# Annotated Speech Corpora Development in Indian languages : A Technical Report

Centre for Development of advanced Computing (C-DAC), Kolkata
Plot E2/1, Block GP, Sector V,
Saltlake Kolkata-700091.
Phone no. 2357-9846/5989
Fax no. 2357-5141
www.cdackolkata.com

## CONTENTS



## Introduction

1.1 Objective: Development of Annotated Speech Corpora for speech research and speech technology for Indian languages.

1.2 Background:

In the present age of Information Technology (IT), information access to all in a convenient manner has gained utmost importance. In the current Indian context, the needed skills of both computer literacy as well as conversance with written English in the standard QWERTY English keyboard-oriented interfaces to Personal Computer (PC), restrict the usage of such essential access device to IT to a miniscule fraction of the population. Since speech is a primary mode of direct communication amongst human beings, it would be highly desirable from added convenience point of view to have direct speech-anchored dialogue between man and machine. Added gain could be the possibility to have such speech based interaction in local languages themselves instead of English. This demands full-scale research and development in the area of spoken language processing including speech recognition and speech synthesis in different Indian Languages. The development of such Speech Research and Technology on Acoustics/ Phonetics of different spoken languages demands large speech databases. The creation of annotated Speech Corpora is therefore of prime enabler for such research.

While corpora of written text exist in most of the Indian Languages, unfortunately none of these exist in spoken versions of them. The absence of standard speech corpora in regional languages has been one of the major retarding factors behind the lack of adequate progress in development of automatic speech processing research, particularly related to man-

machine-man interactions, and the evaluation of developed tools. In addition, standard corpora also can be used to measure current state-of-the-art performance in research areas for particular tasks and highlight deficiencies that require further research.

The major issues for building up such Speech Corpora are the selection of dialect, content, informants and recording environment. Proper annotation and marking of linguistics and phonetic units are necessary. In the selected dialect, it should adequately cover speaking pattern of native informants, normal as well as under different emotions, in different types of sentences and discourses.

Speech mode interaction between man and machine in regional languages has a paramount importance in India as because about 40% of the Indian population is deficient of functional literacy, while about 80% of them have very little or no knowledge of English. Lack of proper access to IT leaves this vast disadvantaged population, out of the fruits of e-technology. An easy and convenient way to empower them is through the establishment of a direct speech-oriented access to them to the global information base.

1.3 Need of Speech Corpora:

Though there exist Text Corpora in major Indian Languages, these are of little significance for speech research and technology. Usually, the spoken language is at large variance with the written one. The spoken lexicon may be widely different from the graphemic lexicon as well as the syntax and grammar. Furthermore, the acoustic parameters of spoken units of sounds and prosodic structure for each of the languages, even if one restricts oneself to the standard dialect, need accurate identification for each of the languages. As all of them are rather nebulous in nature, statistical requirement needs large-scale data analysis. It will be impossible to derive the requisite knowledge base for the speech technology without a well-structured database. It is, therefore, imperative to built standard speech corpora (databases) for all major Indian languages. It will also provide the necessary base-line required for bench-marking various developed solutions in speech technology.

1.4 Methodology: The corpora creation methodology can be divided in following four parts:

(a) Content Selection,

(b) Recording,

(c) Tagging and

(d) Organization

## Content Selection

There exists adequate experience in text corpora building in most of the major Indian Languages. This is primarily useful for Natural Language Processing (NLP), the backbone of text-oriented technology development, such as Machine Translation (MT), Searching and Information Retrieval, etc. Unfortunately Spoken Language Processing (SLP) has an additional dimension, which cannot be addressed purely by mimicking these NLP procedures. Particularly one must respect the importance of prosodic information present in spoken language; and also the gaps/absence of strict grammatical structures of written text in speech. It is, therefore, necessary to respond adequately to the differences between formalisms in textual language against their informality in spoken language while selecting the contents of Speech Corpora.

However, due to primary emphasis of the Speech Corpora contents to various needs of technology development supporting man-machine-man communication instead of general linguistic research, the content selection naturally would be dictated by those specific considerations. In this context, Speech Corpora content can be divided largely into following four parts:

(a) Speech Research,

(b) Automatic Speech Recognition,

(c) Speaker Recognition and

(d) Speech Synthesis

The scematic in Figure 2.1 illustrates how these diverse needs influence the contents selection for speech corpora. The rectangular boxes indicate types of targeted contents; whereas the interconnecting arrowed lines highlight the intricate interactions supporting different technological vis-à-vis the research needs.

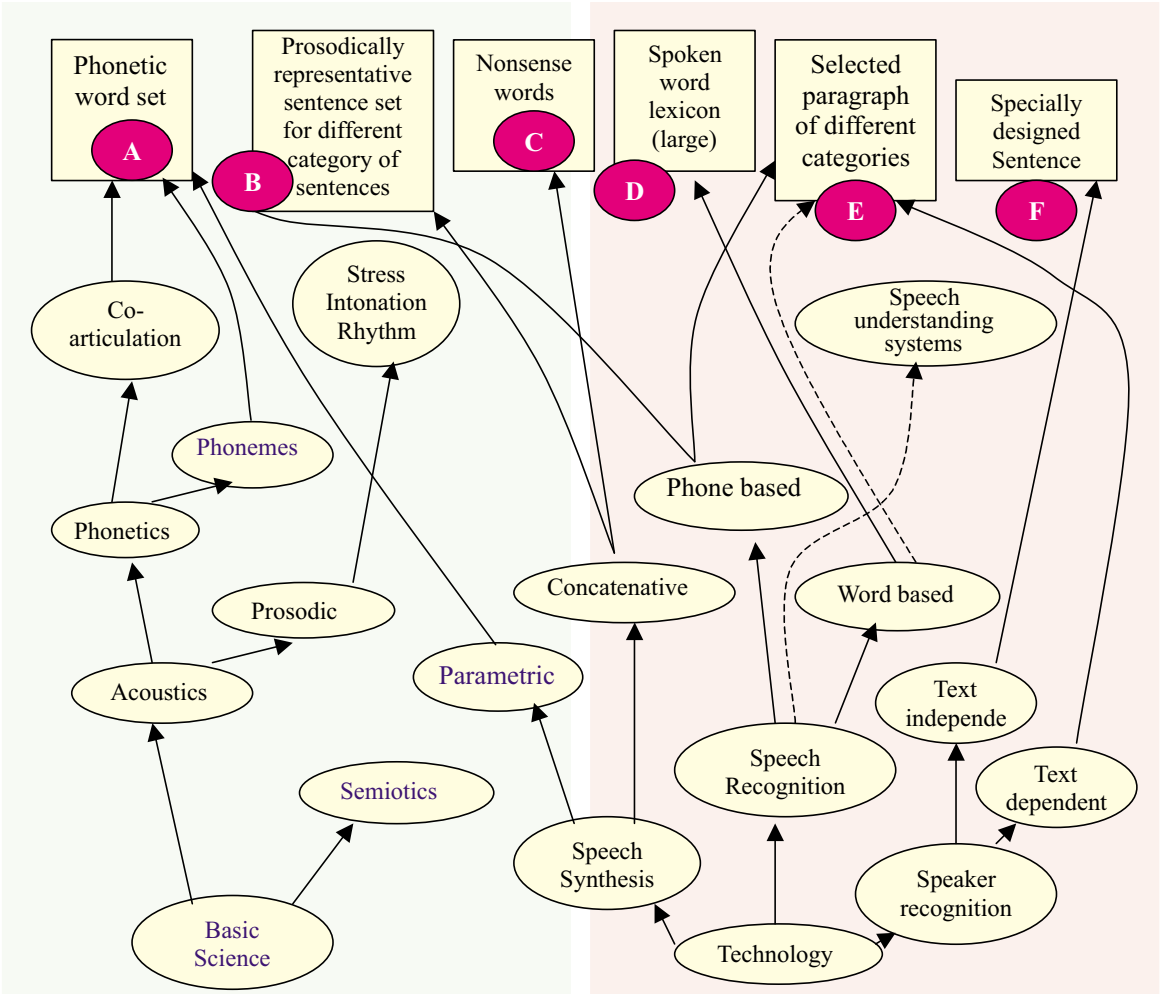


*Figure 2.1: Considerations for Content Selection for Speech Corpora*

2.1 Phonetic Word Set (A in figure 2.1): As the acoustic properties of a phoneme depend significantly on the adjoining phonemes, there need be recorded utterances of these phonemes in all possible bi-phone contexts. In case of Bangla Language, 600 commonly-spoken Bangla Words Set, which contain all bi-phone of the Bangla Phonemes (including the diphthongs), have been selected. In some cases, particularly where phoneme-combination could not be not obtained within the commonly used Bangla Words, requisite onomatopoetic words or non-sense words were included.

Table 2.1 outlines the process of creation of this set; numbers inside the cells points to the respective word housing that indexed bi-phone right at the mid-position.

Table 2.1. Bi-phone context of vowel and consonant for Bangla

| Preceding Vowel | | | | | | | | Con | Succeeding Vowel | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| eE | o+ | o | u | i | e | aA | a | | a | aA | e | i | u | o | o+ | eE |
| 88 | 112 | 110 | 104 | 6 | 111 | 28 | 26 | k | 101 | 6,26 | 104 | 112 | 28 | 113 | | 103 |
| 114 | | 115 | 116 | 117 | 55 | 30 | 77 | kH | 118 | 119 | 55 | 57 | 30 | 31 | | 120 |
| 121 | | 7 | 51 | 122 | 123 | 17 | 124 | g | 125 | 36 | 126 | 127 | 123 | 51 | 22 | 262 |
| 299 | | 90 | 91 | 69 | 137 | 19,68 | 92 | gH | 92 | 128 | 68 | 69 | 19 | 129 | | 89 |
| 36 | | 86 | 93 | 94 | 56 | 34 | 98 | nG | | 85 | 56 | 34 | 93 | 50 | | X |
| 5 | | 47 | 13 | 132 | 57 | 8 | 82 | c | 131 | 28 | 132 | 8 | 133 | 134 | 82 | 88 |
| 142 | | 141 | 140 | 139 | 55 | 10 | 80 | cH | 135 | 31 | 55 | 136 | 81 | 10 | | 138 |
| 145 | 22 | 148 | 148 | 147 | 146 | 23 | 83 | j | 24 | 143 | 22 | 23 | 144 | 25 | | 145 |
| ? | | 52 | 52 | 153 | 152 | 26 | 166 | jH | 26 | 52 | 150 | 27 | 151 | 95 | | 264 |
| 107 | | 25 | 59 | 155 | 154 | 59 | 42 | t. | 156 | 157 | 25 | 158 | 160 | 159 | 42 | 263 |
| 165 | | 164 | 60 | 162 | 161 | 33 | 96 | TH. | 163 | 33 | 48 | 49 | 60 | 96 | | 265 |
| 298 | | 170 | 99 | 169 | 61 | 13 | 100 | d. | 98 | 167 | 97 | 61 | 13 | 168 | | 97 |
| 176 | | 175 | 174 | 63 | 173 | 16 | 62 | dH. | 171 | 16 | 62 | 63 | 174 | 76 | 266 | 172 |
| 282 | | 281 | 280 | 283 | 284 | 277 | 278 | nN | ? | 279 | 283 | 285 | 286 | 287 | | 288 |
| 183 | 3 | 35 | 182 | 4 | 78 | 16,177 | 181 | t | 46 | 108 | 177 | 3 | 178 | 180 | 179 | 87 |
| 71 | | 72 | 186 | 65 | 184 | 71 | 47 | tH | 42 | 71 | 184 | 65 | 185 | 47 | | 267 |
| 194 | | 195 | 54 | 189 | 193 | 11,12 | 192 | d | 187 | 188 | 190 | 189 | 12 | 191 | 112 | 190 |
| 199 | | 70 | 203 | 7 | 202 | 14,15 | 201 | dH | 70 | 7,14 | 196 | 197 | 15 | 198 | | 200 |
| 89 | 2 | 209 | 4 | 208 | 5 | 6,10 | 24 | n | 4 | 5 | 204 | 8 | 205 | 206 | | 207 |
| 20 | | 217 | 216 | 27 | 213 | 35 | 215 | p | 210 | 27 | 211 | 214 | 212 | 35 | | 36 |
| 224 | | 54 | 54 | 303 | 223 | 37 | 222 | pH | 218 | 54 | 219 | 37 | 220 | 221 | | 38 |
| 88 | 45 | 226 | 7,12 | 65 | 102 | 1,2,3 | 118 | b | 20 | 8,10 | 102 | 7 | 225 | 118 | 2,3 | 5 |
| 304 | | 229 | 230 | 17 | 102 | 7,24 | 35 | bH | 229 | 2,17 | 56 | 228 | 227 | 7 | 266 | 88 |
| 235 | | 236 | 237 | 238 | 239 | 23 | 2 | m | 231 | 23 | 232 | 106 | 233 | 234 | | 84 |
| 87 | | 96 | 19 | 8 | 243 | 14 | 4 | r | 14 | 4,20 | 240 | 162 | 241 | 242 | | 274 |
| 275 | | 10 | 246 | 53 | 245 | 27 | 44 | r. | 276 | 27 | 40 | 39 | 41 | 244 | | X |
| 272 | | 66 | 67 | 273 | X | 105 | X | rH. | 268 | 269 | 247 | 270 | 271 | 105 | 67 | 272 |
| 38 | | 30 | 30 | 11,37 | 107 | 16 | 46 | l | 248 | 11,30 | 57 | 32 | 249 | 37 | | 250 |
| 21 | | 11 | 255 | 254 | 78 | 12 | 79 | h | 21 | 12 | 78 | 11 | 251 | 252 | | 253 |
| 289 | | 292 | 293 | 291 | 290 | 42 | 106 | s | X | 295 | 296 | 294 | 305 | 297 | | 107 |
| 109 | | 125 | 54 | 1 | 257 | 43 | 258 | sS | 256 | 1 | 109 | 43 | 54 | 44 | 45 | 302 |
| 300 | | 259 | 260 | 9 | 261 | 74,75 | 73 | y | 301 | 14,29 | 57 | 75 | 74 | 259 | | X |
| X | X | X | X | 526 | 529 | 527 | 532 | w | X | X | X | X | X | X | X | X |

It may be noted that, out of many possible bi-phones, only one bi-phone is housed in a word. Tabular enumeration ensures the completeness of the created word set covering fully all possible bi-phones.

Table 2.2. Bi-phone context of vowel and vowel

| | Succeeding Vowel | | | |
|---|---|---|---|---|
| Preceding Vowel | | i | u | o e |
| | aA | 1 | 2 | 3 4 |
| | a | X | X | 5 6 |
| | eE | X | X | 7 8 |
| | e | 9 | 10 | 11X |
| | o | 12 | 13 | 1415 |
| | u | 18 | 19 | X X |
| | i | 16 | 17 | X X |

Table 2.3 presents the statistics of n-phones available in this word set. It may be seen that quite a large number of tri-phones are there in the word set. Phonetic transcription of the word set is analyzed to derive these kinds of statistics of n-phones.

Table 2.3 Distribution of n-phone in phonetic word set

| Number of phone | Number of unique n-phone | Maximal n-phone | Average ccurrence |
|---|---|---|---|
| Bi-phone | 750 | LA1 | 4.0 |
| Tri-phone | 1797 | PA10C | 1.35 |
| Tetra-phone | 1745 | GULI | 1.06 |
| Penta-phone | 1253 | PROUR10 | 1.02 |

2.2 Prosodic Representative Sentence Set (B in Figure 2.1):

Prosodic patterns depend largely on (a) types of sentences, (b) mode of speaking as well as (c) on the informant. The two modes (namely: (i) text reading and (ii) dialogue) may be of immediate necessity. The sentences are to be selected with the normal use of speech in mind. It is generally believed that every speech uses utterances or sentences to perform some functions, which amounts to creating an impact in the mind of the listener, who, in his/her turn, acts or utters a proper response. Each of these functions carries their own intonation and supra-segmental patterns, not necessarily unique, to strengthen those objectives. On the basis of these characteristics, following thirteen types of sentences need be particularly addressed:

(a) Simple and verb-less
(b) Simple, with verb
(c) Simple, negative
(d) Complex affirmative
(e) Complex negative
(f) Compound affirmative
(g) Compound negative
(h) Passive
(i) Imperative
(j) Exclamatory
(k) Affirmative (i.e.Yes/No)
(l) Interrogative (i.e. Wh)
(m) Created Patter

For each of the above types, texts are selected carefully by an experienced linguist to cover the appropriate prosodic structures of requisite type. The expert in question has therefore to begin from the phrases/ clauses and pursue the modification the pattern that undergoes when placed in a sentence. Hence, all supra-segmental patterns are structure-dependent, and unless one is mindful of the structure of these utterances, these patterns may not be fully covered.

The procedure adopted here is to collect a large number of sentences with varied word and phrase combinations in sentences of various functions and structures from written sources, which use colloquial languages. These sources are popular fictions of contemporary writers, children's books, newspaper reports and articles, etc. A process of random selection is used to select a sentence from a source. More than two thousands sentences have thus been selected initially. Out of these, an experienced Linguist selected 880 sentences covering all the aforesaid categories of sentences. Thus, the basis for selection has largely been the intuition of the Linguist who is also a native speaker of the language. In spite of these rigorousness of collection of sentences from printed texts, there may still be patterns that are yet uncovered within the collection. To further augment the collection, created patterns are included in order to improve the comprehensiveness in this regard.

The above selected sentences represent the prosodic structure mainly in text reading mode. Some created dialogue had also been included to cover some of the prosodic structure in dialogue mode.

2.3 Nonsense words set (C in Figure 2.1):

This set is intended for supporting design of speech synthesis system. This set of words must contain acoustic phonetic characteristics of all phonemes and all combinations of possible bi-phone, so that necessary parametric or raw signal information, for all types of synthesis using di-phone, syllabic, partneme etc. are present. This may enable production of unlimited vocabulary Text-to-Speech Synthesis (TTS) systems. The Non-sense Word is a set of tetra-syllabic nonsense words of the forms CVCVCVCV, CVVCVVCVVCVVC for all possible consonants and vowels combinations. The choice of tetra syllabic words in case of Bangla is necessary because Bangla being a bound stress language with stress occurring normally at the first syllable. Also the last syllable often is either muted or loosely pronounced.

However an additional 8 syllabic form CVNVCVNVCVNVCVNV for two nasals /n/ and /?/ (especially for Bangla, as /n/-/?/ distinction in Bangla is not ascertained at all without in conjunction) with appropriate consonant is also included.

2.4 Spoken Word Lexicon (D in Figure 2.1):

This set is intended for development and benchmarking of large vocabulary word recognition systems. A large collection of most frequently used unique words (excluding the formal classical written words) may qualify for this Lexicon. For Bangla, a set of most frequent 10,000 (containing cluster-units and consonant-clusters) words are selected from the combined collections of the MIT Bengali Corpora of 2,00,000 unique words and around 1,00,000 unique words from published Bangla News Paper Ananda Bazar Patrika. As stated earlier, the formal classical written words are excluded.

2.5 Selected paragraph of different categories (E in Figure 2.1):

Three paragraphs, one each on sports, news and science, containing on an average 25 sentences are selected for this category out of newspaper reporting. These texts also contain commonly used foreign words adapted in the language

2.6 Specially designed Sentence (F in Figure 2.1): This set contains 40 phonetically rich sentences from set B above, for text dependent speaker recognition.

2.7 Digit Corpora (G): Isolated numbers, from zero to ninety-nine, and also denotional words (like hundred, thousand, lac and crores) are put in this set in colloquial form. This set also contains about 400 continuous spoken numbers up to 10 digits. These are recorded both in the digit form (like / æk, shunno shunno, shunno, dui, tin, car/ in Bangla) and in verbal form (like /æk lokkho duso coutrish/ in Bangla). This is aimed to design and testing of speech recognition system applications (like voice dialing and/or voiced data entry systems).

## Speech Recording in Digital format

### 3.1 Recording setup:

Speech Recording for the corpus is done in an acoustically treated speech-studio environment in digital domain using PC add-on sound card and available recording software (like CoolEdit, SpeechAnalyser, etc.) with "Shure" dynamic microphone. The recording hardware is specified in Table 3.1. The recording set up is made as per the following schematic.

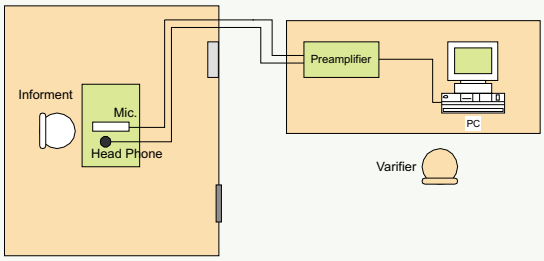


*Figure-1 Schematic outlining Speech-Studio Hardware Setup*

Table-3.1 Studio Specification

| Sl.No. | Description | Specification |
|---|---|---|
| 1 | Distance between speaker and Microphone | 8 inch |
| 2 | Reverberation Time | |
| 3 | Signal to noise ratio | |

Table-3.2 Hardware Specification for Recording

| Sl.No. | Description | Specification |
|---|---|---|
| 1 | Sound Card | Sound blaster live 5.1, Creative Make, Model no SB0080 |
| 2 | Microphone | Shure, Model no. SM63, Frequency response 80 Hz to 20 kHz |
| 3 | 3 Channel Audio Mixer | Frequency response : 30 to 15 kHz Webel Mediatronic make Model no. PAM103 |

3.2 Recording format: 16-bit PCM, Mono, Sampling Frequency 22050Hz.

### 3.3 Informant Selection Procedure:

For recording, informants speaking in Bangla (Standard Colloquial Bengali (SCB)) are selected from different age groups (like 20 to 30, 30 to 40 and 40 above). The selection of the informant is done on the basis of their audition performances on specially designed materials for assessing pronunciations of consonants, vowels, and nasal nominal pairs. For Bangla, the audition material is listed in Annexure.

| Category | Reading material |
|---|---|
| Consonant Pronunciation | Annexure-1 |
| Vowel Pronunciation | Annexure-II |
| Nasal Nominal Pair | Annexure-III |
| Intonation and prosody | Annexure-IV |

After audition test, the informants are to fill up the following Form for their primary metadata. After this an acoustic phonetic expert evaluates all recorded pronunciations to make gradation and comments in the following format

Evaluation Result of the Informant

- Name of the Informant: …….....…. …….....
- Date of birth…….. DD)……(MM)……..(YYYY)…….. Age……
- Sex…….(M/F)
- Informant for the Languages ……..…….
- Whether the specified Language is mother tongue of the informant ……Yes/No
- Whether he/she is original speaker of the selected dialect ……YES/NO

| Category | Reading Material | Total Point | Point obtained |
|---|---|---|---|
| Consonant Pronunciation | Annexure-I | 20 | |
| Vowel Pronunciation | Annexure-II | 10 | |
| Nasal Nominal Pair | Annexure-III | 10 | |
| Normal Reading | Annexure-IV | 10 | |

Remarks in each category of recording

a) Consonant Pronunciation:
b) Vowel Pronunciation:
c) Nasal Nominal Pair:
d) Normal Reading:

- Name of the evaluator…….
- Experience in this field……
- Sex……. M/F
- Evaluation for the Language ……
- Whether the specified Languages is mother tongue of the evaluator ……YES/NO
- Whether he/she is original speaker of the selected dialect ……YES/NO

Final recommendation:
Signature of the Evaluator
Date & Place

3.4 Informant Metadata Format:

If an informant obtains above 90% marks and have a little bit of problem, which can be corrected at the recording time then he is selected for recording and he fills up the final metadata form. An informant number is assigned according to the age group and sex.

[1]. Name of the Informant:……..
[2]. Date of birth……(DD)…… (MM)…… (YYYY) Age…….
[3]. Sex…….(M/F)
[4]. Informant for the Languages……
[5]. 2nd language:……..
[6]. Profession……..
[7]. Whether the specified Language is mother tongue of the informant ……Yes/No
[8]. Whether he/she is original speaker of the selected dialect ……YES/NO
[9]. Educational qualification
[10]. Contact Address
[11].Contact Phone Number

Signature of the informant

Office Use

The evaluation Result:

| Category | Reading Material | Total Point | Point obtained |
|---|---|---|---|
| Consonant Pronunciation | Annexure-I | 20 | |
| Vowel Pronunciation | Annexure-II | 10 | |
| Nasal Nominal Pair | Annexure-III | 10 | |
| Normal Reading | Annexure-IV | 10 | |

Final recommendation by the evaluator:……..

Informant is selected…….(YES/NO) if yes the informant no……

3.5 Recording Procedure:

The above selected contents are prepared as per the categorical order for recording. The contents materials are printed in two sets one is for informant and other is for the verifier. For the phonetic word set each word is embedded in the neutral career sentence (for example, in Bangla /ami ebar word bolchi/). After recording the word is cut out from the sentence and stored as an corpora entry. For spoken word lexicon, words are listed in separate lists containing 1000 words each. For digit corpora, two separate lists are made; one for isolated digits, while the other for continuous digits.

Rests (around ten minutes duration) are usually provided after reading out of each of the lists to enable the readers get acquainted with the succeeding list. This is also followed in the case of recording of the prosodic sentences and dialogues. The lists for the nonsense word-set are prepared one for each vowel in the combinations described in an earlier paragraph in Chapter-2. The informants are particularly cautioned for correct stress-free pronunciation while keeping the intonation flat and to maintain the constant pitch as far as possible. For selected paragraph of different categories and for specially designed sentences, there should be separate lists.

The recoding of speech recognition content is done thrice in different days. The Selected paragraph of different categories for reading is recorded in three different modes; namely: a) normal reading, b) reading in slow dictation mode, and c) with short pause between words.