Lexipedia: A Multilingual Digital Linguistic Database

Rajesha N
Senior Technical Officer,
ldc-rajesha@ciil.stpmy.soft.net

Ramya M
Senior Technical Officer,
ldc-ramya@ciil.stpmy.soft.net

Samar Sinha
Senior Lecturer / Junior Research Officer
ldc-samar@ciil.stpmy.soft.net

Linguistic Data Consortium for Indian Languages
Central Institute of Indian Languages
Mysore, India
www.ldcil.org

***Abstract*: Lexipedia, a multilingual digital linguistic database aims to provide all types and kinds of information that a linguistic item carries in a language, and its cross-linguistic morphemic equivalent in other languages. It provides a wide range of information from graphemic to idiomatic expressions and beyond. In this paper, Lexipedia is conceptualised as a model of human knowledge of language, and its description and architecture is an effort towards modelling such linguistic knowledge.**

I. Lexical Database: issues and limitations

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, electronic/digital dictionary does the same by replacing the format. 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. In other words, other different kinds of information that the language users require are often not featured but are readily available in another dictionary specifically created for it. In another aspect, such dictionary is a mere list of lexical items with its specific information, and does not reflect how human beings store and process such lexical items.

With the advent of newer domains of language use, however, different kinds of resources are conceptualised and designed to store information which serve as database for different kinds of applications and processes. One such electronic lexical database is WordNet, which organises words into sets of cognitively synonymous sets (often called synsets [1] and [2].) It stores lexical items of a language hierarchically and the conceptual-semantic and lexical semantic relationships between these items are determined cognitively. In other words, it is a hybrid of dictionary and thesaurus providing information of the both. However, the major concern for which Princeton cognitive psychologist George A. Miller developed WordNet is to model a database that is consistent with the knowledge acquired about how human beings process language. In addition to it, WordNet is interpreted and used as ontology. Despite its wider use in several applications like Word Sense Disambiguation (WSD), Information Retrieval (IR), automatic text classification, automatic text summarization, etc., WordNet like other lexical databases too has its own limitations.

These databases are designed with certain specific objectives, hence, to access the detailed information about a particular linguistic item one has to access several different kinds of databases specifically meant to provide the required specific information. For example, to access detailed information about a word 'किताब' in Nepali, one has to access WordNet for conceptual-semantic and lexical-semantic relations, pronunciation dictionary, or even separate databases for usage, idioms, proverbial usage, etc. Similarly, if one has to find its equivalent in other languages, one has to scan bi/multilingual dictionary. As it is known, accessing different databases often lead to inconsistency since each database is constructed to fulfill certain objective. Moreover, such databases are primarily not designed to provide different kinds of information that a Natural Language Process system requires. In other words, it is imperative to build a consistent, uniform, dedicated database which serves NLP applications.

In section 2, the paper explores conceptual design and organisation of different fields, which are modularised with respect to specific information. A principled basis of comparing various linguistic phenomenon across languages and to achieve such an objective to avoid miss-comparison, and in creating typological databases are the subject matter of the following section. Section 4 deals with the computational aspect along with the design of the back-end and algorithms to execute various information. One of the input interfaces is also highlighted in building such database. The final section is a summary.

II. Lexipedia: concept and organisation

In view of the above shortcomings of the lexical databases, Lexipedia is conceptualised to provide all and every kind of information that a particular linguistic item in a particular language embeds, and its cross-linguistic morphemic equivalent in other languages. Here, it is imperative to mention that linguistic item includes free forms as well as bound forms. The latter is the result of grammaticalisation, a historical processes resulting various forms, functions and constructions (see [3] and [4]).

Lexipedia is designed to model how humans organise these linguistic items, and in turn how these items are related with each other as well as with its linguistic usage in various other forms, functions and constructions in a language. In other words, it is designed to reflect all kinds of information that a user of a language carries overtly/covertly over the synchronic/diachronic dimension about a particular item in a language, and its morphemic equivalent across languages. Lexipedia, hence, provides wide ranging information on a linguistic item which is organised in modules.

Since, information that Lexipedia provides is wide and vast, it is organised into different modules, where each module provides specific information regarding an item. Having such a modular architecture for information organisation has an advantage as each module can be customised according to the need of the application/users as well as for resource building. These modules are designed as follows:

A. Graphemic

An item's scriptal graphemic information is provided following the script used for a particular language like Devanagari for Hindi, Nepali, Marathi, Bodo, etc.; Srijanga script for Lepcha, etc. It also provides spelling variations if an item has in a particular language. Along with it, transliteration of the item following the LDCIL transliteration scheme and the (broad) IPA transcription are also provided.

B. Audio-video

Audio-video information about a linguistic item is provided at another module. In this module, pronunciation in audio file, and in cases, image/video files are also supplemented. This module is handy in the study of sub-lexical structure of a language as well as for developing pronunciation dictionary, and other speech related applications.

C. Grammatical

Grammatical information forms the basis of various NLP applications. The grammatical categories are noun, pronouns, verb, adjectives, adverbs, adposition, and particles, which subsumes a larger number of other traditionally defined categories like conjunction, interjection, clitics, etc. In Lexipedia, the grammatical information for each category is provided in hierarchical layers. For example, nouns are organised with respect to the categorising device that language employs (gender, classifier, number, honorificity, etc.). To illustrate such a noun categorisation, Hindi and Assamese employ gender and classifier, respectively. Among the Tibeto-Burman languages, Khasi and Lepcha are other two languages which extensively organises nouns on the basis of classifiers. Similarly, verbs are typologised and organised on the basis of their syntactic behavior into types following [5] To cite an example, Hindi verbs can be typologised following [6] In the case of adjectives, the Cinque Hierarchy (see [7]) can be explored for Indian languages.

In addition to this information about the categories, Lexipedia also provides information on different grammatical categories like tense, aspect, mood, aktionsart, case markers, voice, classifier, gender, person, number, clusivity, etc.

D. Semantic

In this module, multiple semantic information is provided for which Lexipedia employs corpora to ascertain meaning both in its synchronic and diachronic dimensions. Such semantic variation is supplemented by the citation of the actual usage from the corpora.

E. Other

Lexipedia also records proverbial, idiomatic, register, domain specific and various other usages of an entry. Hence, it provides information on various uses of the entry in a language also. At the same time, it also provides information on root, lexical stock and etymology of an entry. Similarly, lexical semantic relations are also presented forming ontology of organisation of items in a particular language.

III. Cross-linguistic typology

One of the major decisions regarding providing cross-linguistic information is about the uniformity of phenomenon in question, and to handle various gradient linguistic phenomena in a principled way. Since Lexipedia provides cross-linguistic information across Indian languages, it is imperative to follow a uniform definition of grammatical category across these languages to arrive at true cross-linguistic information on Indian languages. In pursuit of such cross-linguistic uniformity, it is essential to adopt standards that can be applied uniformly across languages and which allow to compare like with like. Moreover, such standard should also ensure that the cross-linguistic study of the phenomenon is not missed out either due to the different labels or we compare different phenomena due to the same label.

In order to achieve such criteria, Canonical approach, which is put forward to account typology of possible words in the realm of typology, and is widely used in the realm of morphology and syntax is best suited. Canonical approach takes definitions to their logical end point and builds theoretical spaces of possibilities, and creates theoretical spaces, to populate them while the languages are still there to be investigated. Moreover, it is also useful to study both what is frequent and what is rare, and in the construction of typological databases.

IV. At the back-end

Since Lexipedia is a multilingual database, and has many-to-many relations across languages, scripts, orthography, fields and entries, it throws an enormous challenge for computational and programming aspects. To accomplish

such linkages, we have basically adopted a model which is based on concept related to the linguistic item. In this model, concept refers to a description of an item in a link language. For our present purpose, owing to pragmatic factors, we have identified it to be English. To cite an example, a linguistic item in Kannada 'kEsarI' (ಕೇಸರಿ) has three set of concepts.

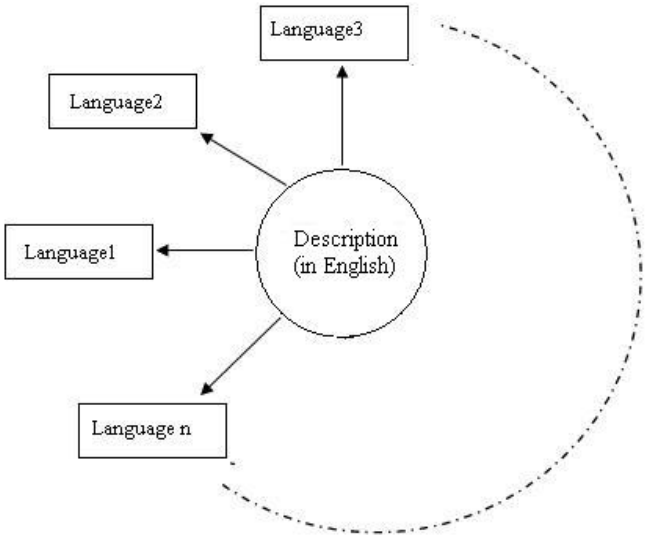
A shade of yellow tinged with orange (SAFFRON).

A flavoring agent (SAFFRON).

A large tawny flesh-eating wild cat of Africa and South Asia (LION).

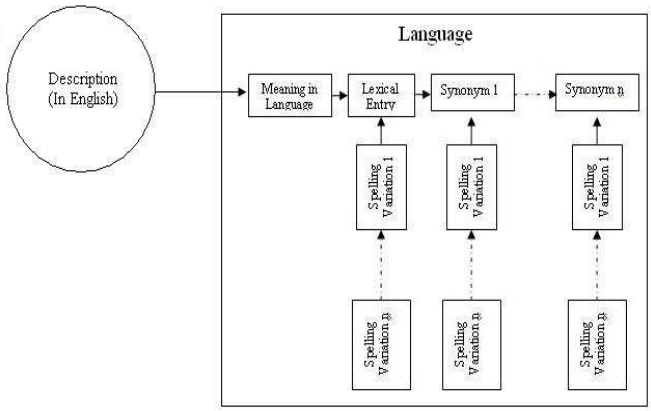
In Lexipedia, rather than following the equivalent items across languages, the descriptive meaning of the item in question is followed. In other words, based on equivalent meaning, items are interrelated, and iterated over different languages. Under such approach, however, it is a known fact that lexical under-specification across languages is encountered. To account such issue, the descriptive meaning of the item in the question will be considered for providing linkages across languages.

Based on the 'descriptive meaning (in English)', the process is iterated in other languages. In other words, we are following indexation of 'descriptive meaning (in English)'.



Conceptual Representationof Linking Languages

In Lexipedia, we have adopted a 'description set model' i.e. based on description (descriptive meaning in English), we provide the entry, meaning (in the language), spelling variation of the entry, and synonyms of the entry. In other words description set consists of description in English, its spelling variations, and synonyms and their respective spelling variations, and meaning in the language where all these items share among each other.



Graphical Representation of Description Set

Other lexical semantic relations are entered manually. IPA, pronunciation, and transliteration (following the LDCIL scheme v0.1) are embedded in the system. To expedite the data entry, we have developed graphical user interface (GUI) which automatically picks 'description set model's' synonyms and spelling variations as an entry and other fields are provided manually.

For the management of Lexipeida, we have devised a methodology that only one language should add fresh concepts (Description in English) at a given point of time. Such language will be called as Primary Language (PL). All other languages will add the entries and other respective fields in their language in correspondence with the concepts given by the PL. We have developed two text data input interfaces for Lexipedia [snapshots are in Annexure I] for both PL and Secondary Language (SL) entry.

V. SUMMARY

Lexipedia attempts to provide wide ranging information, and caters the needs of a user about a specific linguistic item in a language, and its morphemic equivalent across languages. Unlike other lexical databases, it 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, Lexipedia provides specific information of an item at the strata called levels that can be customised according to the requirements. Each level provides specific information.

Lexipedia serves as a linguistic resource hub for Indian languages (at this level of development), however, it can be enriched with other 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. Lexipedia is an effort towards modeling such linguistic knowledge.

ACKNOWLEDGMENT

We would like to thank Dr. B. Mallikarjun, who initially floated the idea of creating multilingual dictionary of Indian languages - a precursor to Lexipedia, and contributed valuable inputs into Lexipedia. We are grateful to Prof. Kavi Narayana Murthy (CIS, UoH, Hyderabad; currently CIIL fellow) for his guidance, help, insightful comments and suggestions on the different issues. We are heartily thankful to our Project Head, Dr. L. Ramamoorthy, whose encouragement, guidance and support enabled us to sum up our efforts so far into words, and other members of the Team LDC-IL for their comments and relevant help.

REFERENCES

[1]. Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM Vol. 38, No. 11*. pp. 39-41.

[2]. Fellbaum, Christiane . 1998.*WordNet: An Electronic Lexical Database.*The MIT Press.

[3]. Lehmann, Christian. 1995. *Thoughts on Grammaticalization*. Munich: Lincom Europa.

[4]. Hopper, Paul J., and Elizabeth Closs Traugott. 1993.*Grammaticalization*. Cambridge, England: Cambridge University Press.

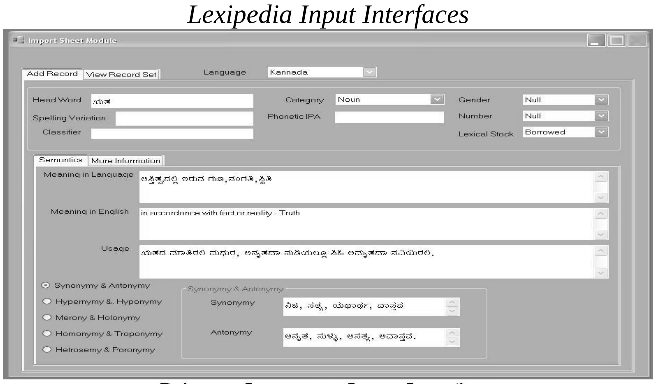
[5] Levin & Rappaport. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Linguistic Inquiry Monograph 26, MIT Press, Cambridge, MA.

[6]. Richa. 2008. Unaccusativity, Unergativity and the Causative Alternation in Hindi: A Minimalist Analysis. Ph.D thesis, Jawaharlal Nehru University, New Delhi.

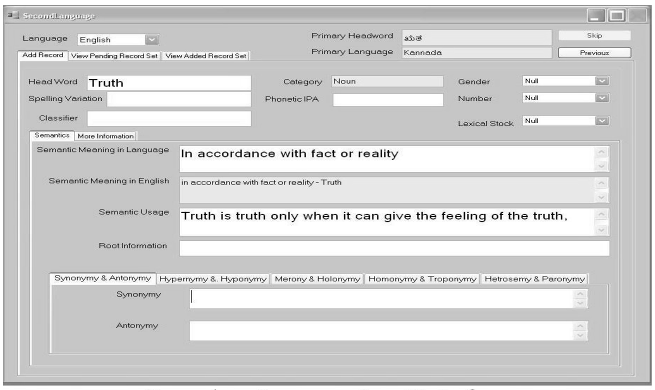
[7]. Cinque, Guglielmo. 1999. Adverbs and Functional Heads. Oxford: OUP

Annexure I

Lexipedia Input Interfaces



Primary Language Input Interface



Secondary Language Input Interface

Output Interfaces developed in First Version.