Known Strangers:
Cross Linguistic Patterns in Multilingual Multidirectional Dictionaries

Dr. Rejitha K. S.
Linguistic Data Consortium
Central Institute of Indian Languages
Manasa Gangothri, Mysore.
ksrejitha@gmail.com

Rajesha N.
Linguistic Data Consortium
Central Institute of Indian Languages
Manasa Gangothri, Mysore
n.rajesha@yahoo.co.in

Abstract

The multilingual multidirectional dictionary gives the linguistic equivalent across the languages. In order to build such a dictionary in electronic form poses considerable challenges to the lexicographer and the dictionary architect. One of the major challenges is linking lexical ambiguity across languages. This paper intends to address that issue along with many other challenges involved in creating such a multilingual and multidirectional dictionary.

1 Introduction:

Dictionaries are compiled with a view to provide lexical and semantic information from thousands of years. Electronic/digital dictionary does the same by replacing the format of the traditional printed dictionaries. An electronic dictionary, though primarily designed to provide basic information such as grammatical category, meaning, usage etc. as the paper dictionaries, they can also provide additional information like pronunciation, motion pictures through multimedia which paper dictionaries cannot.

The expression Electronic dictionary gained momentum in the last quarter of the 20th century as a term for a specialized device - a handheld computer dedicated to storing a lexical database and performing lookup in it. Classical lexicography demands a complex relationship with linguistic theory. So is electronic lexicography with computational linguistics. Electronic dictionaries are a product of this association and they also serve as tools and feedstock for creating other products. An electronic bilingual or multilingual dictionary may be a digitized edition of a conventional reference work perhaps augmented by types of information specific of this medium (recorded pronunciations, hyperlinks, full text search etc.). Alternatively, it may be a system of monolingual dictionaries of different languages interlinked at the level of entries. [Ivan A Derhanski 2009]

If the construction of the multilingual electronic dictionary is not just a collection of digitized versions of printed dictionaries but to offer facilities like multidirectional search, extracting mono-lingual, bi-lingual, tri-lingual dictionaries, root lexicons and even provide backend support for translation systems then designing such a dictionary database throws practical challenges. Especially when such database accommodates multiple languages at one go and provides options for multidirectional search. That means word of any language as source can be sought in one or more target languages catered by the dictionary system.

The creation of a multilingual dictionary database concerns itself with the source of information used for constructing them. Most of such endeavors primarily rely on printed dictionaries or machine readable versions of the same. Currently we have the advantage of electronic corpora which has been built for many Indian Languages over the past decade.

Polysemy is seldom a serious problem in human communication. Lexicographers have traditionally been concerned with the best way to account for the fact that one word can carry several different meanings (Leacock C. and Ravin 2000). Over time, lexicographic procedures have been established that have resulted in the listing of multiple dictionary senses for polysemous words where sub-senses are grouped together with their respective definitions (Henri Béjoint 2000).

This paper addresses how the concepts described in a lingua-franca provides a basis for conducting cross-linguistic research there by facilitating the creation of multilingual dictionary capable of overcoming a number of important linguistic problems.

Proc. of ICON-2017, Kolkata, India. December 2017 ©2017 NLPAI, pages 165–171

The lexical under-specifications and lexical ambiguity are among the major problems. Sometimes one leads to the other. Lexical ambiguity is one of the issues that a lexicographer and the dictionary architect have to face. This paper describes the observations that a lexicographer encounters while handling prototype of 'concept-set-model' architecture.

2 Review of literature:

When we took up the task of building multilingual, multidirectional dictionary for Indian languages, we researched few previous initiatives. The Universidad Politécnica de Madrid's School of Computing have developed a system for building multilingual dictionaries based on multiple term equivalences known as universal words. The system is based on Princeton University's WordNet database. WordNet is a lexical database developed by linguists at Princeton's Cognitive Science Laboratory. The database was designed to inventory, classify and relate the semantic and lexical content of the English language. The system's other mainstay are universal words. The concept of universal word came out of the UNL (Universal Networking Language) Project. The aim of this project is to eliminate the barriers of linguistic diversity by creating a medium of information exchange through which users can communicate in their own language. Similar attempts were done in PanLex Project that aims to help one to express any lexical concept in any language. The endeavor like BabelNet, which is developed with lexicographic and encyclopedic coverage of terms, is a semantic network which connects concepts and named entities in a very large network of semantic relations, called Babel synsets. Each Babel synset represents a given meaning and contains all the synonyms which express that meaning in a range of different languages.

3 Our Approach:

Since our objective is narrow and is to make a corpus based dictionary of Indian languages and not of a semantic net. The paper is about dictionary only. To make a multilingual multidirectional digital dictionary the approach of word to word linking across languages is not practically feasible. Some concepts may never have a word for it because the concept itself could be alien to the language culture. For example, there cannot be an equivalent word for Kannada '*mud̪d̪e*'. *mud̪d̪e* is a kind of edible ball prepared by cooking millet powder used majorly in southern part of Karnataka. So does for 'tulip' flower in Telugu. Since the tulip flower is not native to the culture of the Telugu speaking land. In traditional dictionaries, such cases are dealt with by describing source language word in target language.

Word from language 'A' may have more than one meaning which gets connected to a word in language 'B' which may not share all the meanings of the language 'A'. Sometimes it may have other meanings too which language 'A' word may not have.

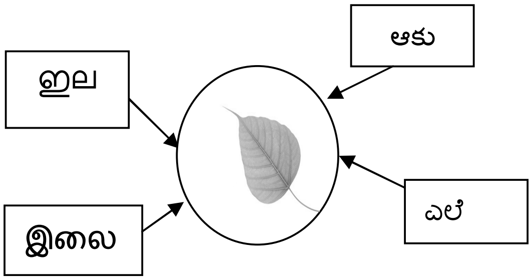
Fixing a language as source and other languages as target may bring only the concepts of the source language culture and omits all possible concepts that other languages may have. A dictionary database based on such limited concepts offers limited descriptions to the end-user, primarily if the end user is searching between two languages which are only target languages in the database architecture. Making a universal word-set is a good start but it will eventually lack the language specific or region specific concepts in the multilingual multidirectional dictionary.

Words borrowed from same origin like Proto-Dravidian or Sanskrit to two different languages may not carry the same concept with them. So it is evident that word to word linking across languages is not a feasible solution even at the stage of polysemy or borrowed words like *tatsamas*. Thus we have to lean back to the basic principles of linguistics where it is the concept that exists as the fact and we label it differently in different languages.

As the Vedic hymn say "*Ékam sáth* víprā bahudhā́ *vadantī*". (The fact exists and the learned one call it by different names -*Rigveda*) The world existed before any language came into existence. When languages evolved with its vocabulary its primary job was to label the things and actions. Those words later fell into different grammatical categories like noun, verb, adjective, adverb etc.

According to Ferdinand de Saussure the signified is the concept, the meaning, the thing indicated by the signifier (Language). It need not be a 'real object' but is some referent to which the signifier refers. The language is built around the concepts that exist in environment.

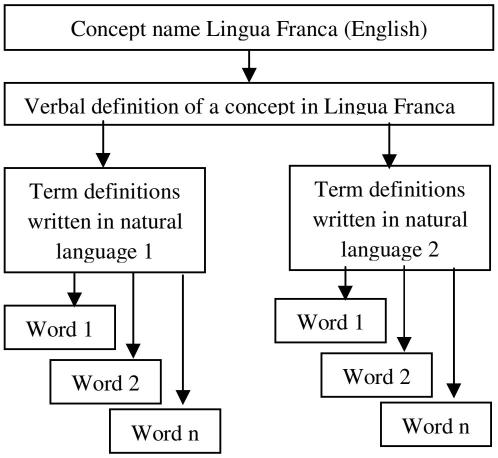
Let us consider the concept 'leaf" and its description as 'The main organ of photosynthesis and transpiration in higher plants'[1]. This concept-set idea of leaf in four Dravidian languages Malayalam, Tamil, Telugu, and Kannada is as follows.



In Terminology, terms i.e. the “verbal definition of a concept” need to be separated from concept names since they belong to two different semiotic systems. The first is a linguistic system while the second is conceptual. Similarly, term definitions written in natural language need to be separated from concept definitions written in a formal language. The former are viewed as linguistic explanations while the latter are considered logical specifications of concept. The result is a new kind of terminology called onto terminology (Christophe Roche, Marie Calberg-Challot, Luc Damas, Philippe Rouard 2009)

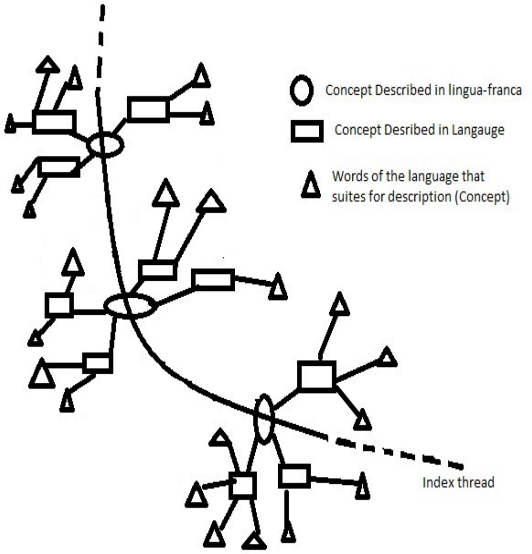
On the similar lines of onto terminology we build our concept set which is a basic data unit for a lexical entry. A concept set is a set which has a concept described in Lingua-franca along with its associated sense in connected languages which in turn connected to related words in Indian languages catered by the dictionary.

Our typical concept set looks as follows.


Concept name Lingua Franca (English)
Verbal definition of a concept in Lingua Franca
Term definitions written in natural language 1
Term definitions written in natural language 2
Word 1
Word 2
Word n
Word 1
Word 2
Word n

The 'concept-set-model' i.e. a Lexical item is entered along with its synonyms and semantic meaning linked with ‘concept’ (descriptive meaning in lingua-franca) into the database. We have chosen English as lingua-franca with its probable word if exists in English. Based on the concept the process is iterated in other languages. In other words, we are following indexation of ‘concept’. Here word is terminal or leaf end of the linkage and not like a node of a semantic network model.

In Central Institute of Indian Languages (CIIL) the concept-set model based dictionary architecture was built in 2010 (Rajesha N, Ramya M and Samar Sinha 2011). We have the advantage of electronic corpora which was built in house for many Indian Languages over the past decade by CIIL. We thought of using the same to enrich our dictionary named ‘*vāgartha*’ (word and sense) that we are building in-house.


Concept Described in lingua-franca
Concept Desribed in Langauge
Words of the language that suites for description (Concept)
Index thread

Since we are following the concept which is the fact that exists, rather than any words to connect with, the challenge of choosing a fixed primary language is also eliminated. For the purpose of management and to avoid confusions, the method of entering the new concepts into the dictionary restricts to one language at a given point of time. Such language will be called as Primary Language. All other languages will add the entries and other respective fields in their language in correspondence with the concepts given by the Primary Language. After a fixed period a different language will become the Primary language so that the dictionary should not miss any concepts which could be a cultural specific item of a language community/region. We devised a system where the concept is fixed and the words act as labels attached to the concepts. The concept and the primary language word associated with it, is shown to the connecting lexicographer. The system gives a facility to transliterate the primary language word if it would be of any help to the lexicographer.

4 Observations:

Looking beyond the well-known issues surrounding the treatment of polysemy in a single language we find even greater problems when it comes to accounting for polysemy across languages. Overcoming these problems is not only important for the design of traditional lexicons but also crucial for the successful implementation Multilingual Lexical Databases. (Hans Christian Boas 2009)

Polysemy can pose problems in intra-lingual and inter-lingual linkages.

4.1 Lexical Ambiguity in a language

In Intra-lingual linking the Lexical Ambiguity words that are not even remotely connected in conceptual sense bring ambiguity to the user. For example the Malayalam word ‘ʋaɽʃam’ has two senses as following.

1. ‘Year- A period of time containing 365 (or 366) days’[1]
2. ‘Rain- Water falling in drops from vapour condensed in the atmosphere’[1].

The Dictionary database architect has to arrange the data without any redundancy in relational database. So the single lexical entry of the word has to be connected with two or more senses here. None of them is a sub-sense of the other.

An end user search of database for Malayalam word ‘*aːɳʈə*’ should fetch the description as well as the synonyms of ‘*aːɳʈə*’; in such a case it will obviously fetch ‘*vaɽʃam*’. But because the ‘*vaɽʃam*’ is connected with other words (like ‘*maɻa*’) in the sense of ‘Rain’, database should not render ‘*maɻa*’ for ‘*aːɳʈə*’.

Tautologous: The organization of data should follow the guideline. i.e., words should be interlinked with all other synonyms and the concept to which it is related. While writing dictionary definitions many lexicographers follow precise guidelines on how to define a word.

In spite of this we find definitions like

Luncher — ‘Someone who is eating lunch’[1]

Magnetism — ‘The branch of science that studies magnetism’[1]

These definitions are logically sound and literally true but they are also tautologous. They use the same words or roots in the definition as are found in the headword. The lexicographer has to understand that the architecture of the database will be such that the definitions are not only for the headword but to all the synonyms to which the sense is connected. All these synonyms are also headword candidates and part of lexicon of that language. So none of those words should be used in definition which leading to tautologous entries.

4.2 Lexical Ambiguity across language

In practical scenario we observed four different types of cross linguistic patterns and two potentially confusing patterns. The following table gives a description of these observations in the multilingual database.

Patterns	Description
A = A	Complete overlapping of word senses
A ≠ A	No overlapping of word senses even if words belongs to the same origin or word conceptualization
A1 = A1 A2 ≠ A2	Semi overlapping of word senses. The word may be having more than one sense in a language-duo of which one is common across language but the other senses may not.

A1 = A1
A2 = Null
	Lexical under specification leading to lexical ambiguity. The word has a meaning in one language similar to the other. In addition to that the same word has a specialized sense in the prior which is absent in the later.

A ≠ A ↙ B ≠ B	Semi cross lexical ambiguity is an extension of no overlapping pattern where a pair of words exists in a language-duo and one of the word in the pair connect with the one which are not their replica
A ≠ A ↙↘ B ≠ B	Full Cross lexical ambiguity is an extension of no overlapping pattern where a pair of words exist in a language-duo but both of the words connect with the ones which are not their replicas

Table: 1 Cross Linguistic Patterns

4.2.1 Complete overlapping:

The complete overlap of word senses; we find "Overlapping polysemy" which refers to cases in which items in two languages have exactly the same meanings. In Indian language scenario, normally some words have same origin like proto-Dravidian or Sanskrit borrowed into different languages.

Let us consider an example of overlapping polysemy among Malayalam '*aṭi*' and Tamil '*aṭi*'. The word carries four senses as follows:

1. To Beat (Verb)
2. The part of the leg of a human being below the ankle joint (Noun)
3. The lower part of anything (Noun)
4. A linear unit of length equal to 12 inches or a third of a yard (Quantifier)[1]

We can observe the varying degrees of polysemy exhibited by them and come to the conclusion that the four senses exhibit "Almost complete" overlapping polysemy patterns. Overlapping polysemy poses no problems for multilingual dictionaries.

4.2.2 No Overlapping:

In contrast to the above we observe common phenomena that the word borrowed from the same source into two different languages may have diverging structure. For example '*la:ɲʧana*' in Kannada and Malayalam exhibit semantic overlap when it comes to the basic sense 'indication of something, highlighting, marking something'. However they differ widely in their meaning extensions when it comes to more narrowed senses over time. In Kannada '*la:ɲʧana*' widely used to describe 'Emblem - A visible symbol representing an abstract idea'[1]. This concept is not carried in Malayalam. But it is carried as 'Indication - Something that serves to indicate or suggest' The Kannada '*la:ɲʧana'* cannot be equated with Malayalam '*la:ɲʧana*' anymore. No overlapping poses an issue to the lexicographer, so that simply looking into the word and not the sense will not help while connecting words.

4.2.3 Semi Overlapping:

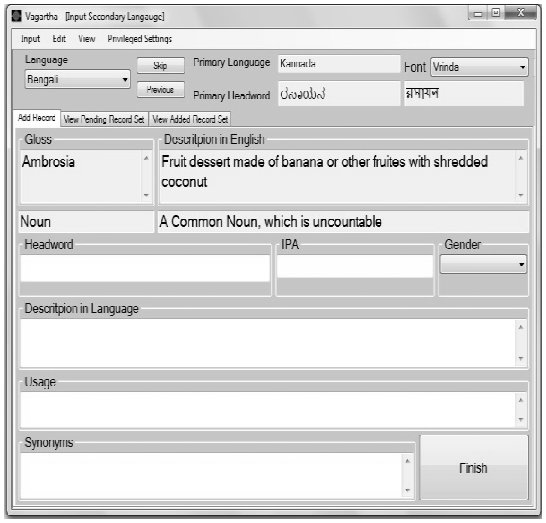
We came across situations in which a word may be having more than one sense in a language-duo of which one is common across language but the other senses may not. For example both Malayalam and Tamil have the word '*kaṯṯi*' and it is used in two senses in both languages. Only one sense is a shared meaning and the other sense is not mutually related. Both words share the meaning 'Knife - A weapon with a handle and blade with a sharp point'[1].

Malayalam '*kaṯṯi*' has a sense 'Burnt - Destroyed or badly damaged by fire' where as Tamil '*kaṯṯi*' has a sense 'Loudly - With relatively high volume'.[1] Simply because these two words are sounds similar and connected in one sense, database architecture should not allow the sense of 'Burnt' to get linked with 'Loudly'. Concept-set-model will take care of it since the words are connected to concepts rather than words. But lexicographer has to be cautious not to jump into conclusions by just looking at the transliterated word that is offered to assist.

4.2.4 Lexical under specification leading to lexical ambiguity:

The fourth type of cross-linguistic phenomenon posing problem for the lexicographer is, cases in which there are no clear equivalents in the target languages. The word has a meaning in one language similar to the other. In addition to that the same word has a specialized sense in the prior which is absent in the later, these cases may lead to zero translations. When the word is outside the culture of the target language and has to be linked, usually lexicographer chooses to borrow the word from source by transliterating the word, (like English word 'tulip' is '*tulip*' in Kannada, as it describes a particular flower.) But in this case lexicographer cannot borrow the word as foreign word for the sake of dictionary entry since it leads to polysemy.

Let's have a look on such a case



Language: Rengali

Primary Longuoge: Kannada

Font: Vrinda

Primary Headword: ರಸಾಯನ রসাযন

Gloss: Ambrosia

Descritpion in English: Fruit dessert made of banana or other fruites with shredded coconut

Noun: A Common Noun, which is uncountable

Headword

IPA

Gender

Descritpion in Language

Usage

Synonyms

Interface created for linking words to concept ‘*rasaːjana*’ of Kannada and ‘*rasaːjana*’ of Bangla exhibit semantic overlap when it comes to the basic senses describing mixture of two or more elements. It is mainly used for the sense ‘Chemical - Material produced by or used in a reaction involving changes in atoms or molecules’[1] in both languages. However they differ widely in their meaning extensions when it comes to more specialized sense. The Kannada ‘*rasaːjana*’ is used to describe ‘Ambrosia- Fruit dessert made of bananas and other fruits with shredded coconut’[1]. This concept is not carried in Bangla. To give equivalent, the Lexicographer cannot borrow it easily since it leads to creating confusion because it is not familiar with the language culture. In such cases lexicographer can just describe the concept in Bangla to convey the meaning to the user. Creating or borrowing a word leads to other complications like social acceptance of something which is not at all part of culture.

In spite of its complexity to find proper equivalents for difficult lexical items across, linguistically it is necessary to account for them within the Database. Without their inclusion, neither humans nor machine will be able to successfully use the database for translation purposes.

4.2.5 Semi Co-lexical pattern

Even though a concept is not a lexical ambiguity we observed a potentially confusing pattern for a lexicographer. This is an extension of no overlapping pattern where a pair of words exist in a language-duo and one of the word in the pair connect with the one which are not their replica

For example ‘*upanjaːsa*’and ‘*kaːd̪ambari*’ are part of vocabulary of Kannada and Hindi. Both words have Sanskrit origin.

Kannada ‘*upanjaːsa*’ is ‘Lecture - A speech that is open to the public’[1].

Hindi ‘*upanjaːsa*’ is, ‘Novel - an extended fictional work in prose; usually in the form of a story’[1].

In Kannada ‘*kaːd̪ambari*’ is ‘Novel’ and in Hindi ‘*kaːd̪ambari*’ is ‘Cluster-of-Clouds’[2]. Both words are present in both languages. But one of the words is having the meaning of the other but the other words are nowhere associated. Lexicographer should not take these words lightly and connect as per their understanding of the word in their language.

upanjaːsa Kannada	≠	*upanjaːsa* Hindi	=	*Kadamabari* Kannada	≠	*Kadamabari* Hindi

The Lexicographer has to take care of the context which appears with the word before connecting it into a sense in their language. Mere identifying the word in their own language will not help them anyway.

4.2.6 Full Co-lexical pattern

This is an extension of no overlapping pattern where a pair of words exists in a language-duo, having same origin but both of the words connect with the ones which are not their replicas. It is also a potentially confusing pattern for a lexicographer.

For example, the words ‘*samɕoːd̪ʰana*’ and ‘*anusand̪ʰaːna*’ is present in both Kannada and Hindi. Both words are having Sanskrit origin. Kannada ‘*samɕoːd̪ʰana*’ carries the sense ‘Research - Systematic investigation to establish facts’[1]. In Hindi ‘*anusand̪ʰaːna*’ is the word for the same sense.

One of the senses that Kannada ‘*anusand̪ʰaːna*’ carries is ‘Modification- The act of making something different in order to achieve desired format’[1]. And in Hindi, word ‘*samɕoːd̪ʰana*’ goes with the sense. The word has other senses like ‘examine’, ‘union’ etc in Kannada.

In this case since both words are part of both the languages vocabulary so the lexicographer has to take extra care to look into the context while connecting. Simply looking into the transliteration form offered by the interface to facilitate the lexicographer will not help and may cause wrong connections.

Conclusion:

As per our observations every word is a new word for the lexicographer. A lexicographer has to take appropriate measures not to get mistaken by looking at the source language word. We mentioned our efforts to ensure appropriate management of the multilingual and multidirectional dictionary project. Once developed, such a dictionary provides a vital resource for cross lingual lexicographers and programmers. At present the data building with the approach of concept set modeling is being carried out. Once the substantial data is entered many more complexities and linking issues may be created. Probable solutions for the same are to be researched accordingly.

Reference:

Christophe Roche, Marie Calberg-Challot, Luc Damas, Philippe Rouard. 2009. Ontoterminology: A new paradigm for terminology. International Conference on Knowledge Engineering and Ontology Development. Madeira. Portugal. 321-326.

Hans Christian Boas (Ed.). 2009. Multilingual FrameNets in Computational Lexicography: Methods and Applications. Walter De Gruyter GmbH & Co. Berlin.

Henri Béjoint. 2000. Modern Lexicography. Oxford University Press. Oxford.

Ivan A Derhanski. 2009. Bi-and Multilingual Electronic Dictionaries: Their Design and Application to Low- and Middle-Density Languages. Language engineering for lesser-studied languages. IOS Press.

Leacock C. and Ravin. 2000. Polysemy. Oxford University Press. Oxford.

Susan J. Behrens, Judith A. Parker (Ed). 2010. Language in the Real World. Routledge. Oxon. 67-88.

Rajesha N, Ramya M and Samar Sinha. 2011 Lexipedia: A Multilingual Digital Linguistic Database. Language in India Special Volume: Problems of Parsing in Indian Languages. 52-55. ISSN 1930-2940.

Timothy Baldwin, Jonathan Pool and Susan M. Colowick. 2010. PanLex and LEXTRACT: Translating all Words of all Languages of the World. Coling: Demonstration Volume. 37-40. Beijing.

[1] Wordweb English Dictionary

[2] Lokbharati Brihat Pramanik Hindi Kosh