Corpus of Spoken Language Data and Issues Involved in Development

DESAI Vaibhav D.  
SDJ International College, Surat.  
vaibhav.sdjic@gmail.com

Dr. JOSHI Hiren D.  
Dr. Babasaheb Ambedkar Open University (BAOU)  
Ahmedabad.  
hiren.joshi@baou.edu.in

Abstract

This paper discusses the main issues involved in the creation of corpus of spoken data. Spoken language data is generally bit difficult to work with. The audio data never give a complete view of the communication made by the speaker. The paper discusses the possible remedies to overcome this problem. The paper also discusses main possible stages to generate computer corpora of a spoken audio data.

Keywords: Corpus, spoken language, audio data, written language, stages of spoken data corpus creation, issues in spoken data corpus creation

1. Introduction

The Information of our interest, we intend to find from unstructured data is generally called a document. Group of such documents is called collection or corpus [1]. A corpus can be defined as a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular natural language or language variety [2].

This paper discusses some of the challenges or issues involved in developing a corpus of spoken language data. Any presentation originally made in oral form is considered as 'Spoken language'. Even though spoken language data can include recordings of scripted speech, the paper focus mainly on the use of recordings of naturally recorded spoken language. Since written language data generally has orthographic words, they can easily be stored in electronic text files. There are some challenges associated with written language data corpus like, layout, font size, indentations, accompanying diagrams and so on. With spoken language data one more difficulty is added: orthographic transcription of a speech event constitute only a partial representation of the original speech event. It is extremely difficult to create an accurate prosodic or phonetic transcription

2. Development of Corpora of Spoken Language

Leech, Myers and Thomas has given a framework to develop computer corpora of spoken language.[3] They have described five stages associated with spoken language based corpora creation.
Table 1. Stages in the development of spoken corpora (after Leech, Myers and Thompson)

<table>
<thead>
<tr>
<th>Stage #</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Recording</td>
</tr>
<tr>
<td>2</td>
<td>Transcription</td>
</tr>
<tr>
<td>3</td>
<td>Representation (mark-up)</td>
</tr>
<tr>
<td>4</td>
<td>Coding (or annotation)</td>
</tr>
<tr>
<td>5</td>
<td>Application (Access)</td>
</tr>
</tbody>
</table>

2.1 Recording

The first stage is the recording or data collection stage. One important activity must be finished before the actual data collection activity starts i.e. consent of those who are involved or owning the base data. The consent could be taken either in written or recorded format. The data collector can follow the guide lines provided by The British Association for Applied Linguistics (BAAL)[4]. Technological advances have made multiple options available to us for audio data collection. From analog to digital era, it is now very much easy and accurate to record audio data in digital format. The advantages with digital data are: 1. More clarity with minimum noise, 2. Easy to record, 3. Easy to transfer using any digital storage.

The EAGLES recommendations also propose that digital recording devices be used, as analogue speech recordings tend to degrade, and are not as easy to access when they need to be studied[5]. One may include video recording along with spoken data. The spoken data alone can’t explain the data entirely, as the speaker can use specific gestures or signs along with delivery of speech. Now, video recording being cheaper it is advisable to record the video along with audio data.

2.2 Transcription

Transcription is a process of creating written or printed version of spoken data. The process of transcription does not necessarily start after recording stage. It may be started in parallel with the recording phase, if needed.

Jane A. Edwards has given three principles to be considered while designing a transcription system,[6]

1. Categories should be:
   a. systematically discriminable
   b. exhaustive
   c. systematically contrastive
2. Transcripts should be readable (to the researcher)
3. For computational tractability, mark-up should be
   a. systematic
   b. predictable

These principal mainly focuses on two things: 1. creation of categories and 2. Readability to both, humans and computers. The entire voice data, including spoken words, pauses and overlaps, has to be represented in a proper symbolic way in the transcript. For example, pause data can be represented in symbolic manner using different approaches as below[7]:

---

VNSGU Journal of Science and Technology – V 5(1) - 2016 - 95
<table>
<thead>
<tr>
<th>Du Bois et al 1990</th>
<th>Short pause</th>
<th>Longer pause</th>
<th>Timed pause</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>..</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MacWhinney (1991)</th>
<th>Short pause</th>
<th>Longer pauses</th>
<th>Timed pause</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>##, ###, #long</td>
<td>#1_5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rosta (1990)</th>
<th>Short pause</th>
<th>Long pause</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;,&gt;</td>
<td>&lt;&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Svartvik and Quirk (1980)</th>
<th>Brief pause</th>
<th>Unit pause</th>
<th>Longer pauses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Edwards has also given direction on the way turn of the speakers be resolved. He has given two ways: Vertical arrangement of each speaker’s turn and columnar arrangement. Vertical being the traditional way is popular. The difficulty with the columnar representation is that it could be much complicated and complex to represent when there are more than 2-3 speakers.

### 2.3 Representation (Mark up)

One important aspect of the transcriptions is its machine readability. As the transcript is readable and understandable by humans, it should also be readable and understandable by the computers. Looking at the huge corpora it would rather be impossible to work with them if they are not readable by computers. For making the computer machine readable, the appropriate method for marking up can be used. Two most common mark up languages are available for the same: 1. HTML (Hyper Text Markup Language) 2. XML (eXtensible Markup Language). Both HTML and XML are forms of SGML (Standardized General Markup Language). XML is having one major advantage over HTML: As the name XML suggests, its and extensible language. User can extend the range of elements, attributes and the entities that are permitted in a document. XML now a days come up as a chosen markup language for corpora.

Even if one follows markup language for representation, there could be a confusion regarding representation of data in multiple ways i.e. data could be represented in multiple ways. To overcome such confusions, TEI (Text Encoding Initiative) has come up with some guidelines on symbolic transcription of symbolic language[8]. The tagset specified in the TEI guidelines covers following components:

- utterances
- pauses
- vocalized but non-lexical phenomena such as coughs
- kinesic (non-verbal, non-lexical) phenomena such as gestures
- entirely non-linguistic events occurring during and possibly influencing the course of speech (e.g. sound of truck reversing in road next to lecture hall)
- writing
- shifts or changes in vocal quality (TEI Guidelines 11.1)

TEI document mainly consists of two parts: a header and a body. The header contains background information about the recording/event and the mark up conventions used in the document. The body contains the transcript tagged data.

### 2.4 Coding (Annotation)

An annotation is metadata (data about data) associated with the base data. In the case of audio data metadata is comment, explanation, presentation mark ups.
Compared to written language data, annotation of spoken language data is much more complicated. There are many features of spoken language data that are worth annotating to facilitate corpus-linguistic research, like, phonological and prosodic characteristics, gestural and interactional and other characteristics as well as capturing the temporal quality of time series data and annotation [9].

2.5 Application (Access)

Most common way to make the audio data corpus available to others is the printed versions of the transcripts. But the printed version of transcripts has so many limitations for analysis purpose that it is preferable to publish electronic version of them. The Oxford Text Archive (http://www.ota.ox.ac.uk/) is the repository for electronically-stored corpora and other forms of data collections in the UK, and there are similar centers in other countries. These are simply repositories, and do not provide an analytical interface to the data. Such open access promises to make the analysis of spoken language data easier for a wider audience, at no extra cost other than the Internet connection. The transcripts can also be downloaded in either HTML or SGML format. Applications to analyze these corpora can be developed by the researcher on need basis.

3. Issues/challenges in creating spoken data corpora

Creating corpora for spoken data is relatively complicated to written data. It have some challenges or issues which needs to be taken care of very precisely.

- Transcription of spoken data sometimes do not carry full information. This not only because of technical difficulties but some visual and tactile features are lost or not captured. The words that appear in an orthographic transcription of a speech event constitute only a partial representation of the original speech event. The analyst prefer to collect video data also along with spoken data so that the missing/misleading content of the data could be corrected.
- For collection of spoken data, clarity of voice is absolutely crucial. The analyst has to ensure the capture of voice data in digital format and with zero of minimum noise.
- Some legal aspects might be there for collection of spoken data. It is sometimes required to take written/spoken consent of the speaker before audio/video recording. Not only the consent itself, but this process can affect the speech quality compare to spontaneous speech.
- Once spoken data is collected, it has to be converted to transcripts. The issue here is to choose spelling convention. One of the major published dictionary has to be selected and its spelling conventions has to be followed.
- When the speaker uses the word other than the native language in his speech, the spelling of that word has to be decided and consistency in this regard has to be maintained throughout the transcription process. For example, if speaker delivers speech in English language and uses the Hindi word mitbhashi in between. Now there has to be a consistent decision to spell this word as meetbhasi or mitbhasi has to be taken.
- Decisions will also need to be taken over how to represent non-verbal data, such as contextual information, paralinguistic features, gaps in the transcript, pauses, and overlaps.
- The transcript has to have a feature of machine readability. Since the size of spoken data can be impossible to analyze manually, most of the time it is the requirement that the transcript should be represented in such a way that it is easily readable or understandable by the computers. Mark up languages can be used to resolve this issue.
- To make the data available to other users or researchers, the data has to be suitable for interchange of information across various software platforms. The information exchange
process is supposed to be automated and accurate. Interpretation of the data must remain consistent on both the sides: sender and receiver.

- XML is one the most accepted markup language used to create transcripts of audio data. Since it is extensible by nature, there has to be a proper guidelines which is to be followed by all the users who uses the transcripts of spoken data. The TEI recommendations provide a comprehensive set of conventions for the underlying representation of data.

- The spoken data corpora has to be available to all the valid users eligible to use it. Maintaining printed version of transcripts is not the best solution for this. There are some open source repository for electronically stored corpora like The Oxford Text Archive. Electronic corpora can be easily accessible and easy to analyze by the valid users.

4. Conclusion:

Creation of corpus of recordings of spoken language is bit more difficult task than written data corpus creation. A well defined process has to be followed to create an accurate spoken data corpus. There are various challenges/issues involved at various stage of this process. All the challenges are to be handled with utmost care and accuracy.

References:


