

A Feature Based Approach to Translating Cuisine Verbs of Telugu and Bangla

*K.Rajyarama
Abhijit Debnath*

Abstract

This paper attempts to establish a feature-based approach to translate words related to the semantic domain of 'cuisine' with particular emphasis on verbs. Identification and listing of features belonging to a particular semantic domain is a challenging task as both deterministic and delimiting criteria need to be developed. Feature matrices thus developed are of immense help in tasks like translation, where cross-linguistic mapping of the maximum number of features will assist in selecting an appropriate equivalent. The main objective of the paper is to unify the established semantic theories like componential analysis, semantic domains and the implementation of the Lambda Calculus.

Keywords: Feature based approach, semantic domains, cuisine, translation, componential analysis, Lambda Calculus.

1.0 Introduction:

One can witness a marked increase in the study of culinary and gastronomic literature from historical, social and linguistic perspectives in the past few decades. Because of increased opportunities to travel, internet as a major resource for accessing culinary practices across the world, availability of hitherto unavailable ingredients, and the effect of globalization, cuisine

literature has gained significance. Translating cook books has gained importance in this process and acquired the nature of a professional activity. Since cooking is an integral part of culture and cooking practices vary significantly from one culture to another, translating cuisine literature is quite a challenging task. The problems of translating cook books are distinctive and deserve a serious study. In this paper we focus on translating cook books from a linguistic perspective taking a semantic factorization approach. The central idea is to develop a set of language-independent semantic features which are capable of handling translation both manual as well as machine, across languages.

We propose to develop a ‘Feature interactive model’ (FIM) based on a set of interactive semantic features for translating the cuisine verbs of Telugu and Bangla, the two widely spoken Indian languages belonging to Dravidian and Indo-Aryan language families respectively. Large scale ‘feature data bases’ representing various semantic domains, play a crucial role in natural language processing tasks like word sense disambiguation, information extraction and retrieval, question and answering, machine translation etc. A lot of research in Indian languages is being aimed at developing grammatical and lexical resources like morphological analysers, parts of speech taggers, parsers and WordNet etc. over the past few decades. The present work is an attempt to develop lexical resources which help in translating at word level.

1.1 Feature based approach has an edge over the conventional lexicons in the following ways:

1. The traditional organization of lexicons as in componential analysis, Katz's theory (1972) and others advocate that word meaning can be defined exhaustively

by a finite set of senses per word. When such a system encounters instances of lexical ambiguity, the above theories presume that disambiguation can be achieved by predetermining the set of contexts in which the ambiguous words may appear. The recursive nature of language, however, renders this approach an unmanageable and humongous task.

2. Further, dictionaries and lexicons are static by nature and they fail to account for the creative use of words in new contexts.
3. Feature-based approach, as proposed in this paper, can overcome these flaws by accommodating the novel contexts in which a word might appear. A new feature that may arise due to the use of a word in a novel context can be accommodated in the feature matrix making it dynamic by nature. A more elaborate explanation is provided in section 3.

The 'Feature Interactive Model' proposed here is based on certain aspects of Pustojevsky's (1995) Generative Lexicon.

1.2 Problems of lexical transfer:

The difficulties in translating lexical items involve:

1. Firstly, what counts as one word in one language can be expressed by two or more lexical items in another language.
2. Secondly, the fragment of outside reality covered by a lexical label in one language doesn't usually correspond exactly to the available equivalents in another language.

3. Thirdly, there may be absence of a lexical item in a structured pattern of lexical field that is conventionally called as lexical gap.

Specifying feature sets as envisaged in this paper would address the problems cited above through establishing ranked hierarchy of preferred choices thereby enabling an effective translation.

Feature sets of a particular semantic domain, cuisine in this particular context, combined with one of the existing models of formal semantics like ‘characteristic function of features’ provides a viable solution for resolving problems of lexical transfer.

The paper is organized into four sections. The second section deals with data drawn from both Telugu and Bangla. The third section discusses the model proposed. Section four concludes the paper.

1.3 The notions of Semantic Domains and Semantic features:

The notions of semantic domain and semantic features as employed in this paper are based on their conventional definitions. ‘... Semantic Domains are common areas of human discussion, such as Economics, Politics, Law, and Science which demonstrate lexical coherence. ... The Semantic Domain associated to a particular field is the set of domain specific terms belonging to it, and it is characterized by a set of domain words whose main property is to co-occur in texts ... (Gliozzo, 2006:4). The implications of this work follow from the usefulness of semantic domains to various natural language processing tasks. At a lexical level, Semantic Domains can be used as a (shallow) model for lexical ambiguity and variability, while at a textual level semantic domains provide meaningful topic

taxonomies that can be used to group texts into semantic clusters. In addition, the inherent multilingual nature of semantic domains allows a uniform representation of both the lexicon and the texts in most of the natural languages (ibid).

Semantic features represent the basic conceptual components of meaning for any lexical item. They help in establishing semantic domains. Words that share certain features may be members of the same semantic domain. Correspondingly, diverging semantic features help establish contrast in meanings of words.

The taxonomy of the cooking verbs and the diagnostic features (the sets of semantic features) adopted in this work with a few exceptions, are modelled on Lehrer (1969). A number of features devised in the present work, may hold good cross-linguistically to describe the processes of cooking at a generic level. Distinctions in culinary culture and practices among the speech communities are reflected in their vocabulary. The cuisine terms of each language possess certain features that are unique and unavailable in other languages. The feature matrices proposed in this paper aim to represent these as well.

A detailed analysis of cooking terms would include syntactic derivations like nominalizations, morphological modifications and collocations and also figurative use like metaphorical extensions etc. In this paper, we confine ourselves to the primary meaning of the terms and do not delve into other extended usages. The verbs chosen for study in this paper are those that are directly used in the process of cooking and hence have [+heat] as a default feature. Verbs denoting ancillary actions to cooking like chopping, slicing, grinding, pounding etc. have been excluded.

2.0 Cuisine verbs of Telugu and Bangla:

The cuisine verbs which form the data for the present study are based on the native speaker's intuitions.

Telugu	Bangla
1. vaMDu 'to cook'	1. ranna (to cook)
2. uDuku 'to boil'	2. shedd ^h o (boil)
3. veepu 'to fry'	3. b ^h apa (steam)
4. kaalcu 'to roast'	4. b ^h aja (to fry)
5. maggu 'to simmer'	5. shenka (to roast)
6. iguru 'boil till evaporation'	6. poraa (to burn)
7. kaagu 'to heat'	7. garamkora (heat up)
8. marugu 'boil for a long time'	8. naracarakora (stir)
9. terlu 'wavy boiling'	9. shedd ^h o (solids)
10. veeDekku 'to heat'	10. p ^h ota (liquids)

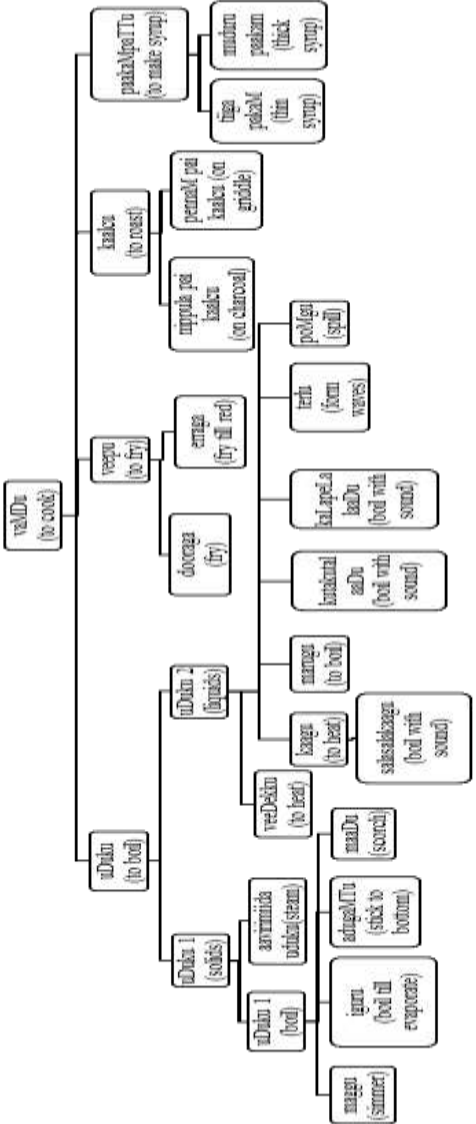
Table 1

The following list of features identifies the verbs in an exclusive manner. The binary values [1/0] imply the presence or absence of the feature.

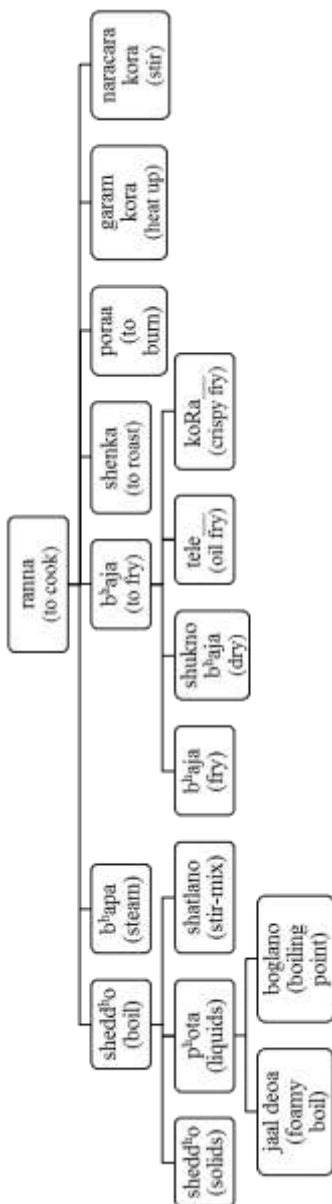
- | | |
|---------------------|---------------------------|
| 1. Heat | c. Steam/vapour [1/0] |
| a. Intense [1/0] | Steam/vapour [1/0] |
| b. Medium [1/0] | 5. Liquid |
| c. Low [1/0] | a. Water [1/0] |
| 2. Flame | b. Oil/Fat [1/0] |
| a. High [1/0] | 6. Viscosity [1/0] |
| b. Sim [1/0] | 7. Vessel |
| c. Low [1/0] | a. Flat [1/0] |
| 3. State | b. deep [1/0] |
| a. Soft [1/0] | 8. Process |
| b. crisp [1/0] | a. Cook [1/0] |
| c. dry [1/0] | b. Fry [1/0] |
| d. Evaporated [1/0] | c. Boil [1/0] |
| e. Melted [1/0] | d. simmer [1/0] |
| 4. Medium | 9. Cooking Time |
| a. Null [1/0] | a. Long [1/0] |
| b. Liquid [1/0] | b. Short [1/0] |

Table 2

The terms in a semantic domain enter into the relations of synonymy, antonymy, and hyponymy. This is evident of cuisine terms as well. The following flowcharts illustrate this.



**Telugu Cuisine Verbs:
Table 1**



Bangla Cuisine Verbs: Table 2

The feature assignment proposed in this paper is exemplified below with the help of the verb *iguru* 'boil till evaporation' drawn from the source language Telugu:

A recipe instruction (Malati Chandur, 1974) in Telugu reads as follows:

... ciMtapamDu pulusu igiripooyi mukka baagaa mettagaa
ayyaaka diMcukoovaali.

Tamarind juice evaporated vegetable piece well soft after
Remove.

'...remove the curry after the tamarind juice
evaporates and the vegetable gets softened well'

In the above example, the word *iguru* 'to boil till evaporation' is used as a complex predicate *igiripoovu* 'to evaporate completely' this has an intensifying function. The context in which *iguru* is used enables us to associate the following features with it.

<p>1. Heat a. Intense [0] b. Medium [0] c. Low [1] 2. Flame a. High [0] b. Sim [1] c. Low [1] 3. State a. Soft [0] b. Crisp [0] c. Dry [0] d. Evaporated [1] e. Melted [0] 4. Medium a. Null [0] b. Liquid [1] c. Steam/vapour [0]</p>	<p>1. Liquid a. Water [1] b. Oil/Fat [1] 2. Viscosity [0] 3. Vessel a. Flat [0] b. Deep [1] 4. Process a. Cook [1] b. Fry [1] c. Boil [1] d. Simmer [0] 5. Cooking Time a. Long [1] b. Medium [0] c. Short [0]</p>
--	---

Table 3

3.0 Feature Interactive Model (FIM): A working hypothesis:

In this section, we describe the various components of the proposed model. This model includes the identification of semantic features associated with each word (the cuisine verbs in particular) followed by building of a feature matrix for each word. Then an interactive device is developed that works as a two-step method,

- a) *Feature Check*: Take any given word as input in any given source language and identify the features built for that word.
- b) *Feature Mapping*: Mapping this feature set onto any specified target language word. This process will be based on conditionally sequenced operations.
 - i) *Equal match*: If an exact match of feature set is found in the target language, then that feature matrix will be considered as first priority and the model will extract the word which has been given that feature matrix. Or else, if an exact feature matrix match is not available, then the model will look for a ranked match.
 - ii) *Ranked Match*: The model will look for a number of feature matrices in the specified target language and rank them for most number of matches with the matrix of the given input word. Thus, it will create a ranked list of matrices. Then it will pick the highest ranked feature matrix and extract the word which has been given that feature matrix.

The most generic term is ranked highest in the hierarchy and the related terms, as they acquire more specificity, are placed at the lower levels. The terms placed at the same level are mutually

incompatible making substitution impossible. However, the substitutability of the terms across the levels is highly constrained. The number of levels crossed to replace a word has to be as minimal as possible.

The distribution of the cooking verbs is restricted by the objects with which they collocate even though the semantic process of cooking is the same. The model proposed will be able to account for the issues of ambiguity and polysemy by taking into account collocational restrictions.

Telugu: *kaacu* 'to boil (liquids)' cannot be replaced either by *vaMDu/vaMTaceyyi* 'to cook' or *uDuku* 'to boil' as *kaacu* is a specific term which is realized only in the case of liquids

paalu/niiru / nuune ... kaacu
milk /water/ oilboil
'boil/heat milk, water, oil...'

In the case of solid substances, the verb would be *uDuku* 'boil', though semantically the process of 'boiling' is subsumed in both the categories.

Similar mechanism is operative in Bangla as evident from the following example.

*du:d^h/jol / caa/kheer ... jaal-deo/*sheddho-koro*
milk / water / tea / kheer boil
'boil/heat milk, water, tea, kheer...'

Likewise, in the case of 'boiling solids' a different collocation is licensed.

*aalu/ dim / gajor ... sheddho-koro/*jaal-deo*
potato / egg / carrot boil
'boil potato, egg, carrot, kheer...'

The formalization of the Feature Interactive Model is a direct implementation of Characteristic Feature Set and Characteristic Functions from the Lambda Calculus (Church, 1965) borrowed from the area of compositional semantics in logic to create a computational model suitable for the interaction. As illustrated in Figure 1, the calculus provides a decision making assistance for the model in the Feature Check and Feature Ranking operations in selecting any given word according to feasible matches of given feature set values (available in binary properties). The device will take the input word in the source language (here Telugu) and find its way to the final binary truth value from which the feature checking module will be able to extract a feature-value matrix and superimpose the same upon all feature matrices in the target language and retrieve the matching or best ranked match on the word and extract that word.

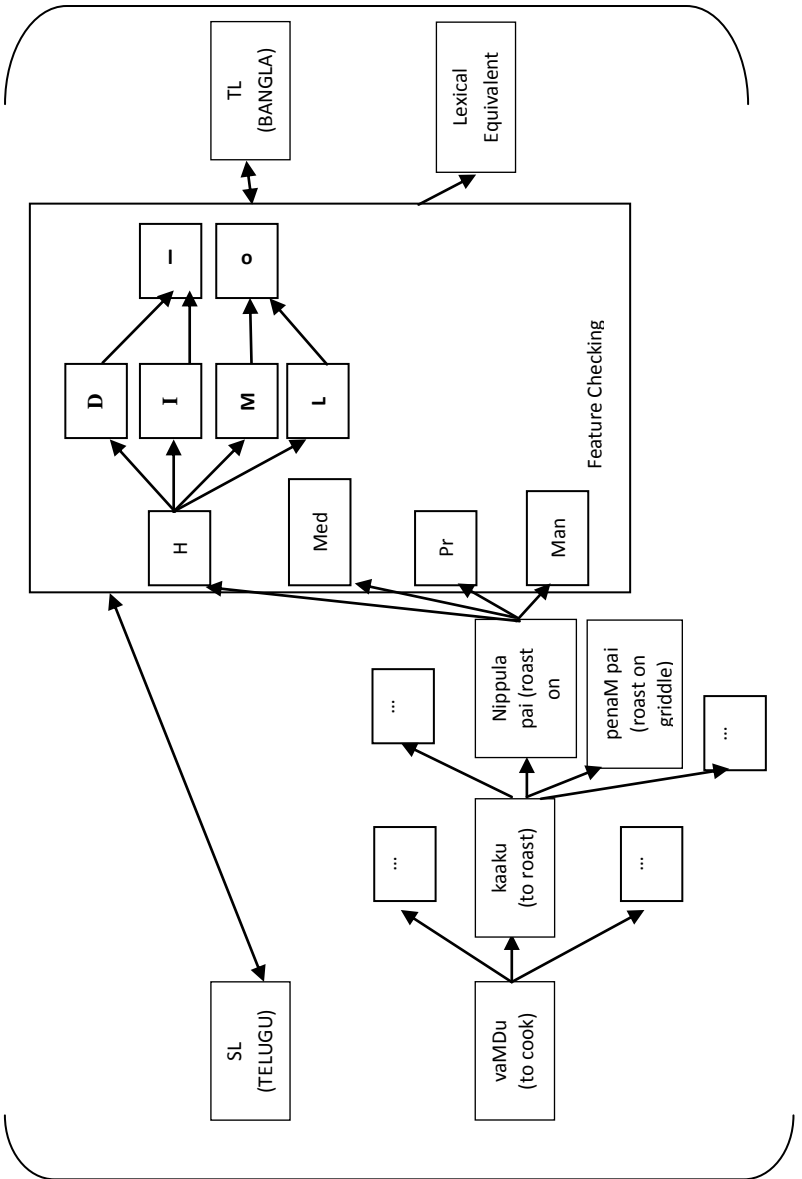


Table 3a

4.0 Conclusion:

The following inferences can be drawn from the above analysis:

1. Constructing feature matrices is based on analysis of the verbs at three levels:
 - a. Coarse grain
 - b. Medium grain
 - c. Fine grain

A fine grain analysis would involve identifying many more features and thereby feature sets that would capture specific processes of cooking. This should enable capture of the processes that are culture specific and facilitate translating language specific and culture specific terms.

2. The loops available at each layer in the model (see Fig 1) accommodate the distinctive feature of recursiveness of natural languages in an elegant manner.
3. The model also gives scope for scalability to other semantic domains.
4. New terms entering the semantic domain would enrich the existing data bases. A new term that enters, brings along with it a number of associated semantic features which would fit into already existing feature sets or result in recognizing a new feature set.

This paper reports the work in progress and checking the formalism proposed, against a corpus would be the next step in implementing it. This work underlines the need for building feature databases on a large scale for various semantic domains in Indian languages, in view of the fact that very few such resources are available. An integration of established linguistic theories like

componential analysis, generative lexicon and those of formal semantics like characteristic functions of features would help evolve robust and efficient NLP tools.

Notes:

1. For the full form of the abbreviations in Table 3, refer Table 2 & 3.

References:

- Church, A. (1965). *The Calculi of Lambda-Conversion* (1st ed.). New York: Kraus Reprint.
- Gliozzo, A. (2006). Semantic Domains and Linguistic Theory. In Proceedings of the LREC workshop "Toward Computational Models of Literary Analysis", Genova, Italy.
- Lehrer, Adrienne. (1969). Semantic Cuisine. *Journal of Linguistics*, 5,39-56.
- Levin, B. (1993). *English Verb Classes and Alternations* (1st ed.). Chicago, Ill. u.a.: Univ. of Chicago Press.
- MalatiChandur. (2015) *VantaluPindiVantalu - వంటలుపిండివంటలు*. Vijayawada: Quality Publishers.
- Pustejovsky, J. (1995). *The Generative Lexicon*, Cambridge: MIT Press.
- Rajyarama, K. and Benarjee, S. (2011). *A Study of Lexico-Semantic features of Cuisine verbs of Telugu and Bangla for Machine Translation*. Paper presented at South Asian Languages Analysis Roundtable (SALA).