

CONCEPT SET MODELING APPROACH TO CONCEPTUALISE MULTILINGUAL DIGITAL LINGUISTIC DATABASE

Ahmad Hweishel AL-Farjat
Applied Science Department
AlBalqa Applied University, Jordan, Aqaba
Ahmed_alfarajat@hotmail.com

Ibrahim Mahmoud Ibrahim AlTurani
Applied Science Department
AlBalqa Applied University, Jordan, Aqaba
Traini110@yahoo.com

Marwan Hweishel Al-Farajat
Asycuda Technical Consultant
United Nation Conference on Trade and Development (UNCTAD)
DTL - Asycuda Programme
Marwan_alfa@hotmail.com

Tareq Ahmad Ali Alzayyat
Department Of Management Information System
University Of Petra, Jordan, Amman
tareq_alzayyat@hotmail.com

Abstract

In this paper, we report the work on developing a multilingual digital linguistic database that aims to provide overall information a linguistic item carries in a language, and its cross-linguistic morphemic equivalent in other languages. It is conceptualised as a model of human knowledge of a language, and its description and architecture is an effort towards modeling such linguistic knowledge. From computational and programming aspects it throws an enormous challenge as it has many-to-many relations across languages, scripts, orthography, fields and entries. To accomplish such linkages among languages and between different types and kinds of information related to the linguistic item, an idea of a ‘Concept Set Model’ is discussed

Keywords: Language, Database, Electronic.

INTRODUCTION:

For more than 2000 years, paper dictionaries are compiled with a view to provide specific information that it aims to provide. Hence, there are several types of dictionaries providing specific information depending upon the type of dictionary. Similarly, an electronic dictionary, though primarily designed to provide basic information such as grammatical category, meaning, usage, frequency, etc., has also got its usage in various other ancillary

tasks in the newer domains of language use. Such electronic dictionary, however, has a major shortcoming as it provides specific information considering the scope, usage, and storage for which it is developed [1].

With the gaining in weight of regional and foreign languages in India from the 11th century onwards, a novel type of lexicon came into being: bilingual and multilingual dictionaries. Amara simhāna Amarakośa emba Nāmaṅgānuśāsana (ಅಮರ ಸಿಂಹಾಸನ ಅಮರಕೋಶ ಎಂಬ ನಾಮಕರಣ ಅನುಶಾಸನ) published in 1970, by Prasaraṅga of University of Mysore, is an example of multilingual thesaurus having Sanskrit as source and English and Kannada as target languages. Thus, even in multilingual dictionaries the correspondence between the working languages is mostly established through an intermediate language – an interlingua – very much in the same way as it is done when connecting two languages by means of a couple of bilingual dictionaries.

Electronic Dictionaries: The expression Electronic dictionary started life in the last quarter of the 20th century as a term for specialised device - a handheld computer dedicated to storing a lexical database and performing lookup in it. As classical lexicography is in a complex relationship with linguistic theory, so is electronic lexicography with computational linguistics, of which electronic dictionaries are a product whilst also serving as tools and feedstock for creating other products.

An electronic bilingual or multilingual dictionary may be a digitised 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. [2]

Multilingual Lexical database is a great help if one often needs to translate similar documents into different languages (a reasonably common situation and bound to become more and more frequent in this age of global communication, especially in massively multilingual societies such as India, European Union). Also adding one more language to a multilingual dictionary tends to be less labour-intensive than creating a new bilingual dictionary thus economically more viable for languages with relatively few speakers and learners.

Some of the Advantages of Electronic Lexical Database is enlisted below;

1. Flexibility in Growth: Potential for Infinite growth in Lexical Entries and can be integrated with voluminous corpus.
2. Multi Purpose: Serve as explanatory dictionary, grammatical dictionary, Dictionary of Synonyms, Antonyms, phraseology, etymology, pictorial including Embedded Audio-Video which is not possible in printed dictionary.
3. User Friendly: Easy look up is possible, since easy to use GUI is provided by which the word to be looked up can be typed directly or by just selecting the word in the text by invoking dictionary by keyboard or mouse events (as in WordWeb) from any word editors.
4. Digital grammatical dictionaries can also extract inflections at least partly and work as morphological analysers and generators upon demand.
5. Easy update facility can be provided under the guidance of a moderator

The major issue which comes into picture at the time of developing a multilingual digital Linguistic database is, interlinking the Lexical entries of different languages in database.

The major issue in Linking Lexical entries across language is the complexity of many-to-many relationship of words. One word of a language may have more than one meaning, but the word corresponding to the same word in a different language may not represent all the meanings of a source language term. At the same time Target word may be representing some more meanings other than the source language term. In multilingual scenario, those other meanings may get linked with a lexical item of some other third language of which there may not be any corresponding term available in first language.

To cite an example, a linguistic item in Kannada 'ke-sa-ri' (ಕೆಸರಿ) represents three concepts.

A shade of yellow tinged with orange- saffron.

A flavouring agent - saffron.

A large tawny flesh-eating wild cat of Africa and South Asia- lion.

When the Kannada word 'ke-sa-ri' maps with its Arabic counter part 'za'farān' (زعفران), which provides the first two sense but the third sense 'lion' is not provided by Arabic 'za'farān', instead Arabic word 'al'asad' (الأسد) is used.

There are situations when a term used in one meaning in a language may exist in the second language but with a different meaning. Thus, linking them is not possible.

e.g.: The Linguistic item 'u-pa-nya-sa' (ಉಪನ್ಯಾಸ) in Kannada means 'A speech that is open to the public - Lecture' in Hindi same 'u-pa-nya-sa' (उपन्यास) means, 'An extended fictional work in prose; usually in the form of a story - Novel'. 'Novel' in Kannada means 'Ka-dam-ba-ri' (ಕಾದಂಬರಿ), but same 'Ka-dam-ba-ri' (कादंबार) means 'Cluster-of-Clouds' in Hindi.

A Language might have borrowed a word with a meaning or a few meanings from some Classical Language instead of borrowing all the meanings, and may have a word homographic in its own language with different meaning or lexical category

For e.g. Linguistic item 'hari' (ಹರಿ) in Kannada which was borrowed from Sanskrit as noun means 'The sustainer; a Hindu divinity worshipped as the preserver of worlds - Lord Vishnu', but in Sanskrit it is used in 36 senses including 'Lord Vishnu', few of them are given below.

'The destroyer; one of the three major divinities in the later Hindu pantheon -Shiva' (noun)

'Solid-hoofed herbivorous quadruped domesticated since prehistoric times - Horse', (noun)

'Any of various long-tailed primates (excluding the prosimians) - Monkey', (noun)

'Limbleless scaly elongate reptile; some are venomous - Snake' (noun)

'The process of combustion of inflammable materials producing heat and light and (often) smoke - Fire' (noun)

Kannada borrowed only the meaning of 'Lord Vishnu' so in developing multilingual dictionary it cannot be linked with all the concepts of Sanskrit. More over in Kannada the same lexical item 'hari' is homograph representing different meanings as follows.

'The motion characteristic of fluids – flowing' (verb)

'Move smoothly and sinuously, like a snake – snake' (verb)

'To separate or be separated by force – tear' (verb)

The above meaning of the Kannada lexical entry 'hari' is not shared by the 'hari' of Sanskrit. So these meanings also cannot be linked in database.

By the above given examples it's clear that Word-to-Word mapping is impossible. Multilingual dictionaries usually select one language as the leading one (or vedette). Data in all other working languages are translated into this one and in this way are connected to each other [3]. If exact word is not available in the target language the user should at least get the descriptive meaning of Source Language word.

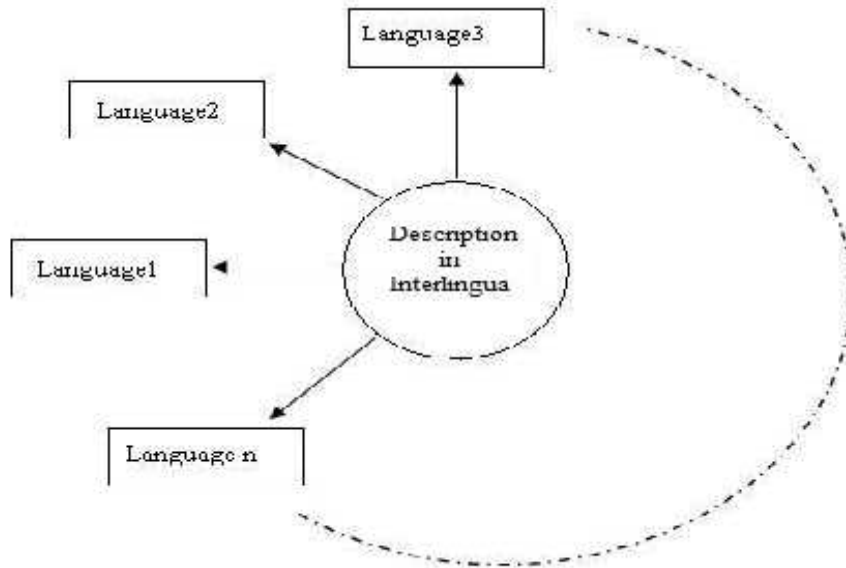
As it is evident that word to word mapping is not possible, and in multilingual scenario no language can be set as Interlingua. There is no natural language exists which can provide all the corresponding words for all the languages used in a multilingual dictionary. But there is no better choice than having an Interlingua for linking multiple languages lexicon entries.

In this approach rather than following the equivalent items across languages, the descriptive meaning of the item in question is followed. In other words, based on equivalent meaning, items are interrelated, and iterated over different languages. Under such approach, however, it is a known fact that lexical underspecification across languages are encountered. To account this issue, an Interlingua language is taken. Even though it may not have corresponding lexical item for a source word, its descriptive meaning will enable the language lexicographer of a different language to give suitable lexical item in one's language.

The Proposed 'concept set model' i.e. a Lexical Item is entered along with its Semantic Meaning and synonyms and spelling variants linked with 'descriptive meaning in Interlingua' to database. Other lexical semantic relations are entered manually. Based on the 'descriptive meaning', the process is iterated in other languages. In other words, we are following indexation of 'descriptive meaning'.

Concept Set can be represented as follows.

```
{
  (Description in Interlingua + Lexical Item in Interlingua + Grammatical Category),
  (Semantic Meaning + Words along with their Spelling Variations sharing Semantic
  Meaning) in Language-1,
  (Semantic Meaning + Words along with their Spelling Variations sharing Semantic
  Meaning) in Language-2,
  -----
  -----
  -----
  (Semantic Meaning + Words along with their Spelling Variations sharing Semantic
  Meaning) in Language-n
}
```



Conceptual Representation of Linking Languages

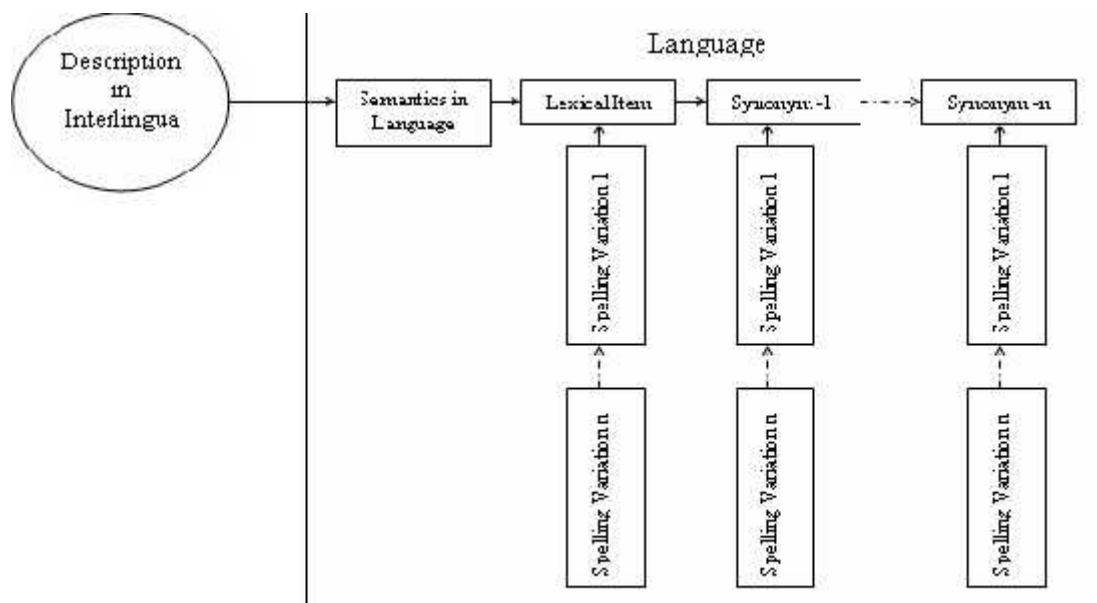
IPA, pronunciation, and transliteration can be embedded in the system. To expedite the data entry, a graphical user interface (GUI) can be provided, which automatically picks 'concept set model's synset along with their spelling variations as an entry. Other fields are provided manually.

As Interlingua is also a Natural Language, Many a times it may give a word which has more than one sense to a linking word. That's why 'description-along-with-lexical-item-and-its-grammatical-category' should be represented in Interlingua with a specific index. The Secondary Languages which are getting linked in Multilingual Dictionary can be able to give their Lexical Items to that Concept-Set.

Many a times the Interlingua may not be able to provide any lexical item for a multilingual language's lexical item. In such cases a detailed description has to be given in Interlingua in concept set.

For e.g. If there exists a word like 'Abhisarika' (ಅಭಿಸಾರಿಕಾ) in Kannada, meaning 'The lady who goes rendezvous with her significant-other', if English used as an Interlingua which may not have a corresponding lexical item for it then the whole Description of 'Abhisarika' in English (interlingua) can be used. Other languages of the multilingual dictionary will give corresponding lexical entry (or description in case no corresponding lexical entry exists).

In this proposed model Interlingua can be any Language, but as it is used only for linking, and has only description, the Interlingua by itself cannot be one among the Multilingual Dictionary Language. The language used for the Interlingua also has to be represented like any other language and has to use the Concept Set as any other Language.



Graphical Representation of Linking of Interlingua-Description with a Lexical Entry of a Language

The major advantage in this approach is; all language of the multilingual dictionaries can enjoy primary language status. A methodology can be devised for a particular period of time, where one language will have primary language status and goes on adding lexical items to Interlingua description, and other languages (Secondary Languages) will give their corresponding lexical items for each Interlingua description added by the primary language. After the time period is over some other language will take over as primary language.

SUMMARY

Multilingual Electronic Dictionaries attempt to provide wide ranging information, and cater the needs of a user to know about a specific linguistic item in a language and its morphemic equivalent across languages. It also provides information at different levels from graphemic to idiomatic expressions and beyond. Its architecture is modular; hence, it can be customised according to the needs of the specific applications/users.

In its conceptualisation and design, specific information of an item is provided at the strata which are called levels that can be customised according to the requirements. Each level provides specific information.

The multilingual digital linguistic database can be enriched with more and more languages, drawing cross-linguistic morphemic similarities and differences between languages. On the other hand, it is conceptualised as a model of what a native speaker of a language knows about an item in his/her language synchronically/diachronically.

REFERENCES:

- [1] Rajesha N, Ramya M, Samar Sinha, 2010, Lexipedia: A Multilingual Digital Linguistic Database, published by "LANGUAGE IN INDIA" Volume-11: 5-May-2011 ISSN 1930-2940 <http://languageinindia.com/may2011/rajesharamyasamar.pdf>
- [2] Ivan A DERHANSKI, 2009, Bi-and Multilingual Electronic Dictionaires: Their Design and Application to Low- and Middle-Density Languages, Language engineering for lesser-studied languages, IOS Press, PP-123
- [3] Igor Boguslavsky, Jesús Cardeñosa, Carolina Gallardo 2009, "A Novel Approach to Creating Disambiguated Multilingual Dictionaries", Applied Linguistics (2009) 30 (1): pp 70-92

- [4] Kavi Narayana Murthy. 2006. Natural Language Processing : An Information Access Perspective . New Delhi: Ess Ess Publications
- [5] Samar Sinha 2011. - (A causerie on churning of speech sounds), NEILS 6 Conference, Tezpur University, Tezpur, Assam 31 Jan - 2 Feb, 2011
- [6] Ammar Merhbi 2009. Corpus Linguistics, Concordance and Data-Driven Learning: An innovative Language Teaching Approach!