Issues while developing a Corpus-based Tamil-English Bilingual Electronic Dictionary for Modern Tamil

Dr.K.Umaraj
Assistant Professor, Department of Linguistics
Madurai Kamaraj University, Madurai
umarajk@gmail.com, www.umarajk.in , + 91 9487223316

ABSTRACT

Technology has developed a lot in India. With the advancement in Technology, corpus based electronic dictionary is emerging as a standard format of dictionary instead of the traditionally valued paper dictionary. Corpus based electronic dictionary as well as other corpus based electronic dictionary resources like e-Wordnet , e-Thesaurus, e-Lexipedia , e-Glossary and e-Lexicon are of immense value. They form a basis of various aspects of Natural Language Processing (NLP).

***Keywords*:** corpus, bilingual dictionary, compilation of corpus based dictionary

1. INTRODUCTION

A bilingual dictionary, as contrasted to a monolingual dictionary, deals with two languages. The lexical units of one language are defined or explained in another language. The basic purpose of a bilingual dictionary is to coordinate with the lexical units of one language with lexical units of another language which are equivalent in their lexical meaning

2. PREVIOUS WORKS

Several websites are available for Tamil dictionaries which can be utilized for teaching learning purpose. For example www.sol.com.sg/classroom/dictioanry/html is a website for accessing the Tamil lexicon. Another website for an online web based English Tamil dictionary is www.geocities.com/Athens/Acropolis/8780. The University of Chicago has developed an online platform for accessing Tamil lexicon http://dsal.uchicago.edu/dictionaries/tamil-lex/. A website to access the Cre-A : dictionary of Contemporary Tamil (on line version) is available in the following web page www.lib.Uchicago.edu/Libinfo/Subjects/SouthAsia/.

The National University of Singapore has developed a Tamil English dictionary which is available online in the website www.irdu.nus.edu.sg/tamilweb/. Another website for Tamil-English, English-Tamil and Tamil-Tamil dictionary www.murasu.com/akaram will be useful for learning Tamil language through online. The Institute of Indology and Tamil studies, Colone, Germany has produced Sanskirt, Tamil, Pahlavi dictionary. This dictionary has 1, 66,434 entries. It was developed by Prof. Malton from University of Colone. "Core vocabulary for Tamil '' is an online dictionary published by the department of South Asia Regional Studies, University of Pennsyvania., Mysore. PAL organization published English –English -Tamil electronic dictionary. It has 22,000 heads words and 35,000 sub words, Tamil Lexicon and Muthu Shanmugam pillai's Tamil- Tamil e_dictionary was available in the Tamil Virtual University website. Winslow and Lifco companies published online dictionaries.. An electronic dictionary for Scientific Technical Terms in Tamil has also been developed by Chellapan Radha. It has different kinds of retrieval and browsing facility. CoRpuaiyal is an online dictionary

contains 20,000 root words. Each entry in the dictionary includes the Tamil root word, its English Equivalent, different meaning of the word, and the associated syntactic category.

3. PAPERS PUBLISHED IN INFITT CONFERENCES AND AROUND THE WORLD

The following papers were published in the INFITT Conference proceeding on Bilingual dictionaries 1. "Compilation of Electronic Dictionary for Tamil" by Prof.M.Ganesan 2. "Role of Electronic Dictionaries in Tamil language Teaching and learning" by Prof.S.Raja. 3) "Tamil Interface to the online Tamil dictionary at Cologne" by Prof. Anbumani Subramanian 4) "The English dictionary of the Tamil verb" by Prof. Harold F.Schiffman and "Moziperyarppuk kalaiyil agarathin payanpaatu" by Ilangkumaran. Apart from the INFITT conference proceedings, the following papers were published around the World 1) "Corpus-based Activities versus Intuition based Compilations by Lexicographers, "The Sepedi Lemma-Sign List as a Case in point" by Gilles –Maurice de Schryver 2) "A corpus based survey of four electronic Swahili-English Bilingual dictionaries".

4. DRAWBACKS OF PREVIOUS WORKS AND NEED OF THE STUDY

The above said previous electronic dictionaries have the following drawbacks.

1. Most of the above said dictionaries are developed for the purpose of Teaching and Learning of the Tamil language.
2. We cannot create the list of Antonyms, list of Synonyms, list of Infinite forms and list of Finite verb forms and their frequency from those dictionaries
3. We cannot get root word, list of inflected suffixes and list of derivation suffixes from those dictionaries.

So, an exclusive corpus based Tamil-English bilingual electronic dictionary for modern Tamil is the need of the hour and yet no one published it for the developers especially for open source developers freely. So the author tries to develop a dictionary for the developers and while developing such a dictionary, a number of linguistic problems arise. Those problems are discussed in detail in this paper.

5. LINGUISTICS ISSUES IN PREPARING CORPUS BASED TAMIL-ENGLISH BILINGUAL DICTIONARY

5.1 Issues 1: Identifying the Head word

Identification of head word from each surface form of the word or of a cluster of words is a major issue. This is particularly challenging in a inflected languages like Tamil. In heavily inflected language like Tamil the examples must be identified on the basis of the stem form and not on the basis of the form that it has in text. For example: There are certain words in Tamil 'mun' 'pin' etc will have more than one surface form like munnaal, munpu, munnee, pinnaal, pinup, pinee etc. But each surface forms have different meaning in English based on the context.

5.2 Issues2: Identifying a compound word

Identifying compounds is a problem. For example for following word in Tamil 'vilaimagal' 'prostitute' shouldn't be split into two units. If we split the compound word "vilaimagal' then we will get different meaning for each head. So it is very important that the meaning and use of such construction should be included in the dictionary. In the same way, identifying the idiomatic expressions may be a problem. In that it is not possible to derive the

meaning of the expression on the basis of what we know of the basic meaning of the individual words that constitute the expression.

5.3 Issues 3: Identifying verb entries

Verb as a head word means what type of entries, we should have taken is a problem. The entire verb root should be taken as a entry or not is a problem. Noun construction and Adjectival expressions are also a problem in dictionary compilation.

5.4 Issues 4: Identifying Homonym and Polysemy

A major problem in dictionary compilation is the accurate representation of polysemy. Polysemy here understood as a case where one lexical unit has more than one semantic interpretation. Homonym is also like polysemy has more than one semantic interpretation. But unlike polysemy, homonyms express unrelated meanings.

5.5 Issues 5: Assigning Grammatical Information

Assigning grammatical information to the head word is another issue in Modern Tamil. The Tamil traditional grammarians classified the words into four types. Prof. Asher (1982:101,102) classified words into 6 types. Lehmann (1989) classified words into 8 types and Prof. R.Kothandaraman(1989) classified the words into 10 types. 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.

5.6 Issues 6: Giving glosses

There is a word 'uL" in Tamil which has much equivalence in English. Inside, into, among, within, etc. So based on the context we have to choose right equivalence.

6. SOLUTIONS FOR ISSUES

6.1 Head words: Proper care should be taken for presentation of head word in the dictionary. The following points should be kept in mind while identifying a head word in a bilingual dictionary.

1. There must be more number of examples for the head word and the examples should be retrieved from the recent corpus. 2. Rare words, Archaic forms, certain type of compounds should get required examples .3.If the head word has multiple meanings like homonym or polysemy , each member of homonym or polysemy should have examples and it is required that the examples are placed in appropriate places immediately after each member of the homonym or polysemy 4.The concordance tool is very useful for us to find out homonyms and polysemy. If the Modern Tamil have a thesaurus similar to Roget's Thesaurus of English which groups words by their similarity of meaning in to 'fields of knowledge', it is very useful for identifying meaning of related words.

6.2 Compounds

In the case of compounds, standards should be established for creating rules for the compound words. We can provide rules for the system, so that the system can better understand what a compound and fixed expressions are and by giving strategies for how to find them in dictionary. For example Transitive phrasal verbs are those verbs which will take object .While Intransitive phrasal verbs are those verbs that never take an object. .Professor Karthikeyan from Tamil University published some rules for the compounds and those rules may be utilized for developing the bilingual dictionary.

6.3 Verb entries

Corpus based linguistic information are relevant for translators as it enhances the quality of the description of the verb entries by providing accurate and new data. The information found in the comparable and translation corpus is worth being extracted and above all thoroughly examined. Therefore, further work implies refining the current analysis, especially regarding the potential equivalents found in the translation corpus.

6.4 Giving glosses

While developing a bilingual dictionary, one has to consider the purpose for which he or she is developing a dictionary whether it is for first langue learning or for NLP application. While giving meaning, based on the context we have to assign right equivalence.

6.5 Assigning POS tags

Regarding POS tags , so far no one standardized the tagsets for Tamil language. Anna University AUKBC center follows one type of tagsets and CIIL Mysore LDCIL project follows another type of tagsets. Based on the applications, the tagsets are varying. Organizations like INFITT to take initiatives for creating different of tagsets for Tamil according to the users and according to the products.

7. CONCLUSION

In this paper I have discussed only a few problems in detail. Still lot problem exists while developing a corpus based electronic dictionary . In future, those problems should be analyzed elaborately.

References

1. Cruse, D.A (2000) Meaning in language, Oxford University Press; Oxford.
2. University of Madras., (1982), Tamil Lexicon, Vol, 1-6 & Supplement, Reprint, Madras.
3. Bilingual Lexicography and Corpus Methods. The Example of German-Basque as Language Pair David Lindemann UPV-EHU University of the Basque Country, Tolosa Hiribidea, Donostia, Spain.
4. Enriching Bilingual Dictionaries with Corpus-Based Data: First Steps Towards an Improved Description of Verbs in General Bilingual Dictionaries Thanks to a Popular-Science Corpus. Amélie Josselin-Leray.