Computer Learner Corpora (CLC) for Teaching Tamil in the Border areas of Tamilnadu

Dr.K.UMARAJ
Assistant Professor
Department of Linguistics
Madurai Kamaraj University
Madurai -21
umarajk@gmail.com
www.umarajk.in
9487223316

Introduction:

Corpus (Plural Corpora) means a large collection of written text or transcriptions of recorded speech chosen to characterize a language or verifying hypotheses about a language. Wikipedia defines a Corpus as a large and structured set of texts (now usually electronically stored and processed) used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. Corpus is a valuable resource for developing Dictionaries, Thesaurus, Teaching packages, Text to Speech Synthesizers, Machine Translation tools, etc. There are lots of corpora available for the Tamil Language and each one has its own purpose. For example, Parallel corpora are used for Machine Translation. Speech corpora are used for Automatic speech Recognition, Text to Speech synthesizer and Speech to Speech Translation. The Computer Learner Corpora (CLC) for the Tamil language are the collections of authentic texts produced by the learners of Tamil language which are stored in an electronic format. It can be used to identify typical difficulties of the Tamil learners in the Border area or the intermediate learners and It also provides a basis for the identification of frequently occurring mistakes of the learners who are learning Tamil language in the Border areas of the Tamil Nadu. Raw learner Corpora are not much useful for Teaching and Learning of the Tamil language in Schools in Border areas.. It should be grammatically annotated. There are a lot issues arises while annotating the Learners Corpora. The present paper analyzes the issues in annotation of Learner Corpora for Tamil language.

Uses of Annotated Learner Corpora for Language Teaching

Student –Centering learning

Corpus based teaching is a student's centered learning. The benefit of such student-centered discovery learning is that the students are given access to the facts of authentic language use, which comes from real contexts rather than being constructed for pedagogical purposes, and are challenged to construct generalizations and note patterns of language behavior.

Finding Collocates

By using corpus we can find out the Collocates. Collocates provide information on word meaning and usage. Collocates can tell a lot about a word by the words that it hangs out with". Collocates are grouped by part of speech and then sorted by frequency. A focus of the lexical approach to language pedagogy is teaching collocations (i.e. habitual co-occurrences of lexical items) and the related concept of prefabricated units. There is a consensus that collocational knowledge is important for developing L1/L2 language skills .For example, posits that ‘learning a lexical item entails learning what it occurs with and what grammar it tends to have.’ Cowie (1994: 3168) argues that ‘native-like proficiency of a language depends crucially on knowledge of a stock of prefabricated units.’ Aston (1995) also notes that the use of prefabs can speed language processing in both comprehension and production, thus creating native-like fluency.

Identifying sentence structures

Due to corpus based teaching the students can understand the following things a)Useful phrases and typical collocations they use themselves b)The structure and nature of both written and spoken discourses c)The different structures of the sentences in a language

Concordance tools

Concordance tools can be used for understanding the language use in different linguistic context. Corpora are useful in this respect, not only because collocations can only reliably be measured quantitatively, but also because the KWIC (key word in context) view of corpus data exposes learners to a great deal of authentic data in a structured way.

Language Testing

Another emerging area of language pedagogy which has started to use the corpus-based approach is language testing. Alderson (1996) envisaged the following possible uses of corpora in this area: test construction, compilation and selection, test presentation, response capture, test scoring, and calculation and delivery of results. He concludes that the potential advantages of basing our tests on real language data, of making data-based judgments about candidates' abilities, knowledge and performance are clear enough.

Design of Corpus

Sinclair(2005) gives the following criteria for corpus designing.

The content of a corpus should be selected without regard for the language they contain, but according to their communicative functions

Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen

Only those components of corpora which have been designed to be independently contrastive should be contrasted

Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination

Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications

Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get to this target as possible. This means samples will differ substantially in size

The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken

The corpus builder should retain as target notions representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of corpus and the selection of its components

Any control of the subject matter in a corpus should be imposed by the use of external and not internal criteria

A corpus should aim for homogeneity in its components while maintain adequate coverage and rogue texts should be avoided

A corpus should aim for homogeneity in its components while maintaining adequate coverage and rogue texts should be avoided.

Based on the above principles the learners corpora should be selected.

Annotation of Textbook corpora

The present study used the following tagsets for annotating the Tamil written corpora used in the border areas of Tamilnadu

1. FW Foreign word
2. JJ Adjective
3. NN Noun, singular or mass
4. NNS Noun, plural
5. NNP Proper noun
6. CAS Case markers
7. POS Postposition
8. PRO Pronoun
9. RB Adverb
10. RP Particle
11. SYM Symbol
12. VB Verb, base form
13. VBD Verb, past tense
14. VBG Verb, gerund or present participle
15. VBN Verb, past participle
16. WP Wh-pronoun

17. WP$ Possessive wh-pronoun

18. WRB Wh-adverb

Issues in Annotation corpora

While annotating the words, several places it is difficult to settle on a single correct set of tag. For example, the word ends with the suffix ‘aaka” . The suffix ‘aaka will act as a particle in one place and case marker in certain places. In the same way, it is hard to say whether a word is functioning as an adjective or a noun. Based on the context only we can determine the function of a word.

Issues in frequency of phrases

In real text book corpus, certain phrases will not occur in real corpus, but the book will have explanations for those phrases. For example the phrase “ maa munivar” is explained as “ uriccol thodar” , but this type of phrase is not occur in real situation. In the same way , for finite verb phrase (vinai muRRu thodar) ,the book will have the example “ kanteen sitaiyai” this phrase is found in literary Tamil only that too in only one place.

There is confusion in explaining phrases Vs sentences. The word “ thodar” is used intermingled. In one place it will refer as sentence and in another place, it will refer as phrase. E.g “ eluvaay thodar” Vs “idaiccol thodar”. This type of

Issues in frequency of finite verbs

In real text book corpus, the participle forms are occurring more than the finite verb forms. But the books have exercises more on the finite verb forms.

Conclusion

With the corpus-based approach to language pedagogy, the traditional ‘three P’s’ (Presentation – Practice – Production) approach to teaching may not be entirely suitable. Instead, the more exploratory approach of ‘three I’s’ (Illustration – Interaction – Induction) may be more appropriate, where ‘illustration’ means looking at real data, ‘interaction’ means discussing and sharing opinions and observations, and ‘induction’ means making one’s own rule for a particular feature, which ‘will be refined and honed as more and more data is encountered. In this paper, Only a few of the issues are discussed and still more has to done in this area of research.

References:

Aijmer, K. (2009) *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Sinclair, J. 1991. *Corpus, concordance, collocation: Describing English language*. Oxford University Press, Oxford.