==
Language in India www.languageinindia.com ISSN 1930-2940 Vol. 23:6 June 2023
==

Development of Verb Frames for Nepali

Krishna Maya Manger, M.A. (Nepali), M.A. (Linguistics)
Assistant Professor
Department of Nepali
University of North Bengal,
Siliguri, West Bengal, India, 734013
krishnamanger@gmail.com

==

Abstract

This paper describes the method and procedures of building Verb Frames for the Nepali language which is developed as a part of doctoral research. A total of 486 Verb Frames have been developed for 200 ambiguous Nepali verbs following the theoretical framework provided by Begum (2017) with some modifications. Our Verb Frame captures lexical, syntactic and semantic information about each sense of a particular verb with due focus on its argument structure and ontology. The significance of Verb Frames is seen in the area of NLP, especially in Parsing and Word Sense Disambiguation. It provides a detailed description of the linguistic attributes of Nepali verbs for scholars who are interested in studying verbs in the language.

Keywords: Nepali, Verbs, Ambiguity, Verb Frames, Argument Structure

1. Introduction

The verb is a core grammatical category of a language that plays a pivotal role in determining the functions of each argument in a sentence. It is a central element in a sentence without which the structure of a sentence seems impossible. That is why the database of verbs which provides linguistic information about a verb has become a major focus in the field of automatic processing of a language at present. There are different approaches for building linguistic resources which could capture all the necessary information about a verb 'ranging from phonological and morphological to syntactic, semantic and pragmatic criteria' (Walde, 2009). In this context, Verb Frame is a kind of knowledge base which can be used as a linguistic resource, particularly in the field of computational linguistics and natural language processing.

The origin of the **concept of verb frame** is presumed to be started from the work of Chomsky (1965) which presented the idea of a sub-categorization frame for the first time. The sub-categorization according to Chomsky denotes 'the ability/necessity for lexical items (usually verbs) to require (allow the presence and types of the syntactic arguments with which they co-occur' (Chomsky, 1965). The notion of sub-categorization looks similar to the notion of valency in the sense that both account for the number and status of arguments in a sentence. But the two are different since sub-categorization in its original meaning did not include the subject.

Though modern theories include the subject in the sub-categorization frame as well, whereas valency was perceived as the number of arguments including the subject from the very beginning. However, this study takes the Verb Frame in a particular sense which has its root in the work of Levin's classification of English verbs (Levin, 1993) which has inspired later development of the concept in many different ways. VerbNet (Schuler, 2005), FrameNet (Baker et al, 1998), PropBank (Palmer et al, 2005), and WordNet (Miller, 1995) are the works which have contributed in the development of the concept, especially for natural language processing. Verb Frame in this study, is a tabular representation that 'captures linguistic information about the syntactic distribution of a verb in a language' (Begum, 2017).

2. Related Work

Levin (1993) presents a large-scale classification of English Verbs which is based on the syntactico-semantic correlation of verbal behavior. It was developed under the Lexicon project of the Centre for Cognitive Science, MIT. Levin assumes that 'the behavior of a verb, particularly for the expression and interpretation of its arguments is to a large extent determined by its meaning' and 'the verb meaning is a key to verb behavior'.

VerbNet (Schuler, 2005) is an online lexicon of English verbs developed under a research project led by Martha Palmer at the University of Colorado. It is a hierarchical domain-independent broad coverage verb lexicon with mappings to other lexical resources such as WordNet, Xtag and FrameNet'. It aims to refine and add subclasses to Levin's verb classes in order 'to achieve syntactic and semantic coherence among members of a class'. Each class of VerbNet contains two types of information about a verb. They are i) syntactic description that depicts thematic roles and selectional restrictions of a verb; and ii) syntactic frames which represent syntactic description of a verb with semantic predicates that shows temporal function.

FrameNet (Baker et al, 1998) is a human and machine-readable lexical database of English developed by the International Computer Science Institute, Berkeley. It is a kind of semantic role labelling that attempts to represent roles in the frame. It contains more than 13,000-word senses with annotated examples. It has more than 200,000 manually annotated sentences linked to more than 1,200 semantic frames which provide a unique training dataset.

PropBank (Palmer et al (2005) is an acronym for 'Proposition Bank' which contains a corpus annotated with verbal propositions and their arguments. It provides resources of sentences annotated with semantic roles which are defined for an individual verb sense where each sense of each verb has a specific set of roles such as Arg0, Arg1, Arg2 and so on.

WordNet (Miller, 1995) is a lexical database of English which groups different words (nouns, adjectives, verbs, adverbs) into sets of cognitive synonyms called synsets. Each synset expresses a distinct concept and is interlinked through conceptual semantic and lexical resources which are navigated through the browser. Synonymy is the main relation among words in the WordNet. However, it also encodes the sense relation of hyperonymy, meronymy, troponymy and antonymy.

Begum (2017) is a doctoral dissertation entitled *Developing a Pilot Hindi Treebank based on Computational Paninian Grammar* submitted at IIIT Hyderabad. In a chapter of her thesis, Rafiya Begum has developed a database of 486 Verb Frames for 300 verbs for Hindi which is one of the pioneering works in the field in India. The model of verb frame presented by her is based on the Paninian Grammatical Framework which has represented the linguistic information (description and dependency relation) about a verb in tabular form. Her work is the main motivation behind this research. The theoretical framework and the methodology devised here followed her work in particular.

3. Methodology

The corpus of 200K sentences was collected using various written sources from personal effort as well as from various NLP projects like the Indian Languages Corpora Initiative conducted in consortium mode led by Jawaharlal Nehru University, New Delhi and Shallow Parser Tool for Indian Languages led by Center for Applied Linguistics and Translation Studies, the University of Hyderabad funded by MeiTy, Government of India. The written corpus from different domains like entertainment, health, sports, news, literature, tourism, and agriculture was collected from both projects. The corpus for the domain of literature is also collected from the website of samakalinsahitya.

The data for this study were extracted from the corpus of the aforementioned sources. The first step of extracting data was to sort out all verb occurrences from 200K sentences. Secondly, unique verb forms looking at their roots were separated. Thirdly, looking at the frequencies of their occurrence, 200 verbs were chosen for the study. A total of 486 unique verb frames have been developed for each sense of 200 ambiguous verbs. The verb senses are verified using the corpus and Nepali dictionaries available in print forms like Poudel (2015), Parajuli (2010) and Lohani & Adhikari (2010). The example sentences for each sense of verbs are extracted from the corpus and Verb Frames are developed following the theoretical model presented by Begum (2017) with some modifications. They contain the dependency relations of the mandatory and desirable arguments with theta role and ontological information.

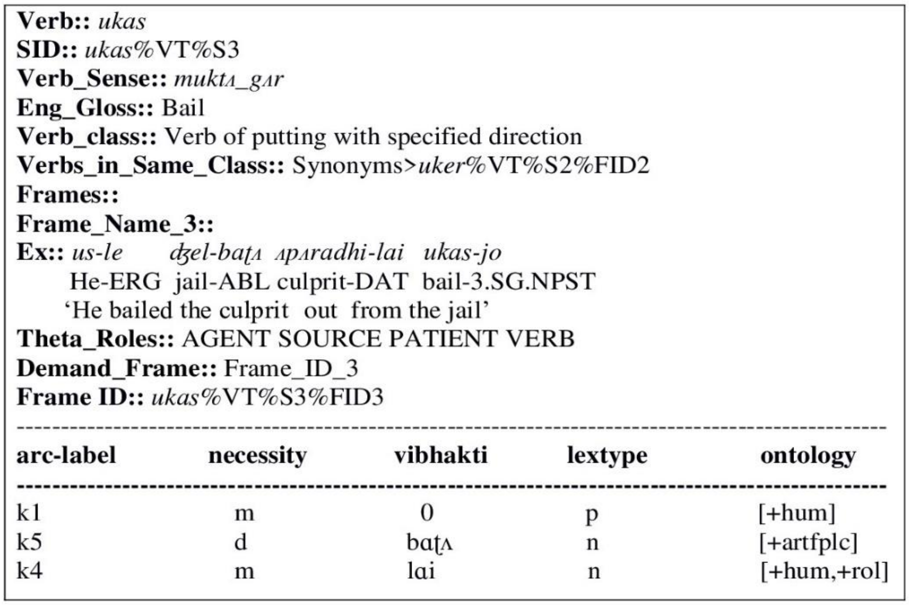
Ambiguity is the primary criterion for the selection of data which means Nepali verbs having two to 15 different senses are selected for the study. Frequently used verbs with more obvious ambiguous senses are preferred for making verb frames. Verb forms with simple past tense are incorporated in this study since each tense and aspect shows different morphology which can be resulted in different senses of a verb. Also, the motive behind this study is to present a small database of verb frames for Nepali which would help to build a large-scale database in future.

4. Components of Verb Frames

The Verb Frame in this study contains two kinds of information about a particular verb such as description and the verb frame. These two pieces of information are provided in a data file which is referred to as a 'verb entry' by Begum (2017: 95). The description part of the verb frame includes the information like Verb name, Sense-id, Verb Sense, English gloss, Verb in the same class, Example sentence, Theta roles and Frame ID whereas verb frame consists the information like Arc-label, Necessity of the argument, *Vibhakti*, Lexical Type and Ontology. Since, Indian languages like Nepali are Subject Object Verb (SOV) dominant phrase order language and the position of arguments to their predicate always remains left in almost all cases and since the dependency grammar treated arguments as children of a verb, their relation with arguments also remain constant in all cases, the fields 'position' and 'relation' are removed in study which are incorporated in the verb frame developed by Begum (2017). Also, new fields like 'verb class' and 'ontology' are added for this kind of information about a verb found to be helpful in identifying a particular sense of a verb in the context.

Figure 4.1 represents the sample of verb frame developed for the Nepali verb *ukas* which has three senses: ‘pull out’, ‘provoke’ and 'bail out'. The figure below demonstrates each component of the verb frame for the third sense of the verb *ukas*.

Figure 4.1 Sample of Verb Frame



Verb:: *ukas*
SID:: *ukas*%VT%S3
Verb_Sense:: *muktʌ_gʌr*
Eng_Gloss:: Bail
Verb_class:: Verb of putting with specified direction
Verbs_in_Same_Class:: Synonyms>*uker*%VT%S2%FID2
Frames::
Frame_Name_3::
Ex:: *us-le dʒel-baʈʌ ʌpʌradhi-lai ukas-jo*
He-ERG jail-ABL culprit-DAT bail-3.SG.NPST
‘He bailed the culprit out from the jail’
Theta_Roles:: AGENT SOURCE PATIENT VERB
Demand_Frame:: Frame_ID_3
Frame ID:: *ukas*%VT%S3%FID3

arc-label	**necessity**	**vibhakti**	**lextype**	**ontology**
k1	m	0	p	[+hum]
k5	d	bɑʈʌ	n	[+artfplc]
k4	m	lɑi	n	[+hum,+rol]

As we see in Figure 4.1 the first field is **Verb,** the name of a particular verb in Nepali which is written in IPA for its appropriate pronunciation.

The second field is **Sense ID** (SID) which is represented as *ukas*%VT%S3. It is a unique identification number for the particular sense of a verb. It consists of verb name, verb type and sense number which are separated by the symbol of percentage (%). In the verb frame above, *ukas* is the verb name, 'VT' is the 'transitive verb' and 'S3' is the 'Sense 3' of the verb respectively. The verb type is of three kinds: Transitive Verb (VT), Di-transitive Verb (VDT) and Causative Verb (VCAUS). Different Senses of a verb are denoted by the convention of S1, S2, S3 and so on according to the number of senses a verb can have.

Verb Sense is the third field which represents each sense of a given verb in Nepali. For example, in Figure 4.1 the third sense of the verb *ukas* is *muktʌ_gʌr* 'bail out'.

The fourth field is **English Gloss** which gives the meaning of a verb in the English language. It is represented as 'Eng_Gloss' in the verb frame. In Figure 4.1, 'Bail out' is an English gloss for a particular sense of the verb *ukas*.

The fifth field represents **Verb Class** which is a semantic class of a verb based on Levin (1993). In Figure 4.1 above, 'bail out' falls under the semantic class of 'Verb of putting with a specified direction'. This kind of information is included because it helps to provide insights into the disambiguation of Nepali verbs since the verbs having the same semantic class behave similarly and have similar argument structures. It also helps to categorize Nepali verbs in certain semantic classes.

The **Verb in the Same Class** is the sixth field which provides information about the synonymous verb frame found in the lexicon. It means that there are several verb framess which has similar senses which are considered synonymous with each other or the verbs in the same class. The verbs in the same class are represented by the unique frame ID they possess. The information about the same class is necessary to see whether or not the verbs in the same class behave similar way.

The seventh field is **Example** which is an example sentence for each sense of a verb which is denoted by the convention of 'Ex'. The example is given in Nepali which is transcribed in IPA. The conventions and the methods of glossing are written following interlinear morpheme-by-morpheme glossing provided by Leipzig Glossing Rules (2015). The following example is given for the verb ukɑs in Figure 4.1 above:

us-le	*ʤel-baʈʌ*	*ʌpʌradhi-lai*	*ukas-jo*
He-ERG	jail-ABL	culprit-DAT	bail-3.SG.NPST

'He bailed the culprit out from the jail'

The eighth field in the verb frame is **Theta Role** which refers to the 'role performed by each argument of a predicate, defined regarding a restricted universal set of thematic functions' (Crystal, 2003: 463). The definitions provided by VerbNet (2006) have been taken as the model for annotating theta roles of each argument.

Frame Name, **Demand Frame** and **Frame ID** are the fields which are all about naming and giving a unique ID to each verb frame for each sense of a particular verb. In Figure 4.1, Frame_Name_3 is the frame name, Frame_ID_3 is Demand Frame and *ukas*%VT%S3%FID3 is frame ID for the third sense of the verb *ukas*. The frame ID consists of two parts: Sense Id (e.g. ucɑl%VT%S1) and the Frame Number (e.g. FID1) which are separated by a percentage sign (%). It helps to identify a particular frame for a particular sense of a verb.

The term **'verb frame'** is used in two senses in this study as in Begum (2017). In a broad sense, it refers to the whole data, i.e., the verb entry (both 'description of the verb' and 'verb frame') and in a particular sense it refers to the tabular form which is a verb frame. 'The actual verb frame is the table given in the verb entry' (Begum, 2017: 99) which represents five pieces of information such as Arc Label, Necessity, Vibhakti, Lexical Type and Ontology.

The first field in the verb frame is the **Arc Label** which represents the dependency relation or *karaka* relation of arguments. Conventions used for each type and subtypes of *karakas* are based on 'AnnCorra: TreeBanks for Indian Languages Guidelines for Annotating Hindi Tree Bank' developed by IIIT, Hyderabad (Bharati et al, 2012). Figure 4.1 demonstrates the method of representing karaka information in the verb frame. The sentence given in the figure has three arguments: k1 (*us-le*), k5 (*ʤel-baʈʌ*) and k2 (*ʌpʌradhi-lai*) respectively and they are captured in the first field i.e. Arc Label.

The second field is **Necessity** which is the information about the necessity of argument in the sentence. In the example sentence given in Figure 4.1, *us* 'he', *ʤel* 'jail' and *ʌpʌradhi* 'culprit' are three arguments of the predicate *ukas-jo* 'bailed out' which have theta roles of 'agent', 'source' and 'patient' respectively.

The third field is ***Vibhakti*** which provides information about the type of case marker or any other postpositional element that comes with a particular argument. In the example sentence, as given in figure 4.1, there are three vibhaktis *–le* (ergative), -*baʈʌ* (ablative), -*lai* (dative) attached with each argument (*us*, *ʤel*, *ʌpʌradhi*) respectively.

The fourth field is **Lexical Type** which contains information about parts of speech of an argument.

Finally, the fifth field consists of **Ontology** which captures information about the semantic properties of the concepts and their relationship with each other. It is useful in determining selectional restrictions of a verb which further helps to disambiguate words having more than one sense. The tag set prepared for 'Shared Task cum Workshop on OntoLex in Indian Languages' by the Center for Applied Linguistics and Translation Studies, University of Hyderabad which was held from 29th November to 1st December 2017 has been used for annotating ontologies of arguments in this study. The ontology for the arguments of the example sentence given in Figure 4.1 are marked as +hum (human) for *us* 'he', +artcplc (artefact place) for *ʤel* 'jail' and +hum, +rol (human, role) for *ʌpʌradhi* 'culprit' respectively.

5. Conclusion

Nepali is in the primary stage of natural language processing and is considered one of the least-resourced languages in terms of NLP research and development. Thus, the development of linguistic resources in every aspect is necessary for this language. In this regard, this study is a small step towards building a knowledge base of verbs which can pave the way for the larger database in future.

Verb frames as a linguistic resource have four basic implications such as i) it is useful as a database for knowledge-based NLP applications; ii) it is helpful to understand the verbal behavior of Nepali verbs; iii) it can be used as a tool for annotating (especially parsing) Nepali verbs; and

iv) it can be used as a tool for word sense disambiguation. Though, this study embodies a very less and limited number of verbs, the verb frames developed through this study can be used as a model of a larger database of verb frames for Nepali which would contribute to the field of NLP tasks and applications like parsing, word sense disambiguation and machine translation.

==

References

Baker, Collin F, Charles J. Fillmore and John Lowe. (1998). 'The Berkeley FrameNet Project'. In *COLIN '98'*. Proceedings of the conference. Retrieved from https://aclanthology.org/P98-1013.pdf

Begum, Rafiya. (2017). *Developing a Pilot Hindi Treebank Based on Computational Paninian Grammar*. An unpublished PhD Dissertation. International Institute of Information Technology, Hyderabad, Telangana

Chomsky, N. (2015). *Syntactic Structures* (2nd Edition). New York: Mouton De Gruyter

Crystal, David. (2003). *A Dictionary of Linguistics and Phonetics* (5th Edition). Hoboken, New Jersey: Blackwell Publishing

Levin, Beth. (1993). *English Verb Classes and Alternations*. Chicago: University of Chicago Press

Lohani, Shridhar Prasad and Rameshwar Prasad Adhikari (Eds.). (2010). *Ekta Concise Nepali-to-English Dictionary*. Kathmandu, Nepal: Ekta Books

Palmer Martha, Dan Gildea, Paul Kingsbury. (2005). 'The Proposition Bank: A Corpus Annotated with Semantic Roles'. In *Computational Linguistics Journal*, Vol. 31:1. Retrieved from https://aclanthology.org/J05-1004.pdf

Parajuli, Krishna Prasad (Ed.). (2010). *Nepali Brihat Sabdakosh*, (7th Edition). Kathmandu, Nepal: Nepal Academy

Poudel, Madhav Prasad. (2015). *Nepali Kriyaharuko Kosh*. Kathmandu, Nepal: Vidhyarthi Prakashan Private Limited

Schuler, Karin Kipper. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Retrieved from https://repository.upenn.edu/dissertations/AAI3179808

==