]* read about it for a bit**
2. **when i go to flori[da]* i mean tennessee i'll look you up**
3. **trimson cri[ed]* i mean crimson tide**
4. **well i don't br* even want to talk about it**

Notice in the first three examples the transcriber felt confident enough of the context to hazard a guess at what the person intended to say. The first two examples are straightforward, the third is what is known as a "spoonerism" and in the fourth, the false start **br*** was not traceable to a word, and hence no information appears in brackets.

2.3.8 Why distinguish between "Cut off" and "False Start?"

Basically the conventions for false starts and cut off speech are the same, although the information yielded by the various transcriptions is different in a significant way.

- Cut off speech at the beginning of a file indicates "barge in" or interruption of the prompt by the speaker. Barge-in can be a significant and non-trivial problem for speech researchers.
- False starts within the file indicate interesting information about the speech at the higher, discourse, level. The type or frequency of false starts might be of interest to linguists or others who want to study disfluencies in continuous speech.
- Cut off speech at the end of a file indicates either: (a) utterance detection was not working properly (b) the speaker chose to talk longer than the allowable time. Both of these issues are significant. The first is important if the system designers were indeed employing a technique to detect when a person has stopped talking. The fact the the file was cut off is a good indication that the technique could use some tuning up. The second issue could be significant for the designers of the interface, who, in order to ensure the interest and cooperation of the caller, seek to develop friendly and natural interfaces. If each person who calls gets cut off by a loud beep when they answer certain questions, perhaps developers would want to consider changing the recording parameters so that the calling experience is more satisfying for the participant. This will make the caller less likely to get frustrated and hang up before the call is completed, representing a loss of valuable data for the researcher.

2.4 Miscellaneous Details

This section describes conventions to be followed at the time aligned and non-time aligned levels.

Table 2.1: List of Exceptions

transcription	example utterance
gonna	i'm gonna go
whacha	whacha doin
wanna	i wanna see
gotta	i gotta go now
kinda	yeah kinda
lota	there's a lota money
nother	that's a whole nother story

Table 2.2: Filled Pauses

transcription	meaning
hmm	thinking word
uh	thinking word
uhm	thinking word
mm	thinking word
mm hmm	yes
hmm mm	no
mm	mm
no nuh uh	no
uh huh	yes
huh uh	no
uh uh	no

2.4.1 Exceptions to "dictionary form" rule

Word transcriptions at the time aligned and non-time aligned levels should contain the words the speaker said. The words should appear in citation form (as they would appear in a dictionary). We know that in continuous speech people often delete or transpose syllables and that coarticulation can alter the percept of a word. In order to simplify the orthographic transcription process and to minimize redundancy with phonetic transcriptions, we have opted to ignore most such articulations, especially those which follow the natural phonological processes of the language.

The words in Table 2.1 are exceptions to the citation form rule. Their dictionarly entry form does not seem to be consistent.

If callers utter the words in Table 2.1, or words of the same form (i.e., *of* becomes **a** in **kinda, sorta, lota**; *going to* becomes **gonna**; *what are you* becomes **whacha**; *want to* becomes **wanna**; *got to* becomes **gonna**) transcribe them as they appear in the list above. Apostrophes are not used to indicate the omission of one or more letters. See Section 2.4.3. The words in Table 2.1 are given special consideration due to their frequency, their widespread acceptance in informal speech, and the significant acoustic variation from the dictionary pronunciation.

Another set of exceptions are the filled pauses. The filled pauses are listed and defined in

Table 2.3: Example transcriptions

actual utterance	**transcription**
it's 'bout time	it's about time
whacha doin'	whacha doing
walkin'	walking
nuts n bolts	nuts and bolts
nuts an bolts	nuts and bolts

Table 2.2. They are also described in Chapter 3.

Table 2.3 contains some example transcriptions for potentially confusing cases.

In summary, if the word is not of the type listed in Table 2.1 or 2.2, but it is clearly an intelligible word, it should be transcribed using standard spelling, as displayed in Table 2.3. If the pronunciation is noticibly different from the "normal" or standard dictionary pronunciation, and it is not simply a result of natural sound change in the language, consider using the <**pron**> tag (Chapter 3). Finer detail, if desired, should be captured in a more detailed transcription, such as a phonetic transcription.

2.4.2 Capitalization

There is no capitalization in English orthographic transcriptions. [1]

2.4.3 Punctuation

The only punctuation mark allowed in English is the apostrophe. The apostrophe is used in the following three instances:

- To indicate possession: **she has susan's slide rule.**
- For contractions: **don't take that it's mine**
- Spelling: **my name's got two r's in it**

The apostrophe is not used for the omission of letters in words like *walkin'* or *'bout*. See Table 2.3.

Periods, commas, dashes, hyphens, semicolons and other punctuation marks are omitted. Square and pointy braces are used in transcriptions for non-speech events, or predictable speech that was cut off. See 2.3.1.

The apostrophe is not used to indicate omission of letters as in **walkin'**, *walking*, or **'bout**, *about*.

2.4.4 Transcribing Spelled Letters

When a caller spells a name, recites the alphabet, or says an abbreviation (see below) transcribe the letters spoken with spaces to separate them:

[1] See Chapter 4 for capitalization rules in other languages. Note Section 4.5.7.

A spelled name
h a r r y

The alphabet
a b c d e f g h i j k l m n o p q r s t u v w x y z

An abreviation or acronym
o h s u (the hospital)
c d player

Note the following transcription for double letters:

my name is terri spelled t e and then two r's i

Transcriptions used to contain dashes to indicate that the speaker did not pause between the letters, but currently that is not done. Without access to a spectrogram, transcribers are not able to make the distinction consistently.

2.4.5 Suprasegmentals

Currently there are no specific notations for suprasegmentals. Pauses are marked, but by convention **.pau** is a period of relative inactivity in the waveform (at least 1000ms long) rather than a discourse unit in the stricter sense. See Section 3.18.

2.4.6 Transcribing Numbers

Write numbers as words. 1959 is transcribed, **nineteen fifty nine**. Dashes (as in twenty-nine) are not necessary in English transcriptions.

2.4.7 The Letter, Number and Exclamation, "o"

The number zero, when spoken as, "o," is transcribed as **oh** . When the speaker is spelling the letter o, write simply **o** ; when "oh!" is said as an exclamation, it is transcribed, **ohh**.

2.4.8 Okay

The confirmation **okay** can typically be written **okay** or **OK** in standard English orthography. Because we do not use capital letters in transcriptions, we will transcribe the word **okay**.

2.4.9 Discarding Files

The criteria for discarding files is task dependent. In general, we discard files in which the caller has hung up without speaking, in which only line noise can be heard, or in which there is no useful speech. Be sure to check the criteria for discarding for your particular application.

The procedure of discarding files varies with the task as well. Different scripts may run slightly differently. The process generally involves moving the file to another directory where it will undergo no more processing. Discarded files are not included in the release of speech data files to researchers.

2.5 Time Aligned Conventions

2.5.1 Word Box Conventions

- **Overlapping boundaries:**
  Lola boxes which have overlapping boundaries are illegal. Geminates will be split where the waveform shows change. If the waveform or spectrogram manifest no particular changes, the geminate will be split in half.

- **Extension of Labels:**
  Labels should only be extended from left to right. If a box is originally created at 30ms, the left boundary will always remain at 30ms, unless the transcriber physically moves the left boundary. Thus, if one drags the right boundary past the left boundary, and sets the right boundary at 10ms, the result will be a backwards label box and resulting error in the lola file.

- **Closures:**
  A closure is the articulation which begins the phonemes: p, t, k, b, d, g, ch and jh. When a word begins with any of these sounds, look for spectral evidence to signal the beginning of the closure. (For the voiced closures of b, d, g and jh, there is normally an appearance of F1 in the spectrogram preceding the burst. For voiceless closures, there is sometimes a small lipsmack which signals the closure of the articulators.) If there is no spectral evidence which signals the beginning of the closure, set the word label boundary 50ms before the burst. This length was chosen as an average length of word initial stop closures in english. For words ending with closures, extend the boundary of the word 100ms after the energy of the preceding phone stops. 100ms was chosen as an average length for word final closures in english.

- **Unrereleases:**
  If a stop comes at the end of a word such as *"cat"*, or *"pot"*, some speakers will release the plosive with a burst (which is visible in the spectrogram), and some speakers will not release the closure. If the final burst is imperceptible, set the right boundary of the word 100ms after the energy in the preceeding phone stops. If the burst is perceptible, extend the right boundary of the word to the end of the burst.

- **Cut off speech**
  For time aligned labeling, transcribe the part of the word that was spoken followed by an asterisk in the word box. Rather than supplying the intended utterance within pointy brackets, write the complete word in the comments window. If the entire word is not known, write the label **\incomplete**, in the comments file. Likewise, if the utterance is ambiguous or unintelligible, use the label **.uu** in the word box, and write the label **\incomplete**, in the comments file.
  See 2.3.6

  Note: There must be evidence in the waveform that a cut off has occurred. Do not use the asterisk unless there is a signal in the waveform, such as a high energy level at the end of the waveform. Do not use the cut off diacritic in cases where you know the person wanted to say something more, but before he/she was able to utter anything the system stopped recording. The speaker is only considered to be cut off when she/he has actually begun to say something, and not only when the transcriber thinks he/she might be about to say something.

2.5.2 Comments Box Conventions

- **Actual vs Citation Pronunciations**
  The comments window at the time aligned word level is used to note actual pronunciations (i.e. partial pronunciations or slang), the mode in which the caller is speaking, and important details which clarify the transcription. All comments should be entered as one word in the comment windows. For example:

  \glot \whisper \outbreath

  Most files will contain one or more comments. For some time aligned word labeling, no comments window will exist. The presence of the comments window is dependent upon the needs for the particular corpus being labeled. If the comments window is not present, disregard conventions pertaining to it, and follow only those associated with the word box.

- **Backslashes**
  Every comment that is not a transcription of something the speaker actually said (as opposed to the dictionary form of the word appearing in the word box) should be preceded by a backslash. For example, the comments **\glot \whisper \outbreath** and **\extraneousnoise** should all be preceded by backslashes.

- **Two Types of Lola Boxes**
  There are two types of label boxes made in the comment window: one kind spans the phenomena it represents and the other simply marks the beginning or end of the event. The former labels should have manually extended right boundaries which span the length of the phenomenon. For the latter, a label box is created and the right boundary is not extended. These latter types of comments usually mark the beginning and end of a long segment, ie, read speech or extraneous noise.

- **Manually Extended Boundaries**
  When the label used is the type which spans a certain phenomena, rather than simply identifying the beginning and end, the right boundary should be manually extended to its proper, time aligned position. When saving the comments file in AutoLyre, use "write lola" rather than "write aligned lola" in order to preserve the position of the labels which were not extended.

Specific Comment Box Labels

- **Actual pronunciations:**
  Word comments help explain the difference between what the speaker actually said and the citation form of the word. For example, if the speaker said "what's dat" write **dat** in the comment box aligned with "that" in the word box. Similary, *"'bout"* written **bout**; etc.. (This information is redundant with the phonetic transcription, and need not be noted unless the pronunciation is very noticeable.). If the comment is a transcription of an actual pronunciation, no backslash should precede it.

- **Read Speech:**
  Judgements are entered in the comment box if the speech is extraneous or read speech.

Spontaneous speech is the default; boundary boxes don't need to be set for this speech. We consider read speech to differ from extemporaneous speech. Thus, if the caller is reading, make a boundary box where the read speech begins and label it **\beginread** Make a box at the boundary where read speech ends and type **\endread**.

- **Extraneous speech, noise:**
  Often noise can be heard in the background of a call. The labels **.bn** and **.bs** (used in the word box) are sufficient to cover isolated occurrances of background speech and noise, when extraneous information occurs during a pause, or during a segment of relative silence. However, when an event is segmented, and extraneous noise or speech is heard in addition to a more prominent phenomenon to be labeled in the word box, the following labels should be used in the comments window:

  To mark the beginning of the extraneous speech event, the abbreviated label is: **\BXS**. To label the end of this phenomenon, use **\EXS**. These are abbreviations for "begin extraneous speech" and "end extraneous speech." Likewise, for extraneous noise, the labels are: **\BXN** and **\EXN**. These labels stand for "begin extraneous noise" and "end extraneous noise." Note that these symbols only need to be used when the noise or speech in the background overlaps foreground events, most commonly, speech by the caller.

- **Answers:**
  When the utterances being transcribed are answers to questions, the comments **\BA** and **\EA** should be used. These are abbreviations of the actual comments, beginanswer and endanswer, adopted for ease of typing. Align only the left boundary of these boxes. This convention has only been used for labeling the Census Corpus.

  This convention identifies the part of the response that is intended as the answer to the questions, as opposed to other speaking which may be in the call. If the utterance contains only an exact answer, with no extra speech, there is no need to use the comments BA and EA.

- **Breath:**
  Labelers should use **.br** in the word box when there is a breath. In the comments box the breath should distinguished further: **\inbreath \outbreath** These labels should span the duration of the breath. These may be abbreviated using the convention "ib" and "ob."

- **Abbreviations:**
  Write the label **\abbreviation** for abbreviated forms of words.

- **Whispering:** For whispered speech by the caller, create a box in the comments file which spans the phenomenon and label it **\whisper** If there is someone in the background whispering, this should be labeled following conventions for extraneous speech. The label **\whisper** should only be used in reference to the primary speaker. See section 2.5.2.

- **Mispellings:**
  If you are unsure of the spelling of a word at the time aligned level. use the label **\spelling** in the comments box.

Chapter 3

Description of Non-Speech Labels

Table 3.1 shows a list of non-speech labels. Table 3.2 shows a list of filled pause labels. Table 3.3 shows a list of other words common to the spoken langauge.

3.1 Overview

A description of each non-speech label follows. Keep in mind that these labels are used in two types of transcriptions, non-time aligned and time aligned. At the non-time aligned level the labels appear in pointy brackets (e.g.. <**bn**>); at the time aligned level they appear with a preceding period (e.g.. **.bn**). The terms "word box" and "comments box" refer to time aligned labeling. Conventions specific to one type of transcription are noted, otherwise assume the convention applies to both time aligned and non-time aligned.

NOTE: Most of the non-speech labels refer to a set of sounds. <**bn**> could be a door slamming, music or general environmental noise. <**ln**> could be clicking noise, static or a buzz from the phone line. There is no mechanism in place for further defining these labels. More detailed transcription conventions could be developed if it were important to specifically define each non-speech event in a file.

3.2 Asterisk

The asterisk is used to denote both cut off speech and false starts at the word level. Due to the complex nature of the use of the asterisk and the obvious implications for ordering of symbols, it has been described in detail in Chapter 2, section 2.3.6 for cut off speech, and 2.3.7 for false starts.

3.3 .blip or <blip>

This label is used when the signal goes completely silent for a period of time due to a bad phone line connection. It is only used when parts of words are imperceptible, and the label should always be connected to a word. It is not considered significant if a blip occurs during a pause.

3.4 .bn or <bn>

This label indicates the presence of background noise, which is a broad category encompassing noise produced in the background of the call. Some common examples are typing, music, papers shuffling, babies crying, etc. At the time aligned level, the label **.bn** is used in the word box, and there is a corresponding label to be used in the comments window if the background noise coincides with the foreground noise, such that both can be heard. See section 2.5.2 for coverage of extraneous noise in the comments box.

3.5 .br or

Breath noise, either exhalation or inhalation occurring any place an actual breath is perceived. Breaths often occur between pauses, but speakers frequently exhale word or sentence finally. Aspiration released at the end of a word which is not a part of the phone should be labeled **.br**.

At the time aligned level, when *.br* is labeled, a corresponding label must be placed in the comments window showing whether it is an *inbreath* or an *outbreath*. See page 17 for more details on labeling in the comments window.

3.6 .bs or <bs>

The label **.bs** is to be used in the word box when speech can be heard in the background. If a person is speaking in the background whether the speech can be heard on the television set, on the radio or in person, it would be labeled **.bs**. **.bs** is not used to for side comments made by the caller. The caller's speech is always considered to be a foreground event. **.bs** is not used for babies crying; that would be labeled **.bn**. The use of **.bs** is similar to the label **.bn**, and it likewise has a corresponding label for use in the comments window at the time aligned level when both background speech and a foreground event can be heard. See 2.5.2 for detail concerning extraneous speech in the comments window.

One special use of the label **.bs** is the following:

i have lived in the u s for uhm martha how long have we lived here <bs: ten years > ten years

In this case the caller has asked his wife in the background how long they have lived in the U.S. The wife's answer is audible in the background, and then the caller repeats her answer into the telephone. This type of construction is only allowable with the <**bs**> label.

3.7 .burp or <burp>

A burp.

3.8 .cough or <cough>

A cough.

3.9 .ct or <ct>

A clear throat.

3.10 .fp (or <fp>), uh, uhm, hmm and others (see table 2)

Sounds uttered which act to fill silence while a speaker is thinking are called "filled pauses." The label <fp> is meant for filled pauses that *do not* appear in table 3.2 (uhm, uh, hmm, mm, etc). <fp> is for utterances that are basically unintelligible. It should be used rarely. Here are some examples to clarify:

Person says "uhm i like coffee." transcribed:
uhm i like coffee
(The word **uhm** appears in the table.)

Person says "er well i don't know" transcribed:
er<sp> i don't know
(**er** does not appear in the table, but can be sounded out.)

Person says "<some odd sound> i mean catfish" transcribed:
<fp> i mean catfish
(There is not an adequate word in the table, and it cannot be sounded out.)

3.11 .glot

This label is only to be used in time aligned transcriptions. It should appear in the *comments* box when excessive glottalization occurs in the speech. If the glottal pulses appear to separate noticeably relative to the rest of the speech, the glottalization label should be used.

Glottalization should be included as a part of the word in the regular word box, and the label **\glot**, should appear in the comments window spanning the period of glottalization. As a general rule, glottalization is not noted in the non-time aligned transcriptions because transcribers do not have access to the spectrogram during transcription. If glottalization can be detected using the waveform solely, the transcriber is encouraged to use the label.

3.12 .laugh or <laugh>

A laugh. If the speaker laughs while saying a word, connect the label to the word:
my<laugh> favorite<laugh> dish<laugh> is<laugh> ribs<laugh>
(This is an authentic example.)

3.13 .ln or <ln>

Line noise refers to clicks, buzzes or periods of static caused by the telephone line. If line noise occurs throughout a the call and can be heard when the caller is not speaking, **.ln** should be

labeled where **.pau** would normally be used. You may have a call in which the label **.pau** is not used but is instead replaced with line noise.

3.14 <long>

This label is only used at the non-time aligned level for elongated or drawn out words. If there is a word which is noticeably drawn out, connect this label to the word that is elongated. Do not use <long> for pronunciations in which the speaker has paused between syllables, as in "na ¡pau¿ cy" but only when sounds in the word are drawn out. When a speaker pauses between syllables of a word there is no special convention to mark this; the word should be transcribed as it would normally be, i.e., "nancy."

3.15 .ls or <ls>

This phenomenon occurs often before a breath or an utterance. If the speaker makes a smacking noise with the tongue and lips, segment and label it **.ls**.

3.16 .nitl or <nitl>

The label **.nitl** is used for speech that is in a language other than what the caller was asked to speak. This is referred to as "foreign speech." Foreign speech is particularly common in multi-language transcriptions when the person is spelling words or giving an address. When a person clearly pronounces a word in a foreign language, the label .nitl should be used.

3.16.1 .nitl at the Time Aligned Level

If the transcriber cannot decipher individual words, the label **.nitl** should span the entire part of the waveform containing foreign speech. If the transcriber understands the foreign language, or can at least make a guess at what is being spoken, the time aligned transcription should be aligned to the waveform just as words in the appropriate language are aligned: in other words, when the transcriber can transcribe the individual words, no special transcription conventions are used at the time aligned level for foreign speech.

3.16.2 <nitl> at the Non-Time Aligned Level

For non-time aligned transcriptions, however, speech not in the language can be dealt with in a more specific way. A transcription of a person speaking Spanish (when another language was called for) might look like this:

<nitl: mi casa es su casa>

where the label <**nitl**> is followed by a colon, then a space, and then the words (as close as they can be deciphered) all within the pointy brackets.

Note: This is the only case in non-time aligned transcriptions where speech by the caller can be transcribed within pointy brackets. See also section **??** for use of the colon in pointy braces to further define non-speech symbols.

3.17 .ns or <ns>

The label <ns> is used for sounds that are made by the speaker's mouth but that are not speech, and for which we don't currently have a specific label. Hiccups, yawns and grunts are examples.

3.18 .pau or <pau>

A period of relative quietness in which the speaker stops to think or hesitates before saying a word. The expression *relative quietness* is used here because there is no actual silence in the speech signal due to line and environmental noise.

For consistency and ease of transcription, <pau> should not be marked unless there is a period of at least 1 second (1000ms) where no other event worth transcribing occurs. If there is a 500ms breath and a 500ms pause, the pause should be ignored.

3.19 .pron or <pron>

The label <**pron**> is used for odd pronunciations or pronunciations that differ from the standard dialect. One native American speaker said "nane" when he clearly intended us to understand *nine*. Another non-native speaker consistently pronounced the letter name z *zed* (British pronunciation). Using <**pron**> allows us to transcribe only words that would appear in a standard dictionary, while indicating that the token varies in some significant way from the most common or a predictable pronunciation.

<**pron**> can be used for dialectical variations (such as zed above). Our English transcriptions assume a standard American English dialect. Each language transcribed has a defined dialect and is transcribed from the perspective of that dialect.

<**pron**> may also be used for speech with a heavy foreign accent. Only use <pron> on words that are so heavily accented as to inhibit intelligibility.

To avoid over use of this label, <pron> should not be used either for subtle or for predictable pronunciation variations. A subtle pronunciation might be a non-native English speaker who tends not to aspirate the voiceless plosives *p, t, k* syllable initially in English. A predictable pronunciation is a pronunciation that is predictable by phonological rule, like *mystery* becoming *mystry*, etc.

There are some tricky cases, where large numbers of people tend to pronounce words one way, often considered wrong by others (who think the former speech is full of errors). One example is Latin *et cetera* pronounced by many as **E k s E t(3r ^**. Another is *escape*, pronounced **e k s k ei p** rather than **e s k ei p**. Another example: *espresso* pronounced **E k s p r E s o** rather than **E p r E s o**. Because it is debatable whether or not these so called mispronunciations are based on phonological rules followed by certain speakers and not other speakers, or whether they are ideolectical (I tend toward the former analysis) the safe thing to do is to attach the <pron> tag to such words of questionable pronunciation.

There are no word level transcription conventions to mark dialectical variation specifically. If dialect information is necessary, a verification of the corpus designed to note dialect should be considered as a separate process.

3.20 .sneeze or <sneeze>

This label indicates a sneeze.

3.21 .sniff or <sniff>

.sniff is to span the period in which the speaker sniffs.

3.22 .tc or <tc>

The label **.tc** signifies a clicking noise made with with the tongue.

3.23 <sp>

At times a caller may utter an unfamiliar word. Rather than place the non-descriptive label <**uu**> in the transcription, labelers are encouraged to sound out the word and produce the most likely transcription. <**sp**> should be attached to the end of all such words to indicate the transcriber's uncertainty of the correct spelling of the word. do not insert a space between the word and <**sp**>. For example, for an unfamiliar street name might be: **toogali<sp>**.

<sp> is commonly used with filled pauses that are not specifically listed in Table 3.2. For example:

i was driving urpl<sp> i mean flying to new york

At the time aligned level, the label would be \spelling used in the comments box (See Section 2.5.2. Obviously if the word can be found in a dictionary the correctly spelled word should appear in the transcription without the <**sp**> label.

3.24 .uu or <uu>

Unintelligible speech is a category of sounds that cannot be mapped logically to any known utterance. If the labeler does not understand what the speaker has said, but he or she is sure that it is speech of some sort, the label **.uu** should be used. This label should actually be used rarely, because usually a guess can be made as to the utterance, and in that case, the transcriber should just do the best he/she can to transcribe what is heard. See the description of the cut off label "*" for another use of the **.uu** label.

3.25 .vs or <vs>

The label **.vs** is to be used for high pitched squeaks produced during speech. Voice squeaks are spectrally distinct, having formants which slope upwards and disappear. They generally occur word initially or word finally when the speaker's voice cracks. There is generally a large enough gap between the voice squeak and the rest of the word to consider the voice squeak as a separate entity. Because of the reliance upon the spectrograms for voice squeak detection it is not required at the non-time aligned level. However, if the transcriber can identify a voice

squeak at the non-time aligned level using only the waveform he/she is encouraged to use the label.

The label **.whisper** is used for whispered speech. Connect the tag to the end of each whispered word in non time aligned transcriptions, or make a box in the comments window spanning the period of the whisper for time aligned transcriptions.

In non time aligned labeling, this tag should always be connected to a word unless there is a period of whispering that is unintelligible. In the latter case, the tag <whisper> can appear by itself.

Table 3.1: Non-speech labels for Word Level labeling

Non-Time Aligned	*Time Aligned*		*Description*
	Word Box	*Comments Box*	
<alt>			grammatically altered word due to mixing of two languages
<asp>	.asp		heavily aspirated p, t, or k or puff at end of word that is not a breath *always connect*
*	*		cut off speech *always connect*
< *beep* >	.beep		a beep sound *connected or not*
< *blip* >	.blip		temp signal blip *connected or not*
<bn>	.bn		background noise *connected or not*
		\BXN	begin simultaneous background noise
		\EXN	end simultaneous background noise
 	.br		breathing noise *never connect*
		\inbreath	inhalation
		\outbreath	exhalation
<bs>	.bs		background speech *connect or not*
		\BXS	begin simultaneous background speech
		\EXS	end simultaneous background speech
<burp>	.burp		a burp *never connect*
<cough>	.cough		a cough *never connect*
<ct>	.ct		a clear throat *never connect*
<fp>	.fp		generic filled pause/false start *connect or not*
	.glot		glottalization
<laugh>	.laugh		laughter *connect or not*
<ln>	.ln		line noise*connect or not*
< *long* >			elongated word *always connect*
<ls>	.ls		lip smack *never connect*
<nitl>	.nitl		not in the language *connect or not*
<ns>	.ns		non-speech *connect or not*
<pau>	.pau		a pause or silence *never connect*
<pron>			an odd pronunciation *always connect*
<sneeze>	.sneeze		sneeze *never connect*
<sniff>	.sniff		sniffing sound *never connect*
<tc>	.tc		tongue click *never connect*
<uu>	.uu		unintelligible speech *connect or not*
<vs>	.vs		voice squeak *connect or not*
<sp>		\spelling	unknown spelling *always connect*
<yawn>			a yawn *connect or not*
		\BA	begin answer (Census Corpus)
		\EA	end answer (Census Corpus)
		\abbreviation	abbreviation
<whisper>		\whisper	whispered speech (usually connect)
		\beginread	beginning of read speech
		\endread	end of read speech

Table 3.2: Filled pause labels for English word level labeling

Filled pause label	English translation (if any)	Description
hmm		bilabial, aspirated beginning
uh		centralized vowel, no nasal
uhm		centralized vowel + bilabial nasal
mm		bilabial nasal hum

Table 3.3: Other miscellaneous words in English word level labeling

Word	English translation (if any)	Description
nuh uh	no	alveolar nasal begins first uh
mm hmm	yes	usually rising intonation
hmm mm	no	usually falling intonation
mm mm	no	usually falling intonation
uh huh	yes	usually rising intonation
huh uh	no	usually falling intonation
uh uh	no	usually falling intonation

Chapter 4

Orthographic Conventions For Multiple Languages

4.1 Overview

Chapter 4 discusses special orthographic transcription conventions developed for languages other than English. **A special note:** Transcription conventions for all languages will follow those found in Chapter 2 and 3. Exceptions include conventions classified as English specific or those superceded by something found in Chapter 4.

The outline of the chapter is as follows:

- Non speech events: a discussion of non speech labels developed especially for languages other than English
- Foreign words: <**nitl**> revisted; a discussion of how to handle foreign or accented words, as well as how to handle newly created words which are a combination of two or more different languages: the <alt> tag
- Filled pauses: a discussion of the filled pauses developed for languages other than English.
- Specal Conventions in each language: a subsection is devoted to each language to describe the special conventions used to transcribe that language, including the romanization for that language and diacritic markings.

4.2 Non-Speech Events

The non-speech labels appearing in Chapter 3, Table 3.1 are used in transcriptions of all languages, unless otherwise noted. For rules of use, see chapter 3.

Two labels do not appear in Chapter 3, Table 3.1 because they are not used in English. They are: <**alt**> and <**so**>.

<**alt**> has been added to transcribe words that have been grammatically altered so that they incooporate elements from more than one language. <**alt**> is described in section 4.3 and following. The <**alt**> tag could theoretically be used in English, but we have not run into a case yet that required its use.

The label <**so**> was developed specifically for Cantonese, and is described in section 4.5.2.

Multilanguage labelers should take note that the non speech labels described in Chapter 3 are to be considered a part of the transcription of languages other than English, even though they are not treated specifically in this chapter.

4.3 Foreign Words

Most people participating in our multi-language project speak English (as well as at least one other language) and live in the United States. As a result, the speech we record is often sprinkled with English words. Especially in countries like India where English is one of the official languages, English heavily influences the speech of the people.

Foreign words are marked with the tag <nitl>. An important question is how to classify words as foreign (**n**ot-**i**n-**t**he-**l**anguage).

We consider a word "foreign" if it is not completely absorbed into the language. For example, *buffet* although originally French has been absorbed into the English language, and a tag such as <nitl> would not be necessary (3.16). But perhaps some would consider "program" less French than "logiciel". It is not always easy to decide if a word is foreign or not, but following are some guidelines for transcribing.

4.3.1 Foreign words spoken with a foreign accent

Foreign words pronounced with the phonology of the language of origin require the <nitl> marker. For example:

> A Mandarin speaker says the name McDonald's with American pronunciation.
> Transcription: **mcdonald's<nitl>**

Section 3.16 describes the use of <nitl> in more detail.

4.3.2 Foreign words (names) spoken without an accent

Place names present difficulties for us because often foreign place names are spelled and/or spoken closely to the way they are spoken in the language of origin.

For a speaker of Vietnamese, is a place name such as *New York* a foreign word? Or is *Paris* a foreign word to a Hungarian? To determine whether place names or other words are foreign (require the <nitl> tag), we rely on the pronunciation of the word.

- If the talker pronounces the proper name the way most talkers of the same native language pronounce the name, the word is not considered foreign, and the <nitl> tag is not used.
- if the speaker attempts to pronounce the name as it is pronounced in the language of origin (assuming that is different than the way s/he would normally say the word in his/her own language), the <nitl> tag should be used.
- if the place name or other word is pronounced identically in each language, the <nitl> tag is not needed.

  For example, in Japanese **McDonald's** is pronounced **m a k o d o n a r u d o**, but an American pronunciation would be **m k d A n l= d z**. The former should be

transcribed in the Japanese romanization as **macodonarudo**, but the American sounding pronunciation would be **McDonald's<nitl>**.

Some languages have developed ways of dealing with foreign loan words orthographically. Loan words in Japanese are written in Katakana script. For Japanese it would be more natural to romanize *McDonald's* as **macodonarudo** than to write the word with standard American spelling. However, in French it seems more natural to write "On se email" rather than "On se i-mel." Transcribers should do what seems most natural for the specific utterance and language. Specific conventions to deal with each language will be developed on a case by case basis, and will be described in this chapter.

4.3.3 Foreign words modified grammatically

At times speakers of one language alter the form of a word to fit the grammar of another language. For example, In Bantu languages the plural morpheme is "ba" and the singular suffix is "mu." When discussing the fate of a single bartender, one Swahili speaker referred to the bartender as a "mutender." He altered the English word bartender to fit the grammer of Swahili—a native speaker of Swahili would understand this to mean a solitary person who tends a bar.

Utterances requiring **<alt>** might raise eyebrows, but the meaning is usually clear to a speaker of both languages.

Words of one language that are altered to fit a grammatical pattern in another language should appear with the <alt> tag attached. The example above would be transcribed as **mutender<alt>**, or possibly **mutenda<alt>**, to better approximate the actual pronunciation.

Another example comes from Japanese. One speaker joined the English verb *lecture* and the Japanese verb *suru*, saying **lecturesuru** to signify *to do lecturing*. This would be transcribed **lecturesuru<alt>**.

It is up to the transcriber to decide how to spell these altered words. It is preferable that the transcription be close enough to the spelling of the original borrowed words in the respective original languages (if known) so that the origin of the alteration is clear. Transcribers should do what seems most natural for the specific utterance and for the language.

Note: **<alt>** should only be used on a word that has changed form. If a French person were to say "Ici, on ne parle pas. On se email." *Email* although used like a French verb, is itself unaltered (it is not conjugated or inflected with French markings) and would not require the **<alt>** tag. It might, however, require **<nitl>**.

4.4 Filled Pauses

Filled pauses are treated like actual words, so they do not appear within pointy brackets. Table 4.1 lists the allowable filled pauses for each language. English filled pause labels appearing in Table 3.2 are also available to transcribers in each language. If a filled pause is uttered that is not in either Table 4.1 or Table 3.2 the transcriber has the following options:

- If it occurs commonly it should be added to this table.
- If it is not common, but can be sounded out attach **<sp>** to the end: **erg<sp>**—no need to create a new label for these rare cases.

- If the utterance is too rare to require a label assignment and you are unable to sound it out transcribe the filled pause <**fp**>.

Table 4.1: Filled pauses for multi-language labeling

OGI ortho	*Worldbet*	Language
ano:	**a n o:**	Japanese
ah	**A**	Czech, Polish, German
ahm	**A m**	German, Czech
e	**e**	Portuguese, Spanish
e:	**e:**	Japanese
e:to:	**e: t o:** or **e t o**	Japanese
em	**e m**	Spanish
eh	**E**	Portuguese, Polish, Swedish
ehm	**E m**	Czech, Swedish
mm (**m**)	**m:**	Japanese, Polish
nnto:	**n: t o:**	Japanese
oah	**> ^**	Swedish
oeoe	**7:**	Hungarian
sh	**S**	French
ung	**^ N**	Mandarin
uh	**^**	Swedish
uhm	**^m**	Swedish
yh	**I_x**	Polish

4.5 Specal Conventions in each language

4.5.1 Arabic

One of the problems in romanizing Arabic for transcription at the orthographic level is that there are different varieties of written Arabic, and no clear standard written version. As of 4/97 we have not begun transcriptions in Arabic.

4.5.2 Cantonese

Cantonese is nearly completely transcribed. In keeping with our goal to provide transcriptions in an ascii based transcription, we have used an already existing romanization as our model. This Yale romanization is treated in *Cantonese a Comprehensive Grammar* by Stephen Matthews and Virginia Yip, 1994.

One modification we made to the romanization is way we mark tone. The tone is transcribed with a number, and is separated from the word by a hyphen. (**baak-3 luhk-6 sahp-6 yaht-6**). There are 7 basic tones in Cantonese:

1. high rising
2. high level

3. high falling
4. mid level
5. low rising
6. mid level
7. low falling

In most cases, there is a straight forward mapping between the romanization and the cantonese character. However there are a few tricky areas. For example, there are some words we have identifed as "spoken only." There exists no written character for these words as they occur in the spoken language only. These words will be identified in the transcription with the tag <**so**>, for "spoken only" connected to the end of the word.

In the sentence: **leih-5 liu<so>-1 jo-2 heui-3 bin-1**, *where did you go?*, **liu<so>** is considered spoken only; it is colloquial and would be replaced with a word such as **jau-2** in formal or written language.

Other words that are spoken only include **je, le, gak** and **ha**. These words are not generally found in dictionaries or formal treatments of the language. However, because they are used so often they are starting to be printed in some newspapers and magazines. Preserving a more traditional view of the language, we attach the <**so**> tag to the romanized word.

The only other time the mapping between the romanization and the original chinese character will not be straight forward is when there are two different chinese characters that are pronounced the same. These words will need to be disambiguated by context.

For example, the character pronounced **leih-5** can be translated in the following ways:

1. Lee: (**leih-5 siu-2 je-2** *Miss Lee*)
2. you: (**leih-5 hou-2 ma-3**, *how are you*)
3. care: (**mh-4 leih-5**, *don't care*)
4. sole: (**yat-1 faai-3 leih-5**, *a piece of shoe's inner sole*)

Now that the romanization/transcriptions are nearly complete we are working on mapping back to the cantonese character in order to disambiguate words words like **leih-5**.

4.5.3 Czech

Written Czech uses a Roman alphabet. However, there are three accent or diacritical markings (the hook, the acute accent and the krouzek, the little circle on top of the letter u). In our transcriptions we have convert these diacritical markings to characters that can be typed on a standard keyboard. See table 4.2 for character conversions.

The first three words in the table are examples of the palatalization marker, or the little hook, háček. The little hook appears on z, s, c, r n, e, d and t. In standard Czech, the palatalized version of lower case d and t are normally marked with an apostrophe; the little hook only appears on the capitalized versions. However, we will transcribe lower case palatalized t and d with the little hook as shown in the table, because there is no real difference between the apostrophe and the little hook in written Czech; the distinction is purely conventional.

Vowel length is marked on Czech vowels by placing an acute accent over the vowel. a e i o u and y have long and short versions. See table 4.2.

The krouz˜ek, or little circle on the u is a historical convention for vowels that used to be long o. u with a circle on top will be transcribed as u+.

4.5.4 English

See Chapter 2 for transcription conventions for English.

4.5.5 Farsi

We have not yet begun to transcribe Farsi. However, Farsi is written with the Arabic script, so the romanization method will likely be similar to that used in Arabic.

4.5.6 French

Standard French is written with the Roman alphabet. French has five characters which are not found on most standard (American) keyboards: the acute and grave accents, and the cedilla, circumflex and diaeresis.

The numbers will be transcribed using dashes as is done in standard French orthography. This avoids ambiguities such as the following: **vingt-deux** "22" and **vingt deux** "20 2."

The French often fill thinking space with **sh**. The label **sh** has been adopted as a filled pause label. (See table 4.1, and table 4.3.)

4.5.7 German

Standard German is written with the Roman alphabet. There are two non ascii characters for which we have devised a conversion scheme: the umlaut and the beta. See table 4.4

Capitalization

Unlike English and, all proper names and all nouns are capitalized in German transcriptions as is done in standard German orthography. It is important to preserve German rules of capitalization because they can, in some cases, disambguate words. German words that are sentence initial are only capitalized if they are nouns or proper names.

4.5.8 Hindi

Hindi has been transcribed orthographically. The Hindi script, for which I have not input Latex fonts, has been romanized. This information is currently being stored off line in my trusty file cabinet under *Languages: Hindi Phonology.*

4.5.9 Hungarian

For special characters, see table 4.5

In Hungarian, the hyphen will be used for the "whether" or "if" condition, as in **Nem tudom hogy lesz-e buli**, *I don't know whether there will be a party.*

Here is an example utterance written first in standard Hungarian orthography and then using our transcription conventions. The sentence means "It's very simple. I cross the main

road, pass the river, and turn right. There is a new supermarket opening nearby."

The sentence: **nagyon egyszerű átmegyek a főeútvonalon és a folyó után roegtoen lefordulok a koezeluenkben nyílik egy új ábécé**

would be transcribed: **nagyon egyszerue' a'tmegyek a foe'u'tvonalon e's a folyo' uta'n roegtoen lefordulok a koezeluenkben nyi'lik egy u'j a'be'ce'**

4.5.10 Indonesian

Standard written Indonesian uses a Roman script. Hyphens are allowed in Indonesian. Nothing else particularly striking (i.e. different from English) happens in Indonesian as I recall.

4.5.11 Italian

Italian has not yet been transcribed. However, it uses a Roman alphabet with a few additional diacritics not found on a standard keyboard. These issues will be dealt with when transcriptions begin.

4.5.12 Japanese

A modification of the Hepburn romanization is used to transcribe at the word level.

Table 4.6 outlines the romanization scheme, unfortunately it does not yet contain the actual Japanese characters. It should still be readable to someone familiar with Japanese.

Note some other salient characteristics:

- Long consonants are transcribed as double consonants, i.e., **kitte** "stamps."
- Long Vowels are transcribed with colons, i.e., **kacho:**, "manager," rather than kacho with a line over the o as in the Hepburn romanization.
- Long e will be transcribed as e: or ei depending on pronunciation. The most common pronunciation is **kire:**, "pretty." But in some cases, speakers say **kirei**, manifesting a strong diphthong. When this occurs, the ei sequence will be used to reflect the sound change.
- The sequence hu is transcribed as fu, as in **fuji** "mount fugi."
- The sequence ti is transcribed as **chi** when it is palatalized, as in, **ichi** "one " and as **tii** when it appears in borrowed words, such as **lemon tii** "lemon tea." Likewise, the sequence di is transcribed **ji** when it is palatalized as in "jikan" **time** and as **dii** in borrowed word such as **birudiingu** "building" The double ii is used in transcribing **tii** and **dii** regardless of the actual length of the vowel because, in this borrowed phonological context, vowel length is in free variation and is not considered phonemically significant.
- Word Boundaries

  1. Particles
     Particles will be considered as distinct words separated by spaces, such as: **watashi wa gakuse: desu**, "I am a student."

2. Classifiers
   Classifiers will be separated from the word they modify by dashes. Note the following examples:
   ichi-ji desu, "one o'clock"
   ip-pun "one minute"
   yok-ka "fourth day"
   yo:-ka "eighth day"

3. Numbers
   Numbers will be written as one word, such as **ju:ni**, "twelve," and **hyakurokuju:**, "one hundred and sixty."

4.5.13 Korean

Korean transcriptions have not yet begun. The issue of what romanization to use will need to be addressed when transcriptions begin, but we will likely use some form of Hangul, such as that used in the dictionary entitled "English-Korean Practical Conversation Dictionary" printed by Hollym in 1984.

4.5.14 Mandarin Chinese

Standard Pinyin is used to transcribe mandarin, with two modifications: 1) tone 2) ü. See table 4.7. Pinyin can be found in many places (Chinese-English dictionaries, for one). Its use is quite widespread.

The four Mandarin tones are identified by the numbers one through four. In orthographic transcriptions these numbers are connected to the word by a dash. The high, level tone is 1, the rising tone is 2 the falling-rising is 3, and the falling tone is 4. A reduced tone sometimes occurs in fast speech. This is usually found in a tone one word whose pitch is drastically reduced or shortened, and which does not reach the target of tone one in actual pronunciation. Because the reduced tone is not phonemic it is not labeled at the word level. The so called "reduced tone" normally happens at the end of a question or in the pronunciation of some function words, like the word **"de."**. [1] [2]

4.5.15 Polish

Although written with the roman alphabet, Polish has a number of characters not found on a standard keyboard. See table 4.8 displays our ascii solution to this potential problem.

4.5.16 Portuguese

Although written with the roman alphabet, Portuguese has a number of characters not found on a standard keyboard. Table 4.9 displays our ascii equivalents of the agudo accent, cedilha, circumflexo and til.

[1] In the phonetic labeling of the OGI Multi-language Corpus, tones were labeled phonetically, as they sounded, not necessarily phonemically. Phonetic labeling of tones will continue in this manner to allow comparison of expected pronunciations in the word level transcriptions with actual pronunciations in the phonetic transcriptions.

[2] In the phonetic level labeling of the OGI Multi-language Corpus the reduced tone was labeled as tone one.

4.5.17 Russian

Russian transcriptions have not begun. Most Russian characters belong the the Roman alphabet, but the remaining non roman characters will need to be Romanized. We will probably do this by employing capital letters which most closely resemble the Russian character.

4.5.18 Spanish

Spanish is written in a roman alphabet. There are two characters not available on a standard keyboard. One is the accent and the other is the tilde. See table 4.10

4.5.19 Swahili

Conventions for Swahili are in the process of being developed. Stay tuned.

4.5.20 Swedish

Swedish is written with the roman alphabet. Table 4.11 displays transcription conventions used when transcribing the umlauts. The fourth character, u umlaut, is used for German loan words.

- The hypen is used in Swedish transcriptions for hypenated names, such as **maja-lena** or **lars-erik**.
- The apostrophe is used in Swedish to distinguish words like **ide'** *idea* (**i d e &**) from **ide** *den* (**i d &**). At least in this case the word with the apostrope is pronounced differently than the word without.

  Bjoernen ligger i sitt ide. Han foar en ide'.
  The bear lays in his den. He gets an idea.

4.5.21 Tamil

4.5.22 Vietnamese

Although written with the roman alphabet, Vietnamese has a number of characters not found on a standard keyboard. Table 4.12 displays our ascii solution to that potential problem. Symbols following a letter usually apply to the previous letter. In the transcriptions, symbols should appear in their original order, from left to right and from top to bottom. (Symbols above the letter will appear first, followed by symbols below the letter; leftmost symbols will appear first and then right.)

Known problems with table 4.12:

- In in the Veitnamese version of the first word, du*o*.c, the hook (glottal stop) symbol is misleading. This marking actually looks like a comma or backward **c** connected to the top right corner of the letter. The o has both a hook and a dot underneath it. This marking is different than the one in the word **tra?**.
- In Vietnamese version of cha^.m, the carat should be above the letter a, not to the left of it.

- In the Vietnamese version of nhie^` , the ^ and ` should both be on top of the e.
- In the Veitnamese version of ngie^`u, the marking should look like the upper half of a question mark and be connected to the top center of the letter. The actual diacritic does closely resemble this glottal stop symbol. **This is a different marking than in du*o*.c**

Table 4.2: Special characters, Czech

Czech ortho	*OGI ortho*	*type*
česky	c~esky	palatalization, hook
ted'	ted~	palatalization, hook
tat'ka	tat~ka	palatalization, hook
otázka	ota'zka	acute accent
u w/ circle on top	du+l	krouz~ek

Table 4.3: Special characters, French

French ortho	*OGI ortho*	*Type*
écoute (listen)	e/coute	acute accent
lève (rise)	le\ve	grave accent
français french)	franc,ais	cedilla
tête (head)	te^te	circumflex
ciguë (hemlock)	cigue"	diaeresis

Table 4.4: Special characters, German

German ortho	*OGI ortho*	*Type*
ä	ae	a umlaut
ö	oe	o umlaut
ü	ue	u umlaut
ß	ss	beta

Table 4.5: Special characters, Hungarian

Hungarian ortho	*OGI ortho*	*Type*
á	a'	acute accent
é	e'	acute accent
ó	o'	acute accent
ú	u'	acute accent
í	i'	acute accent
ö	oe	o umlaut
ü	ue	u umlaut
ő	oe'	double accute accent
ű	ue'	double accute accent

Table 4.6: Japanese Romanization

	a	e	i	o	u
k	ka	ke	ki	ko	ku
s	sa	se	si	so	su
t	ta	te	chi/tii	to	tu
d	da	de	ji/dii	do	tu
n	na	ne	ni	no	nu
h	ha	he	hi	ho	fu
m	ma	me	mi	mo	mu
y	ya	ye	yi	yo	yu
r	ra	re	ri	ro	ru
w	wa	we	wi	wo	wu
ky	kya			kyo	kyu
sh	sha			sho	shu
ch	cha			cho	chu
ny	nya			nyo	nyu
hy	hya			hyo	hyu
my	mya			myo	myu
ry	rya			ryo	ryu
gy	gya			gyo	gyu
j	ja			jo	ju
by	bya			byo	byu
py	pya			pyo	pyu

Table 4.7: Special symbols, Mandarin

Pinyin Orthog	*OGI Orthog*	*Type*
nü	nu^-3	
de	de (no tone)	reduced tone

Table 4.8: Special symbols, Polish

Polish Orthog	*OGI Orthog*	*Worldbet Label*	*Polish Orthog*	*OGI Orthog*	*Worldbet Label*
a	a	a	e	e	e
ą	a~	>~	ę	e~	e~
c	c	ts	g	g	g
ć	c'	tS	i	i	i
ci	ci	tSi	j	j	j
ch	ch	x	ł	l/	w
h	h	x	ń	n'	nj
cz	cz	tS	i	ni	nj
dz	dz	dz	o	o	>
dź	dz'	dZ	r	r	r
dzi	dzi	dZ	rz	rz	Zr
dż	dz:	dZr	ś	s'	s
sz	sz	S	w	w	v
ż	z	Zr			

Table 4.9: Special characters, Portuguese

Portuguese ortho	*OGI ortho*	*Type*
aeróbica (aerobics)	aero'bica	agudo accent
força (strength)	forc,a	cedilha
pôr (to put)	po^r	circumflexo
pão (loaf)	pa~o	til

Table 4.10: Special characters, Spanish

Spanish ortho	*OGI ortho*	*Type*
está (is)	esta'	acute accent
niño (child)	nin~o	tilde

Table 4.11: Special characters, Swedish

Swedish ortho	*OGI ortho*	*Type*
å	oa	umlaut (ring)
ä	ae	umlaut
ö	oe	umlaut
ü	ue	umlaut

Table 4.12: Special characters, Vietnamese

Vietnamese ortho	*OGI ortho*	*Type*
tuọc	du*o*.c	
cám	ca'm	
chạ^m	cha^.m	
câu	ca^u	
nghĩ	nghi~	
nghiê`u	ngiê'u	
tra	tra?	

Chapter 5

Phonetic Level Labeling

5.1 General Comments

"It is vain to do with more what can be done with fewer." —William of Occam

The speech signal is packed with acoustic information. Yet our ear sorts the information into intelligible sounds and words even in the presence of distortions and disfluencies.

Not all of the information in the signal is important, in fact much is filtered out by the human perceptual system. The labeling we do should highlight the information used by people to understand speech. We label what we consider to be the core information–the contrastive sounds of the language, and in addition we mark salient features of the speech which are visually striking, such as glottalization or excessive nasalization.

For efficiency, we draw the line at segments that are:

- visually (acoustically) distinct
- distinct enough both visually and perceptually to allow for consistent and accurate transcriptions by multiple lablers
- worth the time it takes to label

5.2 Overview

This chapter is a general discussion of the transcription process at the phonetic level. Instructions are given to train one to label at the broad phonetic level. Included is information concerning segmentation issues and non-speech labeling conventions.

5.3 Alignment

Phonetic transcriptions are time aligned with the waveform to a certain degree of accuracy. In ambiguous cases the boundaries are set according to established rules. Boundary placement is discussed in detail in this chapter, although we know that a strictly time aligned transcription is impossible, due to coarticulation. Phones are not always signalled by discrete, non-overlapping regions in the waveform. In continuous speech, coarticulation, deletion, and elision cause phone boundaries to overlap. Therefore, because true boundaries do not actually exist in many cases, care must be taken to follow convention in order to ensure consistency.

5.4 Level of Labeling

That which, for the sake of simplicity, we identify as "phonetic labeling" would be more accurately termed broad phonetic labeling with a phonemic basis. For each language, a set of phonemes (distinctive speech sounds within the language) are chosen. These label sets contain all of the phonemes in each language, in addition to spectrally distinct and frequently occurring allophones. The allophones, with a few exceptions, are labeled with diacritics.

5.5 Label Set

In the past, data was transcribed phonetically using OGIbet, a broad phonetic transcription set based on TIMIT. In 1993, a different transcription set, Worldbet [**?**], was adopted for phonetic transcriptions. This label set, developed by Dr. Jim Hieronymus, was chosen because it can more consistently handle transcriptions of non-European languages without multiply defining a given symbol. Whereas the OGIbet label **r** was used to identify both the English retroflex and the Spanish alveolar flap (and other sounds) Worldbet **r** signifies only type of sound, an alveolar trill, where as the retroflex approximate is transcribed **9r**. For this reason, as well as the fact that Worldbet is made up of ASCII characters which are easier to process, Worldbet was determined to be more suitable and extensible for the multi-language labeling effort at the CSLU. See 1.3.3.

The content of the label sets in the following chapter is consistent with those label lists in [**?**], with a few exceptions. All deviations from standard Worldbet are annotated. The differences are minimal i.e., an occasional added symbol.

5.6 Label Assignments

To the unaccustomed user, Worldbet symbols may seem awkward. Due to a desire to create a label set which corresponded to IPA labels and a need to use only ASCII symbols, specific label assignments were, at times, somewhat creative. It was the intent of its author to remain as close as possible to the IPA by using symbols related to the IPA in some visible way. Admittedly, this relationship was "sometimes only true in the abstract sense." [**?**]. It is hoped that the unfamiliarity of this label set will not be a hindrance to its use, as it is a valuable tool when phonetic transcriptions of multiple languages are needed.

5.7 Diacritics

Where as in some transcription sets, such as the IPA, diacritics are seen as super or subscript characters, worldbet diacritics are made distinct from base symbols by use of the linking symbol, the underscore. So a glottalized vowel might be transcribed A_?. For a detailed description of diacritics, see Chapter **??**.

5.8 Necessary Tools

Phonetic Labeling requires a number of skills and tools discussed here.

One must chose a transcription set suitable for the task. There are many philosophies embodied in existing label sets. CSLU uses the Worldbet transcription set, developed by James

Hieronymus, because of its multi-lingual extensibility.

Phonetic training is an important asset because it enables the listener to hear accurately those deviations from the expected pronunciation that are common to fluent speech.

An ability to read and interpret acoustic cues provided in the spectrogram and waveform is important, as are tools for viewing and segmenting of the speech signal. Toolkit has a nice acousic display thanks to recent improvements made by Tim Carmel and Johan Wouters, and the tool is free to academic universities and corporate sponsors.

Knowledge of the language being transcribed is strongly recommended. Labeler reliability studies have shown that agreement is wildly different when labelers do not know the language they are labeling. ([?], [?]).

5.9 Segmentation and Label Selection

Placing exact boundaries is difficult. In continuous speech many boundaries that are intuitively perceived by listeners do not exist when the speech is examined acoustically. Words overlap and phones become coarticulated. Because of this, the boundaries we assign when we label speech are sometimes artificial. Ambiguous cases are specified by rule to achieve consistency.

It is our contention that the ear is the not the most reliable source to use to determine where to place boundaries. Labelers are encouraged to listen to the labeled segments in context, as coarticulation alters the percept of a sound in isolation. For the first pass, the labeler should listen to a phone in a context of at least one phone on each side of the phone in question. Use available acoustic information in the waveform to decide where to place the boundary. (For high frequency, low amplitude phones the spectrogram may be used, as these sounds are difficult to identify in the waveform.) The practice of listening to a segment for a given sound and then extending the boundary until the sound is no longer heard is not generally recommended for determining the placement of boundaries, although it may be necessary as a last resort (in the absence of spectral cues).

Boundaries should be set, whenever possible, where the waveform shows change. Changes in the waveform are generally more reliable than changes in the spectrogram. This is especially true for transitions from low to high amplitude sounds. The spectrogram is computed by averaging the energy in a given sized window and displaying the averaged values. While in most cases this increases the accuracy of the spectrogram, when the energy level goes from near zero (as in a stop closure) to very high (in the stop burst) ghost images appear in the spectrogram. These "ghost bursts" show up on the spectrogram perhaps 10 or 20 milliseconds before the burst appears in the waveform. They are caused by the averaging of the abnormally low values with the very high values. Ghost bursts that appear in the spectrogram should not be used for segmentation, rather the labeler should set the boundary where the waveform shows change.

5.9.1 Stops, Affricates, and Trills

Stops, affricates and trills usually manifest clear changes in the waveform, especially when they follow a pause or a closure. When labelers see movement in the waveform begin they should set the left boundary of the burst label.

When using the OGI speech tools Only only need set the left boundary, as the right right boundary of one label is automatically aligned to the label follows when the file is saved using the "write align" command. Manually extending the right boundary leads to the formation of overlapping boundaries, which, by convention, are not allowed in transcriptions.

Stop bursts that are heavily aspirated are the easiest to mark. Stop bursts with no aspiration may be more easily heard than seen. To identify unaspirated stop boundaries, look for a single pulse in the waveform that is much lower in amplitude than the following vowel.

In fast speech, plosives are often released gently with little or no pressure build up. If a stop burst does not appear readily in the waveform or spectrogram it is possible that one did not occur. Do not allow phonemic expectation to alter interpretation of what is actually in the signal. Some strategies for identifying stop bursts:

1. Set a finer resolution on the waveform. Normally when transcribing a resolution of .001 seconds per pixel is used, but .00025 will make segmentation simpler in this case. If a stop has been released, there will be an irregular portion of the waveform that will distinguish the stop burst from both the preceding closure and following phone. Segment the distinct portion and label it as the burst.

2. Look closely at the formants which follow the stop closure. is there a period where the formants are level and then begin to move into position for the vowel? Segment the portion of the spectrogram where the formants are level and label it as the burst.

3. If there is no acoustic information signaling the burst, assume one did not occur.

5.9.2 Closures

In general, the left boundary of a closure is placed where the energy for the preceding phone stops. The cease of energy is obvious in most cases when the closure follows anything but a pause.

Following are conventions used to label closures:

Voiceless closures

- Voiceless stop closures which occur at the end of a pause and the beginning of an utterance often have spectral evidence to signal the beginning of the closure. Speakers may make a small amount of noise when moving their articulators into the position of the closure. when this occurs, a small pulse is seen in the waveform or the spectrogram at which point the closure label should be placed.

- If there is no acoustic evidence to signal the beginning of the closure, it should still be labeled. The boundary should be set 50ms to ensure that labels are consistent. This length was chosen as an "average length" of a word initial closure in running speech.

- Likewise, when a word, phrase, or utterance ends in a voiceless stop that is not released, the stop closure label should extend 100ms after the energy in the waveform dies out. This value was chosen as an average length of word final closures in running speech.

Voiced closures

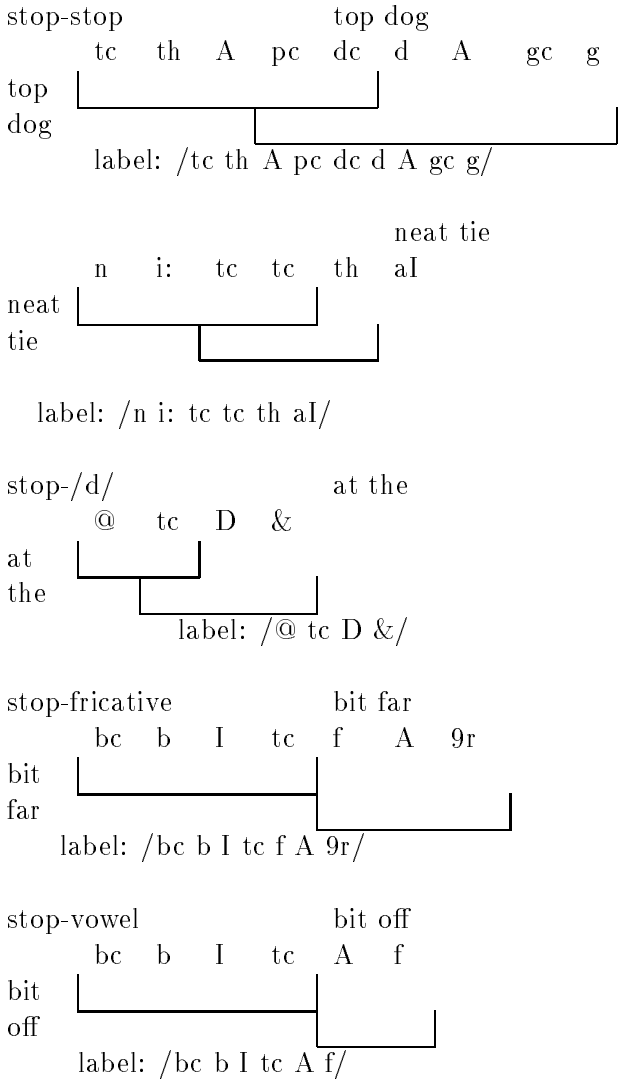
- The voicing in a voiced closure is normally visible in the waveform. Set the left boundary at the point of most radical change in the waveform.

- If the voicing in the phonemically voiced closure is not evident, the closure should be labeled with the devoicing diacritic, and the label should extend 50ms word initially and 100ms word finally. Again, this length is arbitrary, but was chosen as an average length for consistency.

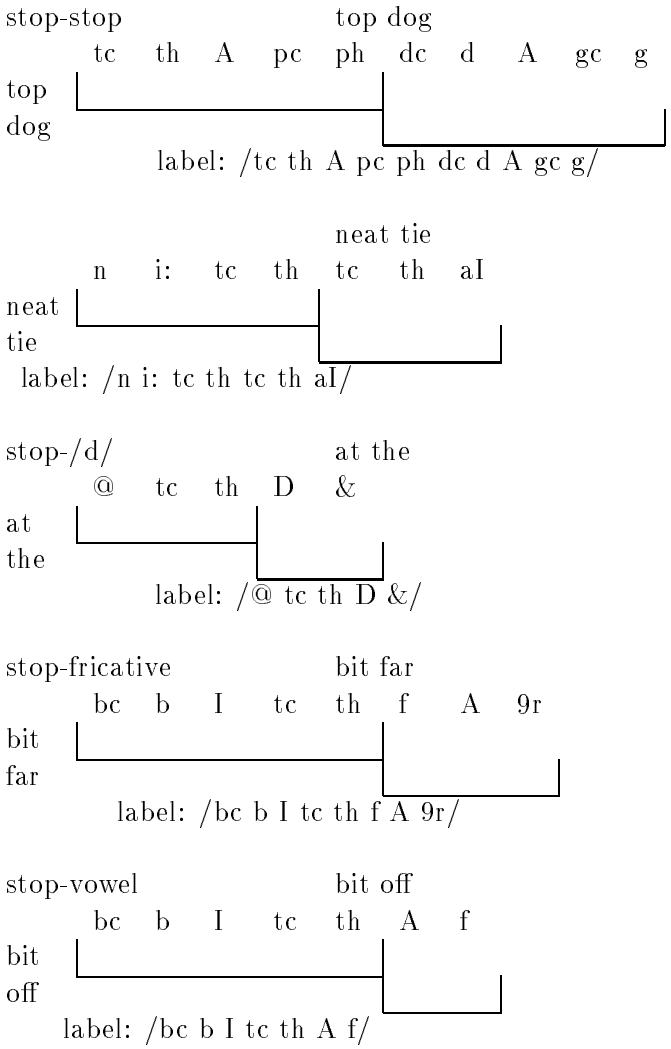
- A voiced stop which follows a nasal often has no visible closure. this is either because the closure is very short or because the voicing makes it indistinguishable from the preceding nasal. As the closure interval is too difficult to isolate, it should not be marked. when the place of articulation is the same for the nasal and the closure, part of the nasal acts as the perceived closure. The velum is closed just before the burst to allow the pressure to build up for the burst. This build up interval can be very short, and need not be labeled. For example, a **dc** preceded by a **n** is usually imperceptible and does not require a label.

Sometimes stops are not released in running speech. sometimes they are. it can be confusing to label them. The following are the conventions for labeling stop closures between words.

Unreleased Closure Segmentation



Released Closure Segmentation



5.9.3 Fricatives

Of the fricatives, sibilants are the easiest to isolate because of the high energy they produce: the onset can be determined by a sudden, heavy increase of random energy in the spectrogram. other types of fricatives are also marked by random energy in the spectrogram, but the amplitude (and visibility) may be very low in the waveform.

Set the left boundary of a fricative label at the point of most radical change in the waveform. If there is little or no clear visible indication of the fricative in the waveform, but you can hear and see indications of a low amplitude fricative in the spectrogram, set the boundary where there is an increase of energy in the spectrogram.

If the fricative is released very strongly, like a burst, label the closure + (epenthetic closure) and the burst the appropriate fricative label.

5.9.4 Nasals

The onset of a nasal is nearly always easy to determine: the waveform rises or drops into a highly periodic, low amplitude signal. The nasal usually carries the same formants of the preceding vowel or other phone, but is lighter in color or intensity. the change in the waveform is very easy to spot because the decreased amplitude causes the height of the signal to be shortened. (That is, the absolute value of the wave has decreased).

At the end of an utterance or phrase, a nasal may trail off gradually until it is indistinguishable from environmental or line noise. Set the right boundary where the first formant in the spectrogram dies. this should coincide with a point of "radical" change in the waveform. The reason we use the disappearance of the first formant to determine the right boundary of a phone is because it is the best clue to voicing. Thus, when the first formant is no longer visible, we know that voicing has ceased, and the phone is said to be terminated. if other formants continue in an obvious way, the devoicing diacritic should be used on the end of the phone, for example, **n_0**.

In some languages, a nasal may be indicated by heavy nasalization of the preceding vowel; it may not be possible to isolate the nasal. This happens frequently in languages like English, where nasalization is not phonemic. Be careful that a separate label is not given to a nonexistent nasal. If there is no indication of a nasal in the waveform, but you hear nasalization, use the diacritic **a_~** on the vowel. Use the nasalization diacritic for the whole vowel even if the nasalization does not carry throughout the entire segment. See **??** for more information on labeling nasals.

5.9.5 Liquids

The onset of a liquid is marked by the disappearance of f3 (after a vowel) or the appearance of f1 and/or f2 (after a nasal or obstruent). There is a corresponding change visible in the waveform. after a vowel, the onset of a liquid may be as gradual as the onset of a semivowel; use the guidelines listed for separating vowels from other vowels or semivowels to determine an appropriate boundary point.

5.9.6 Vowels and Approximates

The onset of a vowel or approximate following a nasal or obstruent is marked by the appearance or darkening of formants in the spectrogram and by increased amplitude and/or periodicity in the waveform; as always, the point of most radical change in the waveform (when it can be determined) is the place to set the left boundary.

After a liquid or strong approximate, the onset of a vowel is marked by the appearance of the f3 and/or f4 and f5 in the spectrogram. (cues from f4 and f5 will generally not exist in telephone speech due to the limited bandwidth.) the waveform will show some change, but there is no typical change to look for. Be sure to use cues from both the spectrogram and waveform, setting the left boundary where the changes coincide. Give preference to the changes in the waveform.

After a semivowel or vowel, it can be practically impossible to determine the exact onset of a vowel. To be consistent, we have chosen to place the boundary in the middle of the transition period. If the formants never level off on either the semivowel or the vowel, divide the segment in half.

5.9.7 Devoiced Vowels

Not all vowels are voiced in rapid speech, are voiced. Devoiced vowels tend to occur after a voiceless obstruent and tend to be shorter than voiced vowels. If a devoiced vowel is suspected, look for the occurrence of f2 and f3 in the spectrogram, disappearance of the first formant, and perhaps increased periodicity in the waveform. Set the left boundary at the onset of these changes, and use the devoicing diacritic. This diacritic, **_0** will be used only on the part of the vowel that is devoiced.

At the end of an utterance or phrase, the final vowel or approximate may trail off gradually until it is indistinguishable from environmental or line noise. Set the right boundary where the first formant in the spectrogram dies. This should coincide with a point of radical change in the waveform.

5.9.8 Geminates

A geminate is two identical, sequential phonemes. In the phrase, "nine nouthetic counselors," there are two geminates, n in "nine" and "nouthetic," and the velar closure and optional burst(s) between "nouthetic" and "counselors." The burst is optional because in this context often the first plosive is not released, and the geminate is realized in the length of the closure. Geminates are generally one and a half times longer than single phones and are usually marked by spectral discontinuity such as lower amplitude. Following are the conventions used to label geminates.

Splitting geminates is useful in speech recognition where there is no facility for between-word phonological rules. One can always automatically merge geminates if there is no need to explicitly represent them.

Lowered Amplitude

Boundaries between geminates may be signaled by lower amplitude. In this case, the phonemes are continuous, but the amplitude decreases when articulation of the second phone begins. There is no intervening pause. Place the boundary between the two phones where the amplitude drops the most radically, labeling both segments with the same phoneme label if there are no other acoustic differences between the two segments.

Glottalization

Glottalization is another acoustic cue signaling the boundary between the geminates. For example, in the phrase "we each have more." the two **i:** phonemes might be separated by a period of glottalized **i:**. Segment the first period of non-glottalized vowel in "we" as **i:**, the glottalized portion as **i:_?** and the final non-glottalized portion in "each" as **i:**. This is consistent with the conventions for labeling glottalization. When converting to the phonemic level, to avoid confusion between how many phonemes there are in the string (in this case how many **i:** phonemes there are), any glottalized vowel can be merged with a preceding or following non-glottalized vowel if one exists.

Allophones

Geminates can be realized as two allophones of the same phoneme. The two segments will be distinct acoustically and visually. For example, in the Spanish sentence *ohala' que pepe este' en la casa* "i hope pepe is in the house", the first of the geminates would usually be realized as

e, and the second as **E**. If labels for these allophones do not exist in the language, segment into two segments marked with the same label, placing the boundary where vowel quality changes. When the more descriptive labels exist for the language use them.

Lengthening

If there is no other spectral cue, geminate phones are signaled by lengthening. Usually geminates will at least be longer than an average occurrence of the phone, but the quality of the phone may remain constant. When this occurs divide the geminate in half and give the same base label for each segment. two labels are given to the segment to retain a phonemic representation of the utterance.

Last Resort

If there is simply no spectral discontinuity and the geminate does not appear markedly longer than a single phone, the geminate should still be split in half and two identical labels should be given to each segment.

5.10 Non-speech Events

The acoustic signal does not solely contain speech events. Sounds such as telephone line noise, laughter, and inhalation do not convey speech information, per se. However, because they are present in the signal, and because recognition systems need to learn to distinguish between speech and non-speech information, perceptible non-speech phenomena will be labeled. See chapter 2 for non speech symbols and usage. In phonetic transcriptions, non-speech labels are used conceptually in exactly the same way as they are used in time aligned word transcriptions. The only differences are that they are time aligned to the waveform and they are preceded by a period rather than being enclosed in pointy brackets.

5.10.1 Pause, Closure, and Epenthetic Closures

Technically speaking, there is no silence in the signal, because there is always line noise or some other type of noise present in the signal. The label **.pau** (pause) is used for periods of time in which the speaker has stopped speaking, or for periods in which there is very little or no apparent energy in the signal. There is no length restriction on the label **.pau**.

The only acoustic difference between a voiceless closure and a pause is the period leading into and coming out of the closure. Due to coarticulation, the formants will reveal the position of the articulators if they have moved into a certain position for the closure. Other than this, the voiceless closure segments could be thought of as pauses as well, as they do span a period of little or no activity in the spectrogram.

The closure labels are used to mark a specific type of articulatory closure: t closure (**tc**), p closure (**pc**), etc. (see chapter 6 for the phonological labels. Closures occur before stops and affricates. as distinct from **.pau** closures do have a typical length because speaker and rate of speaking.

A *Pause* is a discourse unit; it breaks up the sentence and indicates a boundary between sub-sentence sets of words. *Epenthetic closures* are always caused by movement of the articulators from one position to another. They are often referred to as *insertions*, since a closure is inserted into a word.

Some general guidelines to follow (for English; other languages may have other telltale clues that can be used):

- Pause Label a segment with *pause* (**.pau**) if you can "hear" a comma, such as when a person lists a series of items or seems to be collecting their thoughts before going on.
- Closures Use the closure symbols in conjunction with stop or affricate labels. Closures precede all plosives or affricates, and are formed by complete closure of the articulators. Usually this closure is accompanied by a build up of pressure seen in the release of the stop or affricate burst.
- Epenthetic Closures Use the label + when there is no actual pause, but a period of closure in the spectrogram that can not be associated with a phonemic stop or affricate. The segment should be 30 milliseconds or longer. Often it occurs between a nasal and a fricative. Periods of silence between words are usually labeled as pauses and not epenthetic closures. The epenthetic closure label is also used on the closure preceding fricatives that have been realized as stops. See Section 5.9.3.

Chapter 6

Phonetic Labels

6.1 Introduction

The tables in this chapter display those speech labels used in the phonemic/phonetic transcriptions of six languages: English, German, Hindi, Japanese, Mandarin Chinese and Spanish. With the exception of Hindi, these are a part of the OGI multi-language (10-language) corpus. The current data collection has expanded the original 10-language corpus into 22 languages including: Arabic, Cantonese, Czech, Farsi, French, German, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Mandarin, Polish, Portuguese, Russian, Spanish, S wahili, Swedish, Vietnamese, Tamil, and English.

Each language has a unique set of labels. Only labels for the languages that have been labeled at the CSLU appear in this document. Each label set is composed of mainly phonemic base symbols and diacritics that capture phonetic detail. See Chapter **??** for a description of each diacritic, and Chapter 5.10 for the non-speech labels used in phonetic labeling.

The aim of Worldbet is to enable consistent transcription across multiple languages, so that a single symbol is similarly defined across all languages. The label sets are designed to be extensible to all languages, regardless of language family.

Chapter 6 discusses each of the six languages consecutively, with the following subsections:

- **Vowels**. This includes a vowel chart, patterned after IPA vowel triangle, and a table containing all vowel labels with word examples.
- **Diphthongs**. This section charts the diphthong labels used in the language.
- **Notes on Vowels** (optional). This section contains additional information about labeling issues regarding vowels and diphthongs.
- **Consonants**. Included here is a consonant table, which contains only the phonetic labels, and denoting manner and place of articulation. Following is an additional table devoted to providing word examples for label use. [1]
- **Notes on Consonants** (optional). This section contains additional information about language specific labeling issues.

[1] For IPA correspondences of CSLU/Worldbet symbols, see [?].

6.2 English

6.2.1 Vowels

Table 6.1: English Vowel Chart (not including retroflexes)

	front	central	back
high	i:	ux	u
mid	I E @	I_x & &0 ^	U >
low			A

Table 6.2: English Vowel Examples and Description

Worldbet	**OGI**	**Example**	**Description**
i:	iy	beet	high front long
I	ih	bit	mid-high mid-front short
E	eh	bet	mid-low front short
@	ae	bat	mid-low front long
I_x	ix	roses	centralized I
u_x		suit	fronted u
&	ax	above	mid-central short
&_0		to go	voiceless &
5		pot	British
u	uw	boot	high back rounded
U	uh	book	mid-high back rounded short
^	ah	above	mid-low central
>	ao	caught	mid-low mid-back rounded
A	aa	father	low back
3r	er	bird	rhotacized mid-central
&r	axr	butter	short rhotacized mid-central

6.2.2 Diphthongs

6.2.3 Notes on English Vowels

- Rhotacized vowels are included in Table 6.2.
- The carat, ^ is normally found in stressed syllables. It is longer and slightly lower than the reduced vowel, **&**.
- The syllabic retroflexes, **3r** and **&r** have a similar distinction as ^ and **&**, where **3r** is normally found in stressed syllables and **&** in unstressed. **3r** is longer, more tense and slightly lower than the reduced vowel, **&r**.
- Often speakers of American English do not round the high back vowel, **u**. The allophone label **u_i** will be used in these cases.

Table 6.3: English Diphthong Examples and Description

Worldbet	OGI	Example	Description
ei	ey	bay	e ⇒ i
aI	ay	bye	a ⇒ I
>i	oy	boy	> ⇒ i
iU		few	i ⇒ U
aU	aw	about	a ⇒ U
oU	ow	boat	o ⇒ U
i&		here	i ⇒ &
e&		there	e ⇒ &
u&		poor	u ⇒ &

- Note that **i&, e&, u&** and **5** are most commonly found in British pronunciation.

6.2.4 English Consonants

English has bilabial, labiodental, alveolar, alveo-palatal,, velar and glottal consonants, abbreviated as follows in Table 6.4: *bilab*, *l-d*, *int-d*, *alv*, *postalv*, *vel*, and *gl*. See Table 6.5 for a list of the labels with word examples and descriptions.

Table 6.4: English Consonants

	bilab	*l-d*	*int-d*	*alvel*	*alvpal*	*vel*	*gl*
voiceless stops	ph			th		kh	
voiced stops	b			d		g	
flaps				d_(th_(			
voiceless affricates					tS		
voiced affricates					dZ		
voiceless fricatives		f	T	s	S		h
voiced fricatives		v	D	z	Z		h_v
nasals	m m=			n n_(n=		N	
lateral				l l=			
retroflex				9r			
glides	w				j		

6.2.5 Notes on English Consonants

- Aspiration is overtly transcribed on all voiceless stops in English. This contrasts with relatively unaspirated stops in languages like German, where the stop label would be **p**,

Table 6.5: English Consonant Examples and Descriptions

Worldbet	OGI	Example	Description
ph	p	pan	voiceless bilabial stop
pc	pcl	_pan	voiceless bilabial stop closure
th	t	tan	voiceless alveolar stop
tc	tcl	_tan	voiceless alveolar stop closure
kh	k	can	voiceless velar stop
kc	kcl	_can	voiceless velar stop closure
b	b	ban	voiced bilabial stop
bc	bcl	_ban	voiced bilabial stop closure
d	d	dan	voiced alveolar stop
dc	dcl	_dan	voiced alveolar stop closure
g	g	gander	voiced velar stop
gc	gcl	_gander	voiced velar stop closure
m	m	me	bilabial nasal
n	n	knee	alveolar nasal
N	ng	sing	velar nasal
d_(	dx	rider	alveolar flap
th_(	dx	writer	alveolar flap
n_(	nx	banter	alveolar nasal flap
f	f	fine	voiceless labiodental fricative
T	th	thigh	voiceless dental fricative
s	s	sign	voiceless alveolar sibilant
S	sh	assure	voiceless alveo-palatal sibilant
h	hh	hope	voiceless glottal fricative
v	v	vine	voiced labiodental fricative
D	dh	thy	voiced dental fricative
z	z	resign	voiced alveolar sibilant
Z	zh	azure	voiced alveo-palatal sibilant
h_v	hv	ahead	voiced glottal fricative
tS	ch	church	voiceless alveo-palatal affricate
tSc	chcl	_church	tS closure
dZ	jh	judge	voiced alveo-palatal affricate
dZc	jhcl	_judge	dZ closure
l	l	limb	alveolar lateral
9r	r	right	retroflex approximate
j	y	yet	palatal glide
w	w	when	bilabial glide
m=	em	bottom	syllabic **m**
n=	en	button	syllabic **n**
N=	eng		syllabic **N**
l=	el	bottle	syllabic **l**

t or **k**.

- If the phonetic realization of a phoneme is very short, ie, flapped, the diacritic _(will be used. This occurs frequently with /b/, /g/, /n/, /t/, and /d/. The alveolar flap, although a single phonetic percept, has dual underlying phonemic representation, and is transcribed as either **th_(** or **d_(**. This serves to distinguish the phone(s) in the words "writer" and "rider." Although it is not clear whether or not these phones are acoustically distinct, they are transcribed separately to facilitate mapping to the word level from phonetic level transcriptions.

- Spectrally, some fricatives seem to manifest closure/burst segments. These apparent closure-like segments have been analyzed as either periods of decreased frication with low amplitude (in which case the segments are included in the label box for the fricative), or they are analyzed as epenthetic closures (and labeled +).

6.3 French

6.3.1 French Vowels

Table 6.6: French Vowel Chart

	front	central	back
high	i y		u
mid	e 7 E E~ 8 8~	 &	o
low	 a	^	> >~ A A~

6.3.2 Notes on French Vowels

- **y** is rounded **i**. Some speakers have such a high placement of **y** that it is slightly fricated. This is still labeled **y**.
- **7** is rounded **e** (ipa ø) .
- **8** is a rounded **E** (ipa œ).
- Many French speakers do not distinguish between **7** and **8**, but those that do contrast *jeune* **/Z 8 n/** "young," with *jeûne* **/Z 7 n/** "fasts." **??**.
- The nasalized vowels **A~ >~, E~**, and **8~** have the same placement as their non nasalized counterparts, **A, >, E** and **8**.

6.3.3 French Consonants

6.3.4 Notes of French Consonants

- **K**, the voiced uvular fricative is commonly devoiced, especially word finally, *pasteur*. When the phoneme **K** is devoiced, it is labeled **K_0**.
- **X** (the voiceless uvular fricative) was introduced for foreign words, and should not be used when **K** is devoiced.
-
-

6.4 German

6.4.1 German Vowels

Table 6.7 shows the German vowel chart, and Tables 6.8 and 6.9 list the vowels and diphthong labels that are used.

Table 6.7: German Vowel Chart

	front	central	back
high	i: y: I Y		u: U
mid	e: 7 E 8 @	& ^	o: >
low	 a	ax	

Table 6.8: German Vowels Examples and Descriptions

Worldbet	OGI	Example	Description
i:	iy	bieten	high front long
y:	iyw	Güte	rounded i:
I	ih	bitten	mid-high mid-front short
Y	ihw	Mütter	rounded I
7	eyw	Goethe	mid-high front rounded long
E	eh	Betten	mid-low front short
E:	eh	bäte	mid-low front long
8	ehw	Götter	mid-low front rounded short
e:	ee	beten	mid-high front long
a	a	Ratte	low front
a:	aa	raten	low front long
>	oh	Rotte	mid-low mid-back
^	ah		mid-low central
o:	oo	rot	mid-high back rounded long
U	uh	Kutte	mid-high back short
u:	uu	Rute	high back rounded long
&		Gesetz	central short
ax	ah	besser	mid-low central
@	ae	Stähle	mid-low front long

6.4.2 German Consonants

German has bilabial, labiodental, alveolar, palatal, velar, glottal, and uvular consonants. Table 6.11 gives a list of the consonant labels used in German.

Table 6.9: German Diphthongs

Worldbet	**OGI**	**Example**	**Description**
iax	ia	Tier	i ⇒ ax
yax	ia	Tür	y ⇒ ax
7ax	ea	Gehör	7 ⇒ ax
Eax	ea	er	E ⇒ ax
eax	ea	Gewehr	e ⇒ ax
aI	ay	leite	a ⇒ i
aU	aw	Laute	a ⇒ uw
aax	aa	Jahr	aa ⇒ ax
>Y	oy	Leute	ow ⇒ Y
oax	oa	Tor	o ⇒ ax
uax	ua	Ruhr	u ⇒ ax

Table 6.10: German Consonants

	bilab	*l-d*	*alv*	*pal*	*vel*	*uvl*	*gl*
voiceless stops	p		t		k		
voiced stops	b		d		g		
voiceless affricates	pf		ts	tS			
voiced affricates				dZ			
voiceless sibilants			s	S			
voiced sibilants			z	Z			
voiceless fricatives		f		C	x	K	h
voiced fricatives		v					
trill			r			R	
nasals	m m=		n n=	n ~	N N=		
lateral			l l=				
tap			rr				
glides				j			

Table 6.11: German Consonant Examples and Descriptions

Worldbet	OGI	Example	Description
p	p	passe	voiceless bilabial stop
pc	pcl		voiceless bilabial stop closure
b	b	Bass	voiced bilabial stop
bc	bcl		voiced bilabial stop closure
t	t	Tasse	voiceless alveolar stop
tc	tcl		voiceless alveolar stop closure
d	d	das	voiced alveolar stop
dc	dcl		voiced alveolar stop closure
k	k	Kasse	voiceless velar stop
kc	kcl		voiceless velar stop closure
g	g	Gasse	voiced velar stop
gc	gcl		voiced velar stop closure
f	f	fasse	voiceless labiodental fricative
v	v	wasser	voiced labiodental fricative
s	s	Satin	voiceless alveolar sibilant
z	z	Satz	voiced alveolar sibilant
S	sh	Schatz	voiceless alveo-palatal sibilant
Z	zh	Genie	voiced alveo-palatal sibilant
C	cx	Reich	voiceless palatal fricative
x	kx	Rauch	voiceless velar fricative
h	h	hasse	voiceless glottal fricative
pf	pf	Pfennig	voiceless labiodental affricate
pfc	pfcl		voiceless labiodental affricate closure
ts	ts	Zeit	voiceless alveolar affricate
tsc	tscl		voiceless alveolar affricate closure
tS	ch	Deutsch	voiceless alveo-palatal affricate
tSc	chcl		voiceless alveo-palatal affricate closure
dZ	jh	Dschungel	voiced alveo-palatal affricate
dZc	jhcl		voiced alveo-palatal affricate closure
rr	r	brauchen	alveolar retroflexed tap
K	x	brauchen	uvular fricative
r	rr	rasse	alveolar trill
R		Narren	uvular trill
m	m	Masse	bilabial nasal
n	n	nasse	alveolar nasal
N	ng	hangen	velar nasal
n ~		Kognak	palatalized nasal
l	l	lasse	alveolar lateral approximant
m=	em	haben	syllabic bilabial nasal
n=	en	hatten	syllabic alveolar nasal
N=	eng	Haken	syllabic velar nasal
l=	el	Kessel	syllabic lateral alveloar
j	y	Jacke	palatal approximant

6.4.3 Notes on German Consonants

- The alveolar tap, **rr** was added to the Worldbet set.

6.5 Hindi

6.5.1 Hindi Vowels

Table 6.12 is a general vowel chart of Hindi vowels, and Tables 6.13 and 6.14 are lists of the labels used in Hindi.

Table 6.12: Hindi Vowels

	front	central	back
high	i: I	 Ix	u: u U
mid	e: E @	 ^ &	o:
low	 a a:		 >

Table 6.13: Hindi Vowel Examples and Descriptions

Worldbet	OGI	Example	Description
i:	iy	miit "beloved"	high front long
I	ih	mit.ana "wipe out"	mid-high mind-front short
Ix	ix		centralized I
e:	ey	sher "lion"	mid-high front long
E	eh	heh "is"	mid-low front short
@	ae	menaa "sparrow"	mid-low front long
^	ao	rath "chariot"	mid central
&	ax		mid-low central short
u:	uw		high back rounded long
u	uw	muuk "quiet"	high back rounded
U	uh	ruk "stop"	mid-high back short
o:	ow	mot.aa "fat"	mid-high back rounded
>	ao	aur "and"	mid-low mid-back
a	aa	pataa "address"	low front
a:	aa		low front long

Table 6.14: Hindi Diphthongs

Worldbet	OGI	Example	Description
ai	ay	bhaiyaa	a ⇒ i
aU	aw	laut.	a ⇒ U

6.5.2 Notes on Hindi Vowels

- Some vowel length distinctions were not made in Hindi in the first release of the multi-language database. **a** was labeled as short and **u** was always labeled long.
- We added the symbol ^ to Hindi.

6.5.3 Hindi Consonants

Hindi has labial *lab*, dental *dent*, alveolar *alv*, retroflex *ret*, palatal *pal*, velar *vel*, and glottal *gl* consonants.

Table 6.15: Hindi Consonants

	lab	*dent*	*alv*	*ret*	*alveopal*	*pal*	*vel*	*gl*
voiceless unaspirated stops	p	t[		tr			k	q
voiced unaspirated stops	b	d[		dr			g	
voiceless aspirated stops	pH	t[H		tR			kH	
voiced aspirated stops	bH	d[H		dR			gH	
voiceless unaspirated affricates					tS			
voiced unaspirated affricates					dZ			
voiceless aspirated affricates					tSH			
voiced aspirated affricates					dZH			
voiceless fricatives	f		s		S	C		h
voiced fricatives	v		z		Z			
nasals	m		n[	nr		n~	N	
lateral			l					
trill			r(					
unaspirated taps			rr					
aspirated taps			rrH					
glides	w				j			

Table 6.16: Hindi Consonant Examples

Worldbet	OGI	Example	Description
p	p	pal "moment"	voiceless labial plosive
pH	ph	phal "fruits"	voiceless labial aspirated plosive
pc	pcl		voiceless labial closure
b	b	bal "strength"	voiced labial plosive
bH	bh	bhaiyaa "brother"	voiced aspirated labial plosive
bc	bcl		voiced labial closure
t[	dt	pataa "address"	voiceless dental plosive
t[H	dth	path "path"	voiceless aspirated dental plosive
t[c	dtcl		t[and t[H closure
d[	dd	din "day"	voiced dental plosive
d[H	ddh	dhan "money"	voiced aspirated dental plosive
d[c	ddcl		d[and d[H closure
tr	t	t.uut.aa "broken"	voiceless retroflex plosive
tR	th	th.andaa "cold"	voiceless aspirated retroflex plosive
trc	tcl		voiceless dental plosive
dr	d	d.aal "branch"	voiced retroflex plosive
dR	dh	dh.er "lots"	voiced aspirated retroflex plosive
drc	dcl		voiced dental closure
k	k	kaam "work"	voiceless velar plosive
kH	kh	khaanaa "food"	voiceless aspirated velar plosive
kc	kcl		voiceless velar closure
g	g	gam "sorrow"	voiced velar plosive
gH	gh	ghar "home"	voiced aspirated velar plosive
gc	gcl		voiced velar closure
q			glottal plosive
tS	ch	cammac "spoon"	voiceless alveo-palatal affricate
tSH	chh	chat "roof"	voiceless aspirated alveo-palatal affricate
tSc	chcl		tS and tSH closure
dZ	jh	jaan "life"	voiced alveo-palatal affricate
dZH	jhh	jhaarna "waterfall"	voiced aspirated alveo-palatal affricate
dZc	jhcl		dZ and dZH closure
f	f	fatnaa "tear"	voiceless labial fricative
v	v		voiced labial fricative
s	s	siimaa "limit"	voiceless alveolar fricative
z	z	zamiin "land"	voiced alveolar fricative
S	sh	sher "lion"	voiceless alveo-palatal fricative

Table 6.17: Hindi Consonant Examples (continued).

Worldbet	OGI	Example	Description
Z	zh		voiced alveo-palatal fricative
C	cx		voiceless palatal fricative
h	h	hat. "move"	voiceless glottal fricative
m	m	man "mine"	bilabial nasal
n[	n	naam "name"	dental nasal
nr	nr	gun.ii "talented"	retroflex nasal
n~	ny		patalized nasal
N	ng	lihng "gender"	velar nasal
l	l	lataa "tendril"	alveo-palatal lateral
r(	r	savera "morning"	r flap, not retroflexed
rr	rd	per. "tree"	retroflex plosive
rrH	rdh	parh.aaii "studies"	aspirated retroflex plosive
j	y	yaad "person's name"	palatal glide
w			bilabial glide

6.5.4 Notes on Hindi Consonants

- All flapping is allophonic in Hindi, and will be transcribed using the flapping diacritic (_(). Flapping commonly occurs with /b/, /d/, /g/, /n/, and /rr/.
- Worldbet /n$\tilde{}$/ has not been included as a base symbol because it is not phonemic. When a palatized n occurs it will be labeled **n_j**.

6.6 Japanese

6.6.1 Japanese Vowels

Table 6.18: Japanese Vowels

	front	central	back
high	i i:		u u:
mid	e e:		o o:
		& ^	
low	a a:		

Table 6.19: Japanese Vowel Examples and Descriptions

Worldbet	OGI	Example	Description
i	iy	ichi "one"	high front unrounded
i:	iy:	iie "no"	high front unrounded long
e	ey	koe "voice"	mid front
e:	ey:	sensei "teacher"	mid front long
4	uw	uta "song"	high back unrounded
4:	uw:	futsuu "ordinary"	high back unrounded long
o	ow	igo "Igo game"	mid back rounded
o:	ow:	tookyoo "Tokyo"	mid back rounded long
a	aa	san "three"	low central vowel
a:	aa:	apaato "apartment"	low central long vowel
&	ax		mid central long or short
&_0			a common voiceless allophone of &
^	ah		mid central stressed

Table 6.20: Japanese Diphthongs

Worldbet	OGI	Example	Description
aI	ay	hai "yes"	a ⇒ I
oI	oy		o ⇒ I
i4			i ⇒ 4
eI	ey		e ⇒ I

6.6.2 Notes on Japanese Vowels

- Vowel length is phonemic in Japanese. In previous releases of the multi-language corpus long vowels were transcribed with double vowel labels rather than with the vowel label plus colon.
- In word-final position, a nasalized vowel (duplicated from the preceding vowel) may be used instead of the velar nasal, N (see Table 6.21).

- The symbols **4** and **4:** were labeled as **u** and **u:** in previous releases of the multi-language database.

6.6.3 Japanese Consonants

Japanese has bilabial, dental, alveolar, palatal, and velar consonants. These are abbreviated in the consonant chart as follows: *bilab* (bilabial); *dent* (dental); *alv*(alveolar); *pal* (palatal); *vel* (velar).

Table 6.21: Japanese Consonants

	bilab	*dent*	*alv*	*pal*	*vel*
voiceless plosives	p	t[			k
voiced plosives	b	d[			g
voiceless affricates		t[s		tS	
voiced affricates		d[z		dZ	
voiceless fricatives	f		s	S	
voiced fricatives		z			h
nasals	m	n[			N
lateral			l(		
glides	w			j	

6.6.4 Notes on Japanese Consonants

- Length is phonemic for Japanese consonants. Long consonants will be indicated by placing a colon after both the closure and release label, as in [k i t[c: t[: aa].
- The labels /t[s/, /d[z/, /tS/, and /dZ/ are allophones of the phoneme /z/. /t[s/ and /d[z/ are used before /uw/; /tS/ and/dZ/ are used before /i/.
- We added the symbol /r(/ to the Worldbet label set.

Table 6.22: Japanese Consonant Examples and Descriptions

Worldbet	OGI	Example	Description
p	p	papa "papa"	voiceless bilabial plosive
pc	pcl		voiceless bilabial closure
b	b	abu "horsefly"	voiced bilabial plosive
bc	bcl		voiced bilabial closure
t[	t	uta "song"	voiceless dental plosive
t[c	tcl		voiceless dental closure
d	d	eda "branch"	voiced dental plosive
d[c	dcl		voiced dental closure
k	k	aka "red"	voiceless velar plosive
kc	kcl		voiceless velar closure
g	g	igo "Igo game"	voiced velar plosive
gc	gcl		voiced velar closure
t[s	ts	itsu "when"	voiceless dental affricate
t[sc	tscl		t[s closure
tS	ch	ichi "one"	voiceless palatal affricate
tSc	chcl		tS closure
d[z	dz	izu "izu"(place name)	voiced dental affricate
d[zc	dzcl		d[z closure
dZ	jh	iji "taste"	voiced palatal affricate
dZc	jhcl		dZ closure
F	f	fu:fu "couple"	voiceless labial fricative
x	f	hito "human"	voiceless velar fricative
h	hh	voiceless glottal fricative	
s	s	asa "morning"	voiceless dental sibilant
z	z	aza "bruise"	voiced dental sibilant
S	sh	ishi "stone"	voiceless palatal fricative
Z	zh	aho "fool"	voiced palatal fricative
m	m	ama "mom"	bilabial nasal
n[	n	ana "hole"	dental nasal
N	ng		uvular nasal
l(			voiced alveolar lateral flap
w	w	awa "bubble"	bilabial glide
j	y	ayu "sweetfish"	palatal glide

6.7 Mandarin

6.7.1 Mandarin Vowels

Table 6.23: Mandarin Vowels

	front	central	back
high	y i: I	If	u
mid	E	&	
low	@	A	>

Table 6.24: Mandarin Correspondence Chart: Pinyin and OGI

Worldbet	Pinyin	OGI	Description
y	ö	iyw	high front rounded
i:	i	iy	high front tense
I	i	ix	high front unrounded lax vowel
If		uu	very high front, almost fricated
E	a	eh	mid front lax vowel
&		ax	mid reduced vowel
u	u	uw	high back vowel
4r	e	oe	mid-back unrounded retroflexed vowel
2		ox	mid back unrounded vowel
&r		er	mid central retroflex vowel
>	o	ao	mid-low back vowel
A	a	aa	low central vowel
@		ae	low front vowel

Table 6.25: Mandarin Diphthongs

Worldbet	OGI	Description
ai	ay	aa ⇒ i
aU	aw	aa ⇒ U
ei	ey	e ⇒ i
oU	ow	o ⇒ U

6.7.2 Mandarin Consonants

Following is the correspondence chart between the Pinyin, Worldbet and OGI symbols.

6.7.3 Notes on Mandarin Consonants

- The diacritic /_h/ attached to a label normally represents aspiration. However, in Mandarin, it has been attached to vowels when the speaker is whispering (i.e. voicelessness).

Table 6.26: Correspondence Chart for Pinyin and Worldbet: Consonants

Worldbet	**Pinyin**	**OGI**	**Description**
p	b	p	voiceless bilabial plosive
ph	p	ph	voiceless aspirated bilabial plosive
pc		pcl	p or pH closure
t[	d	t	voiceless dental plosive
t[h	t	th	voiceless aspirated dental plosive
tc		tcl	t or th closure
k	g	k	voiceless velar plosive
kh	k	kh	voiceless aspirated velar plosive
kc		kcl	k or kH closure
ts	z	ts	voiceless affricate
tsh	c	tsh	voiceless aspirated affricate
tsc		tscl	ts and tsH closure
tsr	zh	tsr	voiceless retroflex affricate
tshr	ch	tshr	voiceless aspirated retroflex affricate
tsrc		tsrcl	tsr and tsR closure
cC	j	c	voiceless palatal affricate
chC	ch	ch	voiceless aspirated palatal affricate
cCc		chcl	c and chC closure
f	f	f	voiceless labiodental fricative
s	s	s	voiceless aspirated dental fricative
sr	sh	shr	voiceless retroflex fricative
C	C		voiceless palatal fricative
h	h	hh	voiceless glottal fricative
m	m	m	bilabial nasal
n	n	n	alveolar nasal
N	N	ng	velar nasal
l	l	l	alveodental liquid
r+	r	r	advanced /r/
w	w	w	labiovelar glide
j	y	y	palatal glide

Because tones cannot be detected without voicing, they are rarely labeled in these cases.

- The allophone /jw/ has not been included due to its low frequency.
- Mandarin Chinese has a very strict syllable structure. although there can be a great variety of combinations of vowels in syllables, some having monophthongs, some diphthongs and some triphthongs, the consonant structure is comparatively rigid. The syllable pattern is (C)(V)V(V or N). Initial consonants are optional, and only nasals (usually n or N) are permitted word finally.
- The label **tshr** was labeled **tsR** in the first release of the multi-language corpus.
- The label **chC** was labeled **cCh** in the first release of the multi-language corpus.
- The label **x** was labeled as **x** in the first release of the multi-language corpus.

6.7.4 Mandarin Tones

There are 4 tones in Mandarin. They are labeled with the vowel, by number, as the vowel is a reliable marker of the syllable nucleus. Tone are marked on all vowels except whispered vowels and filled pauses.

Tone 1 High level tone
/i:_1/ "cloth"

Tone 2 High rising tone
/i:_2/ to suspect

Tone 3 Falling/rising tone
/i:_3/ chair

Tone 4 High falling tone
/i:_4/ meaning

For more information about tones and their phonemic values, see [**?**], [**?**] or [**?**].

Tone 3 may be shortened in rapid speech; it can change to Tone 1 in a word before a syllable with Tone 2 [**?**], either intra- or inter-word. The tone is labeled according to which phonemic tone it is closer to: Tone 1 or Tone 3.

At this time, the reduced tone (the "light tone") is labeled as Tone 1. This includes most of the Pin Yin "de" and "le" formations.

6.8 Spanish

6.8.1 Spanish Vowels

Table 6.27: Spanish Vowel Examples and Description

Worldbet	OGI	Example	Description
i	iy	niño "child"	high front tense vowel
Ix	ih	simple "simple"	high reduced vowel allophone
e	ey	bebé "baby"	mid front tense vowel
E	eh	pero "but"	mid front lax vowel allophone
&	ax		mid central reduced vowel allophone
^	ah		neutral vowel allophone
u	uw	duda "doubt"	high back tense vowel
o	ow	boda "marriage"	mid back tense vowel
a	aa	papá "papa"	low central vowel

Table 6.28: Spanish Diphthong Examples and Description

Worldbet	OGI	Examples	Description
oi	oy	estoy	o ⇒ i
au	aw	autobús	a ⇒ u
ai	ay	caray	a ⇒ i

6.8.2 Spanish Consonants

Spanish has consonants at seven places of articulation. They are abbreviated in the Spanish consonant chart as follows: *bilab*(bilabial), *ld* (labiodental), *inter* (interdental), *dent* (dental), *pal* (palatal), *vel* (velar), and *gl* (glottal).

6.8.3 Notes on Spanish Consonants

- The label **hs** represents the sylable final replacement of /s/ by aspiration in continuous speech. ($la^h cosa^h bonita^s$, for 'las cosas bonitas.' This phonological process is common in many varieties of Spanish including Andalucian, Caribbean, Pacific Coastal Spanish, and various Central and South American varieties. [2]
- **T** is phonemic in certain dialects of Castellano, including Madrid and other central and northern locations in Spain.
- **Z** and **dZ** are dialectal variants of the unmarked **j** pronunciation of orthographic ll. **dZ** tends to be word initial for dialects that have it.
- The label **s_j** is for the palatilized s in certian dialects of Spanish, especially in Madrid.
- To label /r/ vs. /r(/: if the wave form has more than one obvious burst it is labeled /r/. Trills occur word initially at times, but they are relatively rare in continuous speech.

  Sometimes the **r(** segment will contain a small closure and then a short vowel-burst-like segment. The closure is included in the segment when it occurs.

[2] Courtesy of Barrutia and Schwegler, *Fonética y fonologia Españolas*, 1994.

Table 6.29: Spanish Consonants

	bilab	*ld*	*inter*	*dent*	*pal*	*vel*	*gl*
voiceless plosives	p			t[		k	
voiced plosives	b			d[		g	
voiceless affricates					tS		
voiced affricates					dZ		
voicless fricatives		f	T	s	s_j L	x	hs
voiced fricatives	V		D	z	Z	G	
trills				r			
flaps				r(			
nasals	m			n	n~	N	
approximants		w		l	j L		

Table 6.30: Spanish Consonant Examples and Description

Worldbet	**OGI**	**Example**	**Description**
p	p	poco "little"	voiceless bilabial plosive
pc	pcl		voiceless bilabial closure
b	b	boca "mouth"	voiced bilabial plosive
bc	bcl		voiced bilabial closure
t[	t	tengo "I have"	voiceless dental plosive
tc	tcl		voiceless dental closure
d[	d	dentro "inside"	voiced dental plosive
dc	dcl		voiced dental closure
k	k	coma "coma"	voiceless velar plosive
kc	kcl		voiceless velar closure
g	g	goma "glue"	voiced velar plosive
gc	gcl		voiced velar closure
tS	ch	chica "little girl"	voiceless palatal affricate
tSc	chcl		voiceless palatal affricate closure
dZ	jh	llama "he calls"	voiced palatal affricate
dZc	jhcl		voiced palatal affricate closure
V	bx	sabio "wise"	voiced bilabial fricative
f	f	favor "favor"	voiceless labiodental fricative
T	tx	plaza "plaza"	voiceless interdental fricative
D	dx	cada "each"	voiced dental fricative
s	s	sol "sun"	voiceless dental sibilant
hs	hs	está<u>s</u> "you are"	aspiration replacing **s**
z	z	desde "since"	voiced dental sibilant
s_j	sh	dos "two"	voiceless palatalized /s/
Z	jh	ella "she"	voiceless palatal sibilant
x	hx	jota "letter j"	voiceless velar fricative
G	gx	lago "lake"	voiced velar fricative
r	rr	perro "dog"	alveolar trill
r(	r	pero "but"	alveolar, retroflex flap
m	m	matar "to kill"	bilabial nasal
n	n	nadar "to swim"	dental nasal
n~	ny	niño "child"	palatal nasal
N	ng	cinco "five"	velar nasal
l	l	lana "wool"	dental lateral
w	w	hueso "bone"	labiovelar glide
L	ly	tortilla "tortilla"	palatal lateral approximant
j	y	llorar "to cry"	palatal glide

Chapter 7

Diacritics

Diacritics			
IPA	Worldbet	OGIbet	Type of Diacritic
t^h	_h	-h	aspirated
	_x		centralized
t d	_[		dental
	_(		flapped (consonant)
	_F		fricated stop
	_?*	q	glottal onset
	_?	-q	glottalized
d	_l		lateral release
i	_:	-el	lengthened
d^n	_n		nasal release
ẽ	_~	-n	nasalized
	_NL	.nitl	not in the language
t^j	_j		palatalized
	_r	-r	retroflexion
	_i		less rounded
˛	_w		more rounded
	_=		syllabicity
	_v		voiced
n̥ d̥	_0		voiceless
	_*	-	waveform cut off
Worldbet, as modified at the CSLU			
	_fp	-fp	filled pause
	_ln	-ln	line noise corruption
	_bn		background noise

Table 7.1: Mapping between IPA, Worldbet and OGIbet Diacritics

7.1 Overview

Diacritics are used to show finer detail than the base symbol is designed to give. With few exceptions (mainly vowels and syllabics) base labels represent single phonemes while diacritics provide additional phonetic detail.

A diacritic is separated from the base label by the underscore. The number of diacritics used depends on what is needed to accurately describe the phone. There is no particular ordering for diacritics.

If there is noticeable spectral variation within the same basic phone it may be divided into multiple segments all of which contain the same base label but different diacritics. Vowels that become glottalized should be segmented into (at least) two parts, the first segment with the base label for the vowel, and the second segment with the vowel + diacritic.

A single phoneme can be segmented into more than two segments. Vowels may become glottalized for a period, then heavily aspirated, and finally devoiced. Thus a single phoneme may be represented in the transcription with any number of base labels each reflecting distinct phonetic variations. In order to arrive at a purely phonemic level transcription the diacritics must be removed and adjacent base labels collapsed.

Following is a description of all of the diacritics used in English labeling. Note that this is only a subset of the Worldbet diacritic inventory. For the complete set Diacritics appearing in this chapter are those that were most commonly needed when labeling English. This section will eventually be expanded to include additional multi-language diacritics.

7.2 Aspiration

The diacritic **_h** indicates excessive aspiration on a phone. Aspiration may be evident in relaxed speech on a vowel when the vocal folds are still vibrating but breath increases. If a phone becomes devoiced, the devoicing diacritic (**_0**) is used and aspiration is assumed and therefore need not be explicitly marked.

In order to use the aspiration diacritic on a vowel, the formants of the vowel must remain strong. Aspiration following a vowel that does not contain strong formants should be labeled **.br**, although it may retain some vowel-like quality when heard in isolation.

Predictable aspiration is contained within the base label, i.e. **th** in English contrasts with **t** in German, as English stops are in general more aspirated than German plosives. If an English alveolar plosive were unusually heavily aspirated, the transcription would be **th_h**. The latter transcription convention is rarely used because the base label already specifies aspiration in the phone.

There are certain predictable contexts in which English voiceless plosives do not contain aspiration, such as "st" clusters. Lightly or non-aspirated stops are not marked explicitly in transcriptions because they can be predicted by phonological rule.

7.3 Centralization

The centralized diacritic is used for vowels that have become centralized or reduced but that still perceptually contain elements of the original vowel quality. If the placement of a vowel has moved slightly central, this diacritic will be used. In fast speech the vowel **I** often appears as a very short and somewhat reduced vowel. F2 is still too high for it to be considered ʌ or **&**. It

should be transcribed **I_x**.

7.4 The Flapping Diacritic

The _(diacritic is used for flapped consonants. When stops are flapped they are extremely short, lack closure, and have visible formants. In the spectrogram they often look like a small dip in between vowels and have a slightly lower amplitude than surrounding segments. Flapping is very common with alveolar stops in American English. Usually a flapped segment is not much longer than 30ms, although the actual length depends on the rate of the speaker. The alveolar nasal in English is also commonly flapped.

7.5 Fricated Stops

The diacritic _F is used to indicate that a stop burst has been heavily fricated. Fricated stops are plosives having no true closure, but a sequence of weak frication followed by strong frication. This sequence mimics the closure and aspiration of the regular stop.

7.6 Glottal Onset

The diacritics _? and _?* are in complementary distribution. The diacritic ?* is used to mark the few glottal pulses occurring in word or utterance initial vowels in American English. Glottal onset is a very short "catch" occurring in the pharynx as the vocal folds are beginning to vibrate. Glottal onset is acoustically identical to a glottalized vowel, but it only occurs in the word or utterance initial environment, and it is of limited duration (one to at most three glottal pulses). The glottalized diacritic, _?, may appear anywhere the _?* does not appear, or it may appear in word or utterance initial position if the glottalization is of long duration.

7.7 Glottalization

The diacritic _? is used to label glottalization. Any voiced phone can be glottalized, but vowels are the most commonly glottalized phones.

Glottalized vowels are characterized by a marked slowing of vocal fold vibration. They differ significantly in appearance from non-glottalized vowels. Utterance final vowels are often glottalized as the vocal folds cease moving. Vowel geminates are also often glottalized in the period of transition between the two phonemes. When spacing of vocal striations becomes relatively (and significantly) larger relative to immediate spectral context, the glottalization diacritic should be used.

7.7.1 Marking glottalization on diphthongs

It can be difficult to determine what base label to use if only the initial part of a diphthong is glottalized. It does not seem reasonable to mark a segment **aI_?** unless the glottalized portion sounds like the full diphthong. Often the glottalized portion at the beginning of a diphthong sounds like a single vowel rather than a diphthong.

When the initial portion of a diphthong is glottalized, if there is movement in the formant of the segment and if the quality of the vowel sounds like the entire diphthong, use the entire

diphthong label; for example **aI_?**. However, if the glottalized portion of the diphthong only bears the quality of the initial sound in the diphthong, use the following conventions:

diphthong	base vowel
aU	@_?
aI	A_?
iU	i:_?
>i	**>**_?
ei	ei_?
oU	oU_?

Notice that for **ei** and **oU** the entire diphthong label is used regardless of whether or not the offglide is heard. This is done because there is no specific label for either **e** or **o** in English. Also, the initial glottalized portion of the diphthong **aU** becomes **@_?**, while **aI** becomes **A_?**. The conventions are established this way for optimal descriptiveness, **@** being the most common initial sound in **aU**, and **A** being the most common initial sound in **aI**.

7.7.2 Glottal /t/

A special use of the glottalization diacritic is with the phoneme /t/. In American English, /t/ is often replaced by a glottal stop, as in the word "button." When a glottal stop replaces a /t/, it will be transcribed **th_?**.

7.8 Lateralization

Often in the context of a lateral segment vowels will become lateralized. At times, no distinct /l/ is produced, and the only artifact of the phoneme /l/ is lateralization on the vowel. The **_l** diacritic is used on vowels that have become lateralized when a distinct **l** cannot be seen in the spectrogram. This allows for recovery of a phonemic level transcription without requiring lexical knowledge. If an **l** is visible (usually a segment with slightly lower amplitude having similar formant frequencies as neighboring vowels) the lateralization diacritic should not be used as it is a predictable coarticulatory effect.

When the **_l** diacritic is used with the base label **l** it signifies lateral release. This usually looks like a vertical bar on the spectrogram and sounds like a clicking noise. It is a common articulatory effect caused by tongue movement during the release of the lateral phone **l**.

7.9 The Lengthening Diacritic

Length thresholds are difficult to set due to variable speaker rates. We quantify length respective to each speaker. If a given phone is significantly longer than similar phones in similar contexts by the same speaker, the **_:** diacritic should be considered.

The diacritic **_:** indicates relative lengthening. It is frequently used for vowels that are prolonged for emphasis or for a filled pause. The elongated symbol is often combined with the filled pause diacritic like this: **A_:_fp**. Based on the labeling we have done, we have determined that vowels longer than 300ms should receive this diacritic.

7.10 Nasal Release

The Nasal release diacritic, _**n** may only be applied to a base label that is a nasal. The nasal release is spectrally similar to the lateral release, and it also is caused by movement of the tongue as the nasal is released.

Often nasals are released with a trailing vowel. The vowel should be transcribed with a label describing the quality of the vowel, usually ∧. The word "ten" pronounced [tc th E n ∧] is a common example.

7.11 Nasalization

The diacritic _∼ denotes nasalization.

The nasalization diacritic is normally used on vowels or diphthongs. Nasalized vowels and diphthongs are spectrally distinguished by a split or separation in the first formant. This diacritic will be used when nasalization is not predictable. When nasalization can be predicted by phonological rule (i.e., when in the context of a neighboring nasal) it is not labeled.

Nasal deletion is a common reduction occurring in fast speech. Sometimes the only acoustic clue that a nasal phoneme occurred is nasalization on the vowel (as in some pronunciations of "mountain.") If acoustic or auditory evidence remains signaling nasality but no distinct nasal is evident in the signal, the nasal diacritic should be used on the vowel. Although this is a predictable environment for a nasal vowel, the nasalization diacritic should be used so that a phonemic level transcription can be reproduced without lexical knowledge.

7.12 Rhotacization

The diacritic _**r** is used to indicate r-coloring or rhotacization. Many vowels become retroflexed when they are followed by /r/, as in "beard, bared, bard, bored, poor, tire, and hour." [1] In the above examples rhotacization does not usually appear on the beginning of the vowel; F3 is at the level appropriate for that vowel. When retroflexion begins, F3 suddenly dips to a lower level. The _**r** diacritic should only be applied to the portion of the vowel in which F3 has reached a level indicating retroflexion for that speaker (2,000Hz for the average male voice).

There are two retroflex base labels: **3r** or **&r** in the English set. Traditionally these syllabic retroflex vowels have been distinguished from other retroflexed vowels because in most cases the retroflexion affects the entire vowel. Occasionally vowels with a different placement than **3r** or **&r** will behave like syllabic retroflexes in that the entire vowel will be retroflexed, but this is relatively rare. If it occurs, the _**r** diacritic may span the entire phone.

'tr" clusters often have retroflection in the aspiration, notable mainly because of third formant movement going into the vowel. Another common coarticulation effect in "tr" clusters is palatalization. Segment and listen for retroflexion on the /t/ and check the position of F3 at the onset of the vowel. If the level of F3 is low, consider retroflexion. If F3 is still quite high, but aspiration is heavier than normal on the /t/ burst, consider palatalization.

[1] Examples taken from [?].

7.13 Voicing and Voicelessness

The diacritics **_v** and **_0** are used to label voicing of normally voiceless segments and absence of voicing on normally voiced segments, respectively.

In a normally voiceless consonant, if there is voicing throughout the entire segment, use the voicing diacritic, **_v**.

Similarly, if a phonemically voiced consonant is completely devoiced, use the voiceless diacritic, **_0**. These two diacritics are not used for partially devoiced or partially voiced consonants because the threshold would be too subjective and labeling would lack consistency.

Because vowels have a higher amplitude, carry stress and are generally more clearly articulated than consonants, they should be segmented more precisely. If a portion of a vowel is devoiced and a portion is voiced, the portion that is devoiced should be segmented separately and given the devoicing diacritic. Note that the /z/ must be completely devoiced for the devoicing diacritic to be used where as the devoiced portion of the vowel is segmented separately regardless. When **_0** is used, aspiration is generally assumed and need not be explicitly marked.

7.14 Labialization

The diacritic **_w** is used for rounding of both consonants and vowels. Lip rounding may be indicated in a spectrogram by a drop in both F1 and F2. Velar stops are a common place for labialization, as in English 'quick' or the Cantonese name 'Kwai'; look for formants that emerge from the stop aspiration at an extremely low level. The phoneme /s/ is another common target of labialization: the frication energy will drop to 2000 Hz or lower, causing /s/ to look like [S].

7.15 Unrounding

The diacritic **_i** indicates unrounding on a normally rounded segment. Many American back vowels become unrounded before high front vowels. The high back vowel **u** in particular is often unrounded in dialects of American English.

7.16 Palatalization

.

The diacritic **_j** denotes palatalization. It signifies that the place of articulation of the phone is approaching the palatal place of articulation. When applied to a base label it does not necessarily mean that the phone is produced at the hard palate.

Palatalization of velar stops (especially before high front vowels) is a very common assimilation process in the world's languages, as is palatalization of /t/ before /r/ in English. When a /t/ is palatalized, it is usually more heavily aspirated and it sounds more like **tS**. The change is most obvious in the waveform where the aspiration is quite intense; the waveform looks very similar to that of a **S** or **tS**.

Palatalization often occurs in the transition between high front vowels to back vowels. It happens mainly with velar stops, and it destroys the velar pinch that would normally be seen between F2 and F3. In the sentence "We used to stand in queues," there will likely be palatalization of the /k/ in queues, as the mouth makes the transition from a high front vowel

to a back vowel. If there is no visible velar pinch in the spectrogram and the /k/ sounds palatalized (possibly approximating the sequence **kh j**), use the **_j** diacritic.

7.17 Cut Off Speech

Due to the nature of CSLU's telephone speech data collections, often a person's speech will be abruptly cut short. This has lead to the development of a cut off diacritic used when the signal ends abruptly. If a phoneme occurring at the end or beginning of a file has been cut short the diacritic _* should be used to distinguish it from spectral segments manifesting complete articulations.

Appendices

Appendix A

Terminology

A.1 Description

Appendix A defines linguistic terms used throughout this document. These definitions will be especially useful when examining the charts in Chapter 6. We have not deviated from standard linguistic theory in our use of these terms.[1]

A.2 Phonemes Versus. Allophones

A phoneme is a "distinctive sound" in a given language, which acts to contrast words. [?]. An allophone is "a predictable phonetic variant of a phoneme." [?]

Examples:

/b/ is a phoneme in English. To determine if a given phone is an phoneme, see if a minimal pair can be found. A minimal pair (or set) is two distinct words which differ in only one "meaningful" sound. The two words "bat" and "pat" illustrate that /b/ and /p/ are distinct phonemes in English. Other examples are "sought" and "fought," Notice that pronunciation is a factor, but not spelling.

/p/ is a phoneme in English which has a number of phonetic (allophonic) realizations. /p/ in the word "perfect" is aspirated, and spectrally averages 50-60ms in length. /p/ in the word "spectacle" is generally unaspirated, and the average length is shorter than its aspirated counterpart. Both phones are bilabial plosives, and both have the same percept for native speakers of English. As the percept is the same for both the aspirated and unaspirated bilabial stops, there is no "meaningful" difference between these sounds; they are allophones.

Another allophonic form of /p/ is unreleased. Many speakers do not release word final /p/ as in the word "stop." This would contrast with the word "petunia," where word inital /p/ is always released.

A.3 Consonants:

A.3.1 Place of Articulation

The following terms are used to describe place of articulation of consonants: [?]

[1] For a more complete coverage of Linguistic terminology, see [?] and [?].

Bilabial: Produced by bringing the two lips together.

Labiodental: Produced by moving the lower lip to the upper front teeth.

Interdental: Produced with the tip of the tongue between the upper and lower teeth. These are also called "dental."

Alveolar: Produced with the tip or blade of the tongue raised to the alveolar ridge.

Retroflex: Produced with the tip of the tongue and the back of the alveolar ridge. [**?**]

Palatal-Alveolar: Produced with the blade of the tongue and the back of the alveolar ridge. [**?**]

Palatal Produced with the front of the tongue and the hard palate. [**?**]

Velar: Produced by raising the back of the tongue to the velum.

Uvular: Produced with the back of the tongue at the uvula.

Glottal: Produced at the glottis.

A.3.2 Manner of Articulation

The following terms are used to describe the manner of articulation:

Stop Airflow is completely cut off by the closure of the articulators; pressure builds up and is released. Stops occur in the initial sounds of the words *buy*, *toy*, and *dog*.

Nasal Nasals are produced when the soft palate is lowered and air is allowed to flow through the nasal passage. Examples are the final sounds in *ram* and *cocoon*.

Fricative Fricatives involve partial closure, such that the air flowing between the articulators is turbulent. The words *float*, *veil*, *sing* and *zip* begin with fricatives. The first two, /f/ and /v/ are weak fricatives, and /s/ and /z/ are strong (strident) fricatives.

Affricate A combination of a stop followed by a fricative. The words *church* and *judge* begin with affricates.

Tap or flap Produced by a single tap of the tongue against the alveolar ridge. In the American dialect the phone in the middle of the word *letter* is a tap.

Trill Produced by multiple, rapid taps of the tongue. The Spanish and French *r*'s are often trilled.

Approximant or semivowel Sound produced when one articulator is close to another, but not close enough to produce a turbulent airstreem. The words, "world" and "yes" begin with approximants.

A.3.3 Additional Features

Voiced Sounds produced when the vocal cords are vibrating.

Voiceless Sounds produced when the vocal cords are apart.

Aspiration Audible release of air during production of a phone.

A.4 Vowels

The tables in Chapter 6 describe each vowel using the terms **high, mid, low, front, central,** and back. These words describe where the vowel is produced in the mouth. The following chart illustrates the location denoted by each term. **Front, central**, and **back** describe the part of the tongue that is the highest during production, while **high, mid**, and **low** describe the height of the tongue.

To say that a vowel is high front, for example, means that the front (blade) of the tongue is high in the mouth when the phone is produced. A mid back vowel means that the back (body) of the tongue is raised to a "mid" height (i.e. raised slightly).

Table A.1: Vowel Chart

	front	central	back
high			
mid			
low			

This table is a blank vowel chart which could be filled with any of the similar charts of chapter 8. Its form mimics the cardinal vowel triangle, which graphically displays vowel place of articulation.

Bibliography

[1] George D. Allen. The phonascii system. *Journal of the International Phonetic Association*, 18:9–25, 1988.

[2] A. Evison A.P. Cowie. *Concise English-Chinese Chinese-English Dictionary*. Oxford University Press, Oxford, England, 1986.

[3] International Phonetic Association. Report on the 1989 kiel convention. *Journal of the International Phonetic Association*, 19:67–80, 1989.

[4] W. J. Barry and A. J. Fourcin. Levels of labelling. *Manuscript*, 1990.

[5] Ronald Cole, Beatrice T. Oshika, Mike Noel, Terri Lander, and Mark Fanty. Labeler agreement in phonetic labeling of continuous speech. In *ICSLP Conference Proceedings*, august 1994.

[6] Bernard Comrie, editor. *The World's Major Languages*. Oxford University Press, Oxford, first edition, 1990.

[7] CSLU. Ogi speech tools user's manual. Technical report, Center for Spoken Language Understanding, Oregon Graduate Institute, 1993.

[8] Victoria Fromkin and Robert Rodman. *An Introduction to Language*. Holt, Rinehart and Winston, Inc., New York, fourth edition, 1988.

[9] J. Hieronymus, M. Alexanberd, C. Bennett, I. Cohen, D. Davies, J. Dalby, J. Laver, W. Barry, A. Fourcin, and J. Wells. Speech segmentation criteria for the scribe project. *Manuscript*, 1990.

[10] James L. Hieronymus. Ascii phonetic symbols for the world's langauges: Worldbet. Technical report, Bell Labs, 1993.

[11] Peter Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, third edition, 1993.

[12] Terri Lander, Ron Cole, Beatrice Oshika, and Mike Noel. Multi-language speech database: Creation and phonetic labeling agreement. In *Eurospeech95 Conference Proceedings*, september 1995.

[13] Terri Lander, Beatrice Oshika, Ron Cole, and Mark Fanty. Multi-language speech database: Creation and phonetic labeling agreement. In *ICPhS Conference Proceedings*, august 1994.

[14] Ian Maddieson and Kristin Precoda. Upsid and phoneme. *manuscript*, 1992.

[15] M.I.T., Massachusetts. *Massachusetts Institute of Technology 6.67s Speech Spectrogram Reading*, July 1985.

[16] Yeshwant Muthusamy, Kay Berkling, Takayuki Arai, Ronald Cole, and Etienne Barnard. A comparison of approaches to automatic language identification. In *Eurospeech*, sept 1993.

[17] John J. Ohala and Brian W. Eukel. Explaining the intrinsic pitch of vowels. In R. Channon and L. Shockey, editors, *In Hone of Ilse Lehiste*. Foris, Dordrecht, 1987.

[18] S. Seneff and V. W. Zue. Transcription and alignment of the timit database. *TIMIT CD-ROM Documentation*, 1988.

[19] Timothy J. Vance. *An Introduction to Japanese Phonology*. State University of New York Press, Albany, New York, 1987.

Index