Experiences in Building the Konkani WordNet
Using the Expansion Approach

Shantaram Walawalikar
ILCI - Konkani Team
Goa University
goembab@yahoo.co.in

Shilpa Desai
Dept. of Computer Science
& Tech., Goa University
sndesai@gmail.com

Ramdas Karmali
Dept. of Computer Science
& Tech., Goa University
rnk@unigoa.ac.in

Sushant Naik
ILCI - Konkani Team
Goa University

Damodar Ghanekar
ILCI - Konkani Team
Goa University

Chandralekha D'Souza
Dept. of Konkani
Goa University
chanda@unigoa.ac.in

Jyoti Pawar
Dept. of Computer Science
& Tech., Goa University
jdp@unigoa.ac.in

Abstract

WordNet can be described as an electronic lexical database available on-line as a powerful resource to the researchers in the area of computational linguistics, text processing and many other related areas. Currently, the necessity of building WordNets has been felt for all the Indian Languages to aid in multi lingual machine translation and cross lingual information retrieval to promote tourism, farming, education and other related areas for overall growth and development of the nation. IIT Bombay, India has developed a number of tools, resources and facilities by which WordNet of any language can be constructed through what is known as the expansion approach. Projects to create WordNet in most of the Indian languages using this approach with Hindi WordNet as the base are currently in progress. In this paper we report our experiences of creating a WordNet for Konkani language using the expansion approach with Hindi as the source language and Konkani as the target language. The Konkani WordNet is in the initial stage of development. The 1969 Hindi core synsets have been incorporated in the Konkani WordNet. The Offline Synset Linking Tool developed by IIT Bombay is being used for this task.

1. Introduction

WordNet can be described in short as a massive structure of words in a graph like form. It is an electronic lexical database available as a powerful resource to the researchers in the area of computational linguistics, text processing and many other related areas. Since 1987 when WordNet first appeared globally, it has come a long way, getting itself moulded as per the ongoing requirements of the users and making use of the advancement of technology viz. Computer Science and Communications. Indo WordNet is India's contribution to this global effort and the steps towards the development of Konkani WordNet शब्दभारं shabdamAleM is a part of this initiative.

The layout of this paper is as follows – Section 2 discusses the characteristics of Konkani language. A brief description of the Hindi WordNet, the expansion approach used to create Konkani WordNet, observations made during the WordNet creation process and challenges faced are given in section 3. Section 4 concludes the paper with a discussion on the future work plan.

2. Characteristics of Konkani Language

Konkani language is one of the twenty two languages included in the eighth schedule of the
constituent of India. It is also the official language of the State of Goa. Konkani is an Indo-European (Indo-Aryan) language derived from Sanskrit through Prakrit and is influenced and enriched by various other languages like Marathi, Kannada, Malayalam, Hindi, Portuguese and English. Though Devanagari script is recognised as official script of Konkani, it is also written in Roman and Kannada scripts. Old Konkani literature is also found written in Malayalam and Urdu scripts.

The first edition of Konkani grammar titled ‘Arte da Lingua Canarin’ was written somewhere in 1617 A.D. by Fr. Thomas Stephens (Asmitai, 2008; Cunha, 1958). It was enlarged by Fr. Diogo Ribeiro and revised by four priests of the Society of Jesus, and printed in 1640. This is considered to be the first published grammar not only of Konkani but of any Indian language. Monsignor Sebastiao Rudolpho Dalgado was the first known Indian lexicographer of Konkani as those preceded him were all European missionaries. He contributed to the development of Konkani with his three important works ‘Konkani – Portuguese Dictionary’ (1893), ‘Portuguese – Konkani Dictionary’ (1905), and ‘Bouquet of Konkani Proverbs’ consisting of 2177 proverbs.

2.1 Pronunciation

Shennoi Goembab (1949) in his book, ‘Konkanichi Vyakarani Bandavol’ discusses pronunciation in detail. Konkani pronunciation for अ, ए, ओ, औ have additional pronunciations besides the original Sanskrit pronunciations. अ in पणस paNasa ‘jackfruit’ is known as स्वरित svarita in Vedic Sanskrit. ए and ओ also have open pronunciations. These open pronunciations must have been influenced by Pali language. These are found in other Indian languages like Bengali, Bihari, Gujarati, Kannada, Telugu, Tamil, Malayalam, etc, but it is not found in Marathi.

In Konkani, according to the pronunciation of a vowel in the same word, the meaning changes e.g. पौर perisha ‘guava fruit or guava tree’, मौर mura ‘peacock, sl. or peacock, pl.), वौष्ठ्व voMvaLa ‘kind of flower – minusops elengi flower or its tree’.

2.2 Number

Konkani has two numbers - singular and plural (Sardesai, 1986). The derivation of plural form from singular form is dependent on gender and phonetic characteristic of singular form. In some cases, the change in pronunciation of the vowel denotes change in number, e.g. दौतीर dotara ‘doctor or doctors’, फाटर phAtara ‘stone or stones’, देय dera ‘brother-in-law or brothers-in-law’, औंठ oMTha ‘lip or lips’.

2.3 Gender

Konkani has three genders - masculine, feminine and neuter. However, in some cases feminine nouns are also addressed as neuter e.g., कमला kamalA AMgaNAMta kheLatAleM ‘Kamala was playing in the courtyard’. Here, the verb खेलताले refers to the neuter gender whereas Kamala is otherwise a feminine noun. It is also interesting to note that two synonymous nouns may have two different genders, e.g., रूख rUkha ‘tree, masculine’ and झाड jhADa ‘tree, neuter’.

2.4 Word Structure

Konkani is a highly inflected language (Almeida, 1989). Nouns and pronouns are inflected for number and case. Verbs are inflected for person, number, gender, tense and aspect. Adjectives are inflected for gender and number. The Structure of Konkani word (Goembab, 1949; Borkar, 1986) can be depicted as under:

**Nominal Base** (N.B.) + **Nominal Inflection** (N.I.)

पुस्तकांचे । पुस्तकांचे}}{{N.B.}} + **PustakAcheM । of the book**

(पुस्तकांचे । पुस्तकांचे) + **N. B.** + **Postposition**

रामाकडल्यान {{रामाकडल्यान}} + **rAmAkaDalyAna । from Ram**

(रामाकडल्यान {{रामाकडल्यान}}) + (N. I.) + (N. I.)

पुस्तकांतांचे {{पुस्तकांतांचे}} + **PustakAmtaleM । from the book**

(पुस्तकांतांचे {{पुस्तकांतांचे}}) + (N. I.) + **Postposition**

पुस्तकांतांचे {{पुस्तकांतांचे}} + **PustakAmtaayAna । from inside the book**
2.5 Verb Base

The verbal base of Konkani has three sources (Goembab, 1949), present active base, present passive base and past passive participles. The roots are either active or passive in sense, the passive being intransitive and the active being transitive. The following is a sample of these forms separated with base form of verb:

### Non Perfective

#### Intransitive

The verb: धांलव प ‘to run’

<table>
<thead>
<tr>
<th></th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>धांलतां</td>
<td>धांलतात</td>
</tr>
<tr>
<td>1st person</td>
<td>धांलता</td>
<td>धांलतात</td>
</tr>
<tr>
<td>2nd</td>
<td>धांलता</td>
<td>धांलतात</td>
</tr>
<tr>
<td>3rd</td>
<td>धांलता</td>
<td>धांलतात</td>
</tr>
</tbody>
</table>

In the present tense, gender has no effect. But the verb endings change as we go to all other cases and are differentiated below with the respective affixes in the sequence of masculine, feminine and neuter.

#### Imperfect

<table>
<thead>
<tr>
<th></th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>धांलरां ल्रीं ल्रं</td>
<td>धांलरे ल्ल्मं ल्रीं</td>
</tr>
<tr>
<td>1st person</td>
<td>धांलरा ल्री</td>
<td>धांललरे ल्ल्मा</td>
</tr>
<tr>
<td>2nd</td>
<td>धांलरा ल्री</td>
<td>धांललरे ल्ल्मा</td>
</tr>
<tr>
<td>3rd</td>
<td>धांलरा ल्री</td>
<td>धांललरे ल्ल्मा</td>
</tr>
</tbody>
</table>

#### Past

<table>
<thead>
<tr>
<th></th>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>धांवललरं ल्रीं ल्रं</td>
<td>धांवललरे ल्ल्मं ल्रीं</td>
</tr>
<tr>
<td>1st person</td>
<td>धांवललरा ल्री</td>
<td>धांवलललरे ल्ल्मा ल्री</td>
</tr>
</tbody>
</table>
### 2.6 Contextual Word Usage

There are different Konkani words for the similar sense denoting variety of shades.

**Example 2.6.1:** An example of this is the meaning of the noun ‘stink’ in Konkani being गुड्डालऩ guD.hDaPa ‘stink’. It is used in the following variations - पोमवसANa ‘smell of spoil fish’. हिंलSAsAnA ‘natural smell of fish’. खातसANa ‘smell of urine’. घामवसAnA ‘smell of sweat’. कानुत्ताण kAnuTaTANa ‘smell of utensil in which food preparation of onion is made’.

**Example 2.6.2:** There are verbs which depict many shades of the word ‘beating’.

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Konkani</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st person</td>
<td>खाल्रो</td>
<td>to beat someone by putting under one's feet</td>
</tr>
<tr>
<td>2nd person</td>
<td>खाल्रो</td>
<td></td>
</tr>
<tr>
<td>3rd person</td>
<td>खाल्रे</td>
<td></td>
</tr>
</tbody>
</table>

**Past Perfect**

<table>
<thead>
<tr>
<th>Person</th>
<th>Konkani</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>खाल्रो</td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>खाल्रो</td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td>खाल्रे</td>
<td></td>
</tr>
</tbody>
</table>

#### 2.7 Homographic Words:

In Konkani, we also come across homographic words i.e. two words written alike but have different meanings.

**Example 2.7.1:** पेर त or त ‘pair’ as in English the meanings are 1. guava fruit 2. joint of finger 3. part between two nodes of a stem.

**Example 2.7.2:** The word for mango tree is आंफो aMduM and the mango fruit (sl.) is also आंफो आंफो आंफो aMduM bAjArAMta chaDa Ambo AyalonA ‘This year there was not much mango fruit in the market’.

### 3. Expansion from Hindi to Konkani

#### 3.1 Hindi Wordnet(HWN)

The Hindi WordNet (Narayan. et. al., 2002; Miller, 1995) on which our expanded model is based has currently 32950 synsets covering 77800 unique words. Out of these synsets, 13830 synsets are linked with the synsets of the Princeton WordNet. The synsets are constructed abiding by the following three principles -

(i) **Minimality** - use the minimal set of words to make the concept unique

(ii) **Coverage** - The maximal set of words-ordered by frequency in the corpus - to include all possible words standing for the sense.

(iii) **Replaceability** - The example sentence should be such that the most frequent words in the synset can replace one another in the sentence without altering the sense.

#### 3.2 Konkani Wordnet(KWN) Creation Process

Konkani WordNet is created by using Expansion Approach. In this approach, instead of reinventing the wheel, the readily available Hindi WordNet synsets developed by IIT Bombay, are referred to. They are, one by one, understood by the lexicographer and the
corresponding synsets in Konkani, expressing the same sense are created. Thus the HWN and KWN have identical glosses and examples as far as possible. This is being followed by many other Indian languages so that the resultant WordNet will take a shape of IndoWordNet.

According to Vossen (1996), the MultiWordNet Model seems less complex and guarantees the highest degree of compatibility across different WordNets. In the development of any WordNet a large number of subjective and sometimes far from accurate decisions are involved. Hence, building two different WordNets independently for two different languages, will display differences. Expand model tends to reduce these subjective choices and resultant discrepancies. It also to some extent helps in highlighting potential inconsistencies existing in the WordNet of the source language. But this does not mean that expansion model is without any drawbacks. As Vossen (1996) points out it forces "an excessive dependency on the lexical and conceptual structure of one of the languages involved".

In Konkani at present (at the time of writing of this paper) all the core synsets are linked covering around 3500 unique words. These synsets are classified according to parts of speech (nouns, verbs, adverbs and adjectives).

### 3.3 Observations:

Hindi and Konkani being close languages and with the sentence structure of both being 'Subject Object Verb' (SOV), there was not much problem in maintaining identical concepts.

Our observations during the WordNet creation process can be subdivided under the following 8 broad categories –

#### Hindi English incorrect linkage

Some of the details are presented below –

**Example 3.3.1:** Id. 2897- चिपकली Chipakall - एक रैंगलेयला जंतु जो प्रायः दौड़ते पर दिखाई देता है eka reMganevAlA jaMu jo prAyaH dlvAroM para dikhA_I detA hai ‘a crawling creature mostly seen on walls’. Hindi and English synsets have wrong linkage. English should have been House Lizard instead of Gecko.

**Example 3.3.2:** Id. 3016- व्यवहार की यह प्रकृति जो लगातार दोहराय से प्राप्त होती है vyavahAra kI vahA prakRRitI jo lagAAtAra doharAva se prApta hotI hai ‘behavioral characteristic acquired due to constant repetition’. Hindi concept is understood as “habit” by us while it has been linked to “custom” in English.

**Example 3.3.3:** Id. 3464- जिसे क्षयति मिली हो jise khyAti milI ho jise khyAti milI ho ‘one who is famous’. Hindi concept suggests that the concept is “famous” while English synset is “popular”.

**Hindi concept/gloss definition not clear**

Synset details of two such examples falling in this category are given below –

**Example 3.3.4:** Id. 231 concept in Hindi reads as किसी देशका वह विभाग जिसके निवासियोंकी शासन पद्धती. भाषा, रुझन सह व्यवहार आदी और राष्ट्रीय व्यक्तियाँ और स्वतंत्र हो kisi deshaka vahA vishAva prApyaH kI shAsee bhuuna aur svataMtra ho ‘that territory of any nation, of which the residents have administrative system, language, customs, tradition, etc. different and independent from others’.

For this the synsets are प्रदेश pradesha, राज्य rAjya, प्रति prAMta. Here the mention of ‘administrative system, language, customs, traditions etc. different and independent from others’ is superfluous. Since linkage to English synsets was available this was referred to. The English concept reads as 'the territory occupied by one of the constituent administrative district of a nation’, with synsets as state, province. This is more appropriate concept.

**Example 3.3.5:** In Id. 882 the Hindi concept reads as संध्मा का जब चर्कर लौटते आगे गो खुड़ों के घूल उड़ती है saMdhyA kA vahA samaya jaba charakara lauTanevAlI gau_oMke khuroMse dhUla uDatI hai ‘that time of the evening when the dust from the legs of cattle returning after grazing spreads in the air’.

The synset for the same is गोधुलिवलं godhUli belA. The synset literally translates as 'the time of dust from cattle'. Etymologically this word 'godhuli' (go = cow; dhuli = dust) may have a origin of coinciding this time with the return of grazing cattle who come running through the
dusty lands and the red dust gets mingled in the whole atmosphere around. But this description need not be a part of concept. Simple definition like ‘a short span of time before and after the sunset’ will meet the true sense of the concept and also abide by the principle of minimality. English synset was not available for this.

**English concept/gloss definition not clear**

Although our source language is Hindi, we had referred to English synset to get a better idea of the concept during which we made these observations. Following is an example

**Example 3.3.6:** Id. 3052 कोई बस्तु खरीदने या बेचने पर उसके बदले में दिया जानेवाला धन ko_I vastu kharIdane yA bechane para usake badale meM diyA jAnevAlA dhana ‘Money paid when any goods are bought or sold’. English concept given as ‘cost of bribing someone’ not appropriate to convey the meaning of price.

**English synset missing**

As stated earlier in section 3.1, only 13830 synsets are linked with the synsets of the Princeton WordNet. Hence we found many synsets falling in this category.

**English example missing**

In some of the cases where the English synset was linked the examples were found missing.

**Hindi example could have been better**

We felt that more examples that would overall enrich the WordNet and improve the accuracy of the applications using the WordNet can be used.

**English example could have been better**

Same observation as above can be made with respect to the English examples.

**Recursive definition of concepts**

It was also observed that in certain concepts the definition was recursive, i.e. the synset itself was referred to in the concept.

**3.4 Challenges Faced:**

**Linking culture specific concepts**

The customs and culture played a challenging role. We have experienced in this exercise that very culture specific concepts do not have their parallels in other languages. The linking of such synsets to other languages remains a question.

In Hindi region chhapati vendor possibly comes door to door selling his chhapatis (a thinly made bread like eatable prepared from wheat flour). Konkani speaking populace is not familiar with this scene but they have met a vendor popularly known as पदेर padera ‘bread seller’ visiting residential areas.

There was another such example of a type of saree of length nine yards popularly known as णववारI NavavArI (or नउवारI na_uvArI in Marathi) which women from Goa, Maharashtra and other parts of India wear. Major part of the population may not be aware of this concept.

**Linking of contextual words**

Using the expansion approach, certain synsets may totally get omitted because of the variety of shades of meanings of different words as mentioned in section 2.6 above.

**Coverage of synsets**

The question also arises with respect to the coverage for some of the synsets. Many words though the meaning of them is known to the people, are not in parlance or common in literature; one may find them possibly in poetry. The glaring example could be of सूर्य Surya ‘the sun’. Many people know that रवि ravi, आदित्य Aditya are other names of the sun. Likewise there are many words which are used for the sun in Puranas (ancient literature). Whether we have to cover these is a question. The role of metaphorical usage of words – Should they be included in the synset? E.g. सुंगट suMgaTa ‘prawn’ is commonly used metaphor in Konkani to mean a slim girl.

**Linking a concept not present in the source language**

The concept of a nine yard saree - Synset of this concept is not available in HWN. Hence Marathi WordNet has already created synset for this concept. Id number has also been assigned by MWN of its own. Since the other member languages would not know the existence of this synset, they would duplicate this under different IDs. Hence, centrally controlled system for issuing IDs will have to be established.
Coining of new words
Another issue that remains to be resolved is how far the lexicographer can be given liberty to coin new words. This issue comes up if a language does not have a word for a concept (typically happens for culture specific situations). This question will come after the other alternatives like transliteration and multiword expression (short phrases) are explored.

Computational concerns: Interface, efficiency of access and storage
Interface of Offline Synset Linking Tool could also show the relations like hypernymy, hyponymy, antonymy already defined for source language synsets so that if the same does not correspond in the target language it could be changed.

4. Conclusion and Future Work
WordNet has been a very essential constituent for any linguistic study. Hence creation of WordNet for Konkani language has been started. The expansion approach has been found most convenient to speed up the exercise. The software tool provided was also found adequate for the purpose. Though only the core synsets have been linked for the time being, the project is taking momentum and the rest of the synsets will also be linked with greater speed than earlier.
The Konkani WordNet शब्दभारंशब्दभारं is at the initial stage of creation. Currently only the concepts, synsets and examples have been dealt with. However, it is required to check all the semantic relations like synonymy, hypernymy, hyponymy, meronymy, holonymy, troponymy, entailment etc. Even the concept of gradations will have to be introduced as in Hindi WordNet. It is felt that the existing examples from the HWN should also be strengthened with our own additional examples. These examples could be taken from any of the existing Konkani Corpora. When the project gets completed it will be a useful tool for the computational studies of the Indian Languages and a valuable asset of the Konkani language in particular.

Acknowledgement
We wish to express our gratitude to the Indian Institute of Technology, Bombay (IITB) Hindi WordNet Team for providing the tools and guiding us in our process of creating the Konkani WordNet. We thank the Indian Language Corpora Initiative (ILCI) 11(12)/2008 – HCC (TDIL) project Team members for giving inputs and support to this Konkani WordNet creation process. We also acknowledge that we were able to carry out the work using some of the equipments that were purchased from the AICTE funding under RPS scheme 8023/BOR/RPS/091/06/07.

5. References
Sardesai Madhavi 1986. Some aspects of Konkani grammar. Department of Linguistics, Deccan College