Application of Corpus Linguistics in Language Technology

Dr.K.Umaraj
Assistant Professor
Department of Linguistics
Madurai Kamaraj University
Madurai -21
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. Sinclair (1991) " a collection of naturally occurring language text chosen to characterize a state or a variety of language".
- Sinclair (2004), " a collection of pieces of language text in electronic form, selected according to external criteria to represent as far as possible, a language or language variety as a source of data for linguistic research"

Introduction

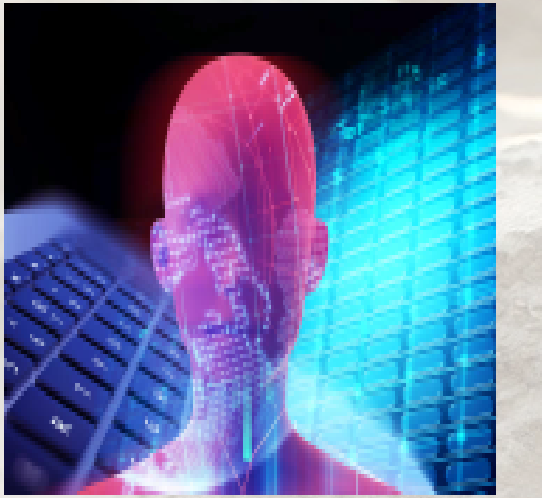
- Finally we can say corpus as an electronically accessing systematic set of texts used for the language research and for teaching and learning purpose.
- Corpus is a valuable resource for developing Dictionaries, Thesaurus, Teaching packages, Text to Speech Synthesizers, Machine Translation tools, etc.
- There are lots of corpora available for the English Language and each one has its own purpose.

Types of Corpora

- Copus
- Parallel Corpus
- Tagged Corpus
- Multi language Corpus
- Speech Corpus

- Parallel corpus are used for Machine Translation, Reading comprehension and Lexical acquisition.
- Speech corpora are used for Automatic speech Recognition, Text to Speech synthesizer and Speech to Speech Translation. and through speech corpus, one can identify the right pronunciation for a particular words, primary and secondary stress, duration of word and intonation patterns. MICASE (Michigan corpus of Academic spoken English).

Speech to Speech



Learners Corpora

- Learner Corpora (CLC) for the English language are the collections of authentic texts produced by the learners of English language which are stored in an electronic format.
- It can be used to identify typical difficulties of the English learners and It also provides a basis for the identification of frequently occurring mistakes of the learners who are learning English language. Raw learner Corpora are not much useful for Teaching and Learning of the English language in Schools. It should be grammatically annotated.
- Lexitutor
- SACODEYL (European Youth Language Pedagogical Focus)
- Linguee (www.linguee.com)
- International corpus of learner English(ICLE)
- Edie is used Lexical study especially the color terms

General Purpose and learners corpus

- British National Corpus
- Brown Corpus
- Penn Tree Bank Corpus
- Childes (Transcription of children)
- Helsinki corpus (written text from earlier period to modern period)
- Colts corpus (London teenage English)

General Purpose Corpus

- Cre-a Corpus
- Collins cobuild dictionary corpus
- Other Dictionary developer have corpus for dictionaries.
  Google , Microsoft and other big companies are collecting corpus for developing Language Technology tools.
- For linguistics analysis, CIIL, TVA, CICT, Amrita, Anna University are collecting data.
- Prof. Deivasundram , Prof. Rajendran, Prof. Ramamoorthy, Prof Vasu Rangathan and Prof. Ganesan and other Linguists are working in the corpus linguistics

Uses of Corpora-Course Planning

- Frequency and register information is very helpful in course planning choices. By conducting an analysis of a corpus which is relevant to the purpose a particular class, the teacher can determine what language items are linked to the target register.
- While frequency is certainly not the only determinant of what to teach and in what order . It can indeed help to make learning more effective. It helps to present the vocabulary, grammar, and functions students encounter most often in real life.

Developing Materials

- **Learners** can also go through the lists word by word in frequency order, finding words that they aren't familiar with. This is a great way to fill in gaps in their vocabulary.
- **Teachers** can assign the students to learn a certain block of words each week and then have a short quiz at the end of the week. At the end of the semester, they'll know that their students are at least familiar with a certain frequency range of words.
- **Student –Centering learning**
- Learners corpora helps to develop student centric learning. In 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.

Collocates

- **Finding Collocates**

By using Learners corpora 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 sentences and concordances

- **Identifying sentence structures**
- By using learner corpora ,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. Learner 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. Tamil Concordance.
  - சங்க இலக்கியத் தொடரடைவுக் கருவி Online U.Ve.Sa. Classical Tamil Corpus
  - சங்க இலக்கிய அகராதி (online Sangam dictionary)
  - சங்க இலக்கிய இணையவழிக் கல்வி (online Sangam Literature via online)
  - சங்க இலக்கிய கற்றல் கற்பித்தல் (Teaching and Learning of Sangam Literature)
  - சொல்லடைவுக் கருவி (Indexing tool)

.

Language Technology tools

- Text to Speech, Speech to text
- Spell checker – sandhi, Grammar Cehcker
- OCR and Tamil typing softwares
- Sentimental analyzer

Identifying sentences and discourses

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.

Grammatical studies of specific linguistic constructions

- If the words in the corpora are tagged grammatical , it will be very useful teaching and learning.
- Grammatical tagging or Part-of-Speech tagging is a process of automatically assigns tag to each word in a corpus.
- POS tagging is a basic preprocessing task for all language processing activities.
- Different approaches have been used to automate the task of POS tagging for English and other languages..

Grammatical studies of specific linguistic constructions

- POS tagging are done by the machine learning techniques, where the linguistical knowledge is automatically extracted from the annotated corpus.
- Tamil being a Dravidian language has a very rich morphological structure, which is agglutinative.
- lexical roots followed by one or more affixes. So tagging a word in a language like Tamil is very complex.

Grammatical studies of specific linguistic constructions

- The main challenges in Tamil POS tagging are solving the complexity and ambiguity of words. That is a word may belong to more than one category. For example, run is both noun and verb. Taggers use probabilistic information to solve this ambiguity. Ideally a typical tagger should be robust, efficient, accurate and reusable.
- இலக்கண வகைக் குறியீடு

POS tagsets for annotation for Textbook corpora

For achieving POS tagging, deciding and creation of tagset is very important. There are several POS tagsets for Indian languages created by number of research groups. Commonly using tagsets are given below

- CC Coordinating conjunction
- CD Cardinal number
- DT Determiner
- EX Existential there
- FW Foreign word
- JJ Adjective
- JJR Adjective, comparative
- JJS Adjective, superlative
- MD Modal

Tags for Textbook corpora

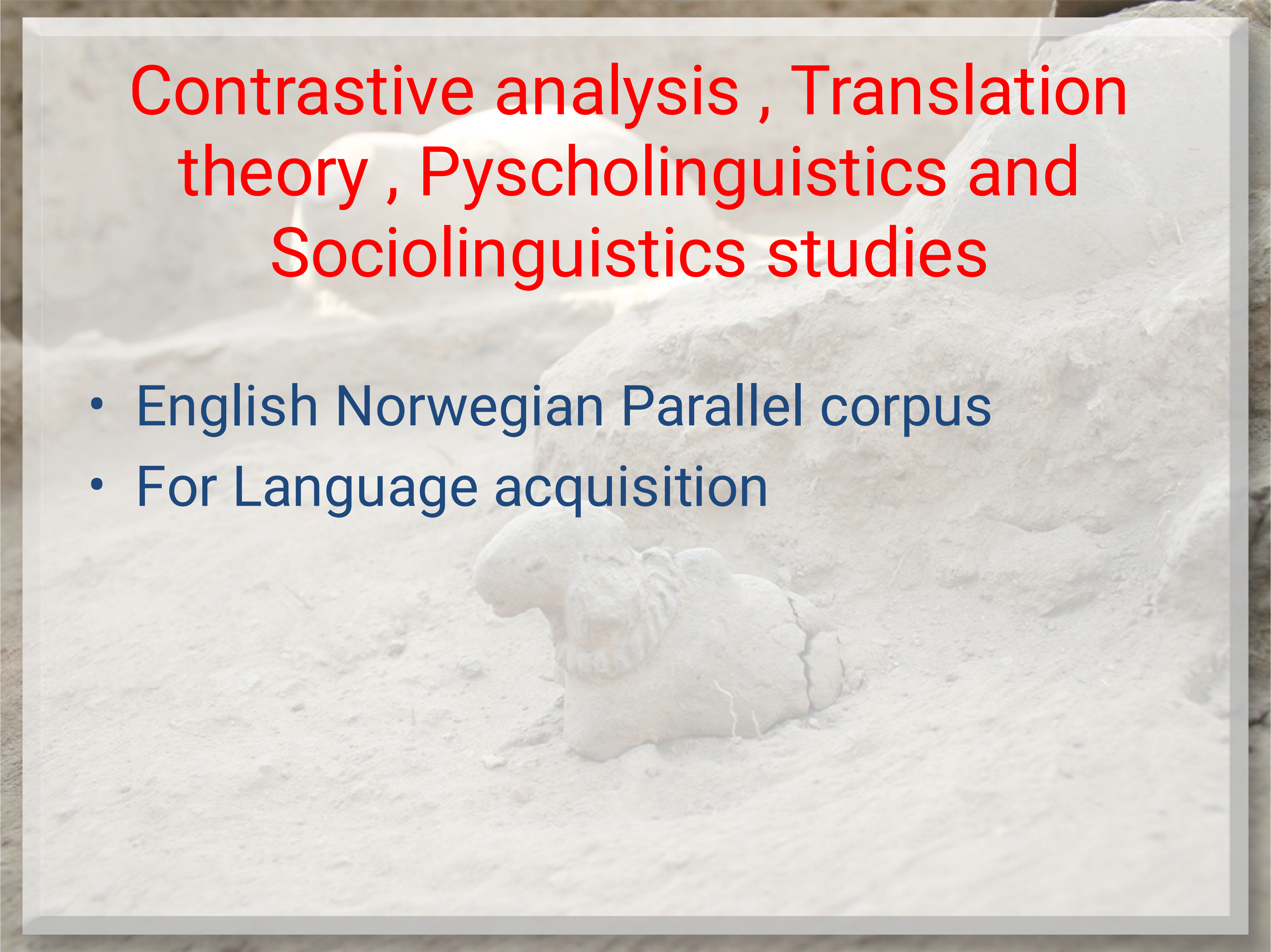
- NN Noun, singular or mass
- NNS Noun, plural
- NNP Proper noun, singular
- NNPS Proper noun, plural
- PDT Predeterminer
- POS Possessive ending
- PRP Personal pronoun
- PRP$ Possessive pronoun
- RB Adverb
- RBR Adverb, comparative
- RBS Adverb, superlative
- RP Particle
- SYM Symbol
- TO to

POS tagging of Textbook corpora

- UH Interjection
- VB Verb, base form
- VBD Verb, past tense
- VBG Verb, gerund or present participle
- VBN Verb, past participle
- VBP Verb, non-3rd person singular present
- VBZ Verb, 3rd person singular present
- WDT Wh-determiner
- WP Wh-pronoun
- WP$ Possessive wh-pronoun
- WRB Wh-adverb

Other uses of corpus

- A grammar of Contemporary English
- Oxford English Grammar
- Longmann Grammar off spoken and written



Contrastive analysis , Translation theory , Pyscholinguistics and Sociolinguistics studies

- English Norwegian Parallel corpus
- For Language acquisition

Major Problems in Corpus development for Tamil

- Lexical Ambiquity , Syntactic and Semantic ambiguity
- 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.

Major Problems in Corpus development for Tamil

- **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”.

Major Problems in Corpus development for Tamil

- 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. This will be reflected in the learners corpora.

Major Problems in Corpus development for Tamil

- Idetifying tags to the head word is an issue in Sangam Literature. The Tamil traditional grammarians classified the words into four types. Asher (1982:101,102) classified words into 6 types.

- Lehmann (1989) classified words into 8 types and .Kothandaraman.R(1989) classified the words into 10 types.

Major Problems in Corpus development for Tamil

- Due to, different approaches in classification of words by grammarians, each dictionary follows their own way of assigning the grammatical information to a particular word. The lexical entry அஃதான்று *’aktaanru’* is marked as adjective in Tamil Lexicon and it is marked as verb in Maree’s Dictionary. Similarly the word அக்கிய *’akiya’* is marked as verb in Maree’s dictionary. But it is an adjective

Major Problems in Corpus

- In Virtual university annotated corpus the word '*naLi*' was tagged as verb and In Prof. Agesthialingam *pathirrupattu* index , it is marked as Relative participle. In following sentences, the words மிகு, உமிழ், உயர், அணி, புகழ் etc are marked as verb by virtual University. But it is not so based on the context.

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 Lecture Only a few of the issues are discussed and still more has to done in this area of research.

References

- **References:**
- Aijmer, K. (2009) *Corpora and Language Teaching*. Amsterdam: John Benjamins.
- Biber, D., Johansson S., Leech G., Conrad S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. London: Longman.
- Biber, D., Leech, G. and Conrad, S. (2002) *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.
- Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. Harlow: Longman.
- Sinclair, J. 1991. *Corpus, concordance, collocation: Describing English language*. Oxford University Press, Oxford.