



Sinhala Lexical Resources and WordNet

Dulip Herath
Language Technology Research Laboratory
University of Colombo School of Computing
Sri Lanka



Overview

- Lexical Resources in Sinhala
 - Dictionary Based
 - Corpus Based
- Sinhala WordNet
- Our Approach
- Issues



Lexical Resources in Sinhala

- Dictionary Based Resources (Phase 1)

- Lexicon consists of ~ 30,000 words
- POS tag information (num/gen/per/tense/voice/...)
- Synonyms / Meaning
- Tamil / English translation equivalents (common words)
- Sources:
 - National Institute of Education Dictionary (based on the main Sinhala Dictionary)
 - Department of Official Languages Dictionary (based on technical glossaries and commonly used words)

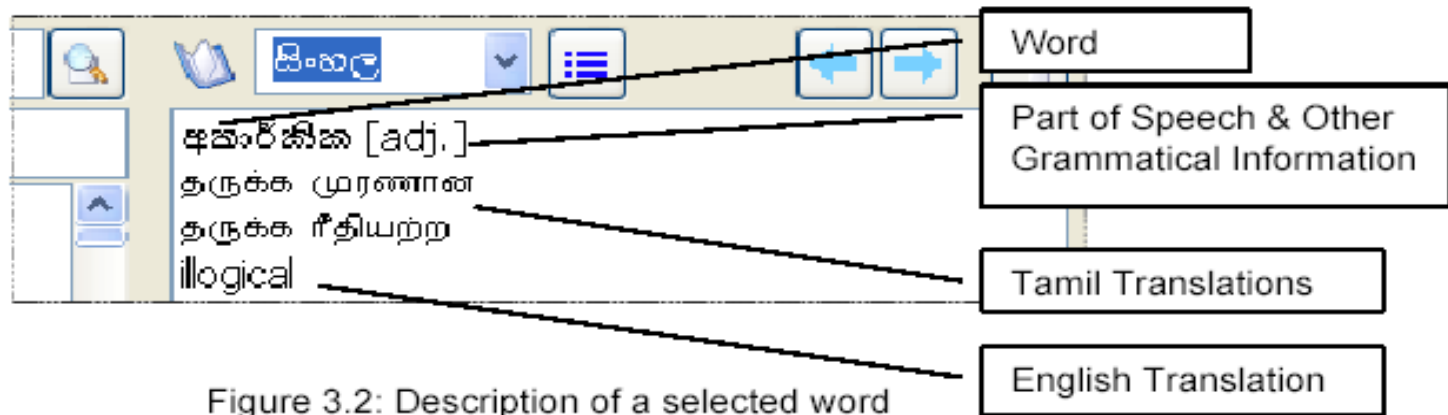
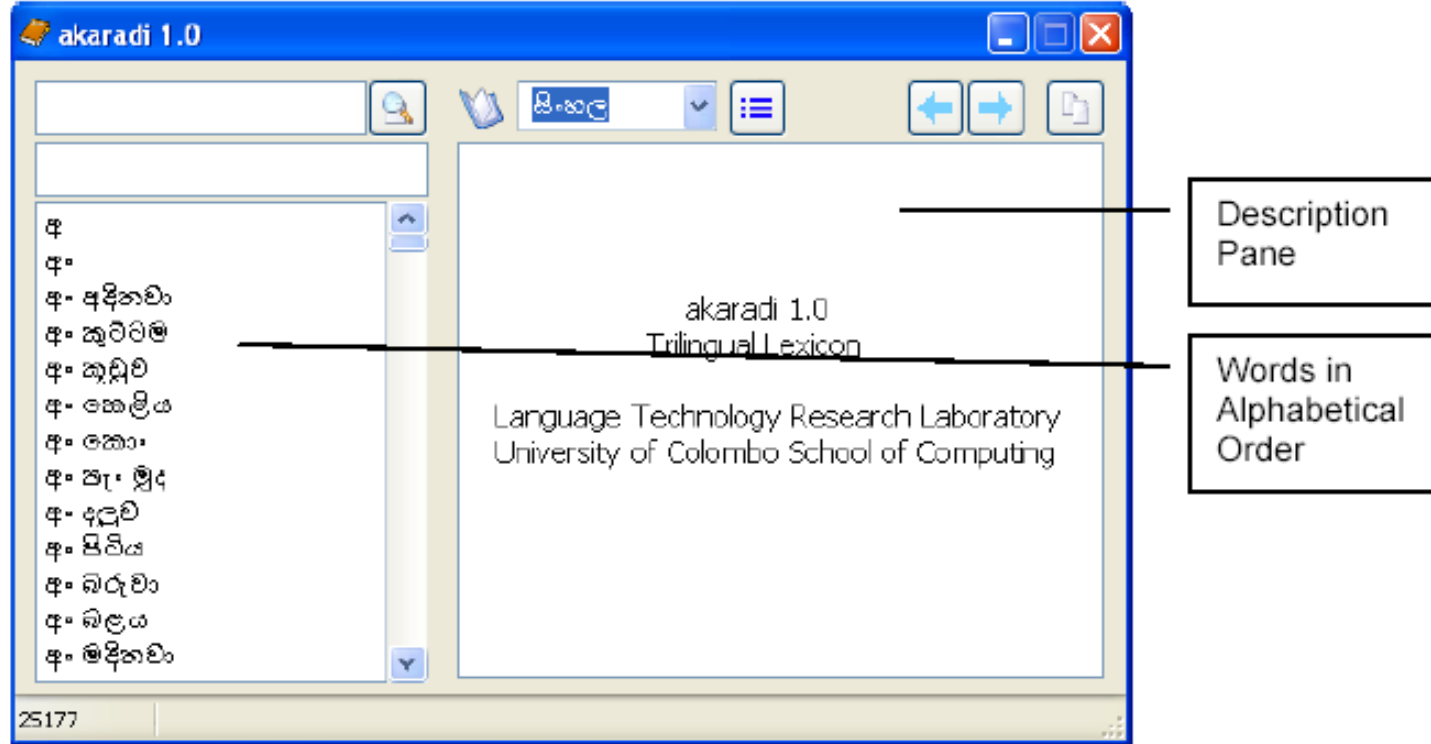
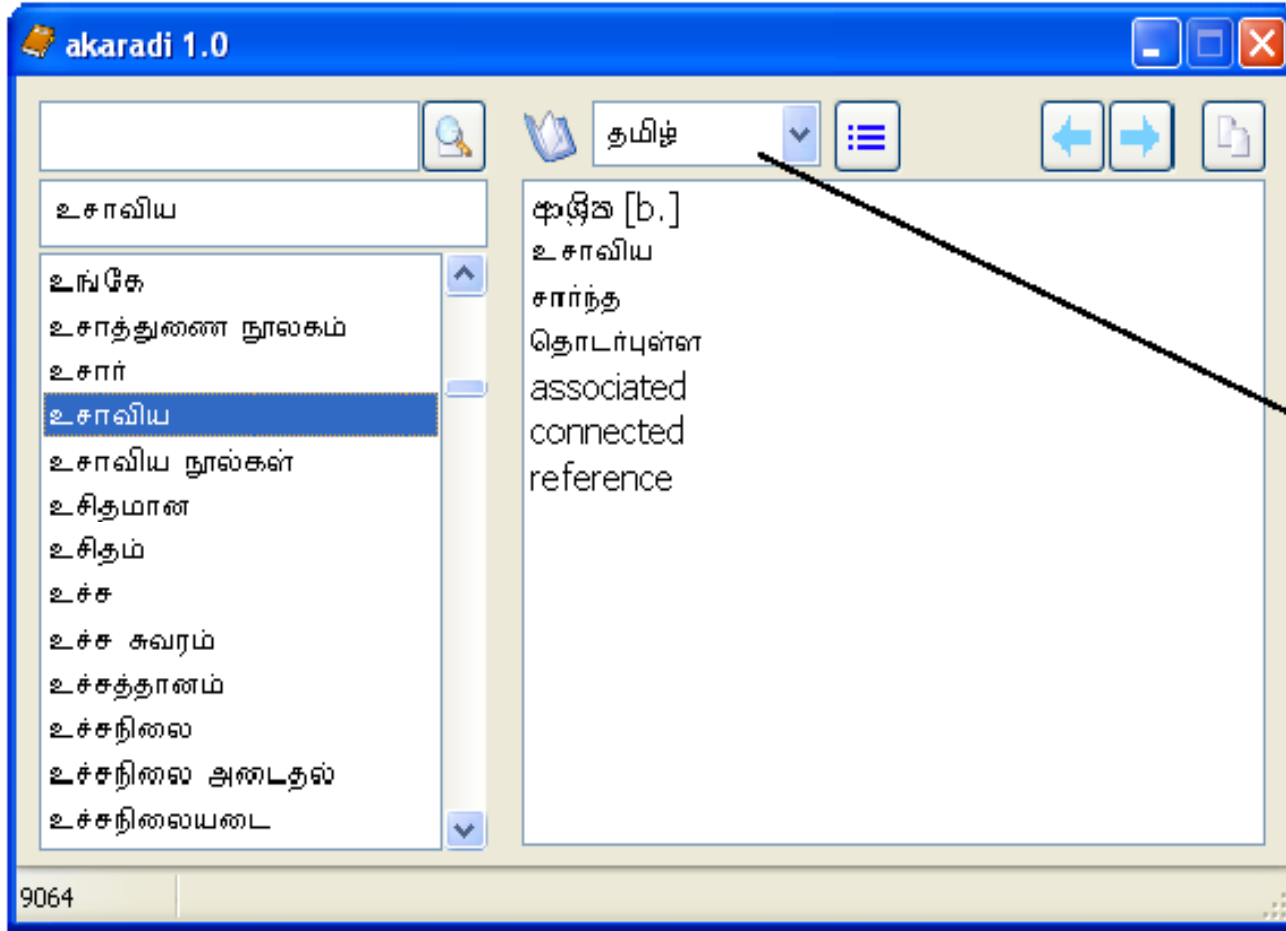


Figure 3.2: Description of a selected word





Lexical Resources in Sinhala

■ Corpus Based Resources (Phase 2)

- Obtained a list of words from the 10 million word corpus (440,000 word forms)
- Classified with respect to grammatical category (13 broad classes)
- Categorized with respect to inflectional / declension paradigms (Nouns: 21 and Verbs: 5)
- Augmented with city/village, person names obtained from the Dept of Census & Statistics
- Used in MS Speller Lexicon (Coverage: 92%)
- Two-level Morphological Analyzer is being built on this resource (with XEROX Finite State Technology)



Sinhala WordNet

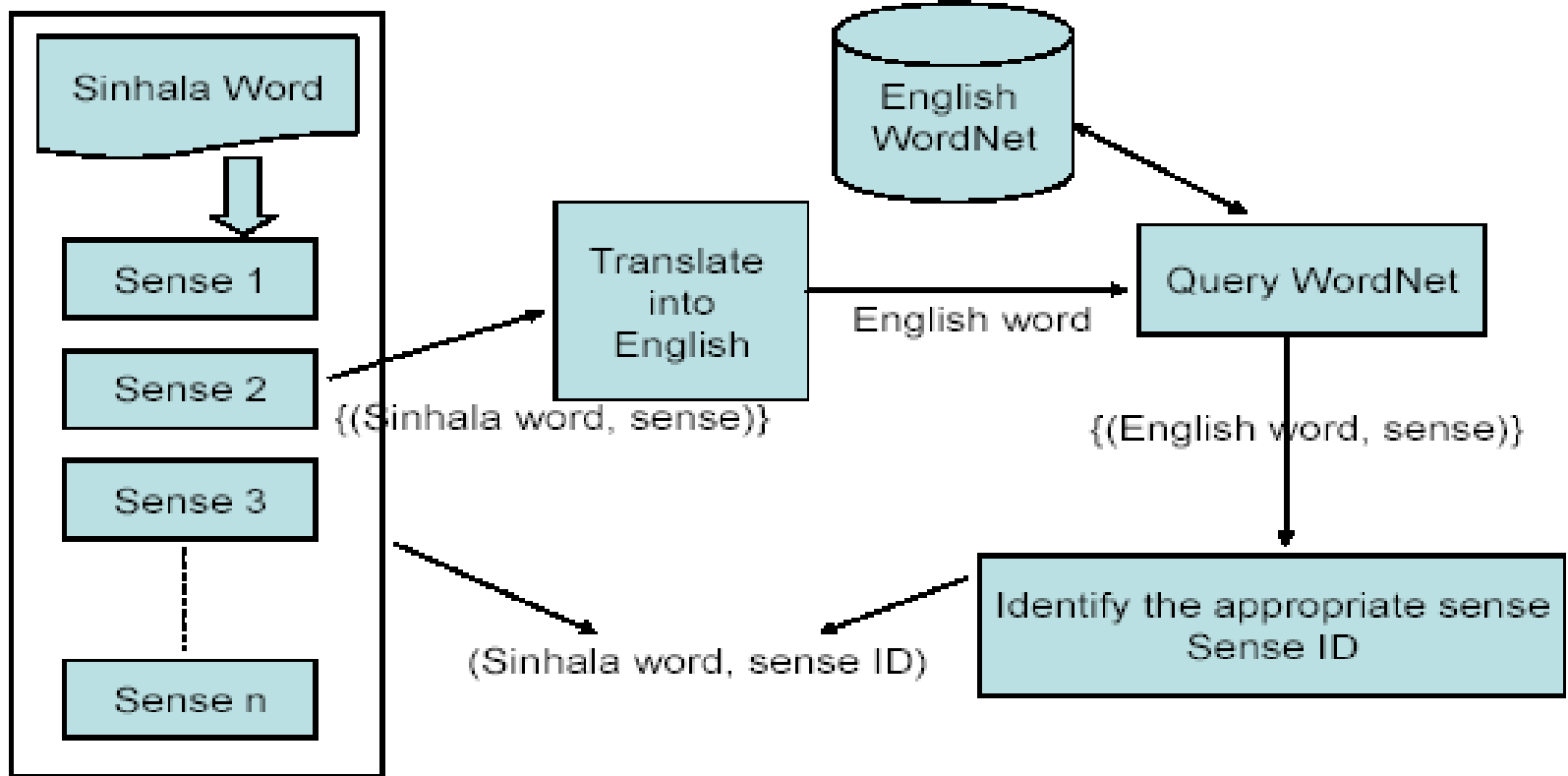
- WordNet is a concept-based lexical resource
- The basic building block is synset (synonym set)
- WordNet defines synsets and relations among synsets
 - Synonym/ Antonym Hypernym / Hyponym, Holonym/Meronym
- Used in many important NLP tasks:
 - Word Sense Disambiguation, Lexical Similarity, Machine Translation, Information Retrieval, Text Summarization



Our Approach

- Select the most frequently occurring 500 words of Sinhala (based on corpus statistics) excluding function words
- Identify the senses of each word in three ways:
 - *Corpus based*: observe the real instances of the word
 - *Dictionary based*: look up a comprehensive dictionary with senses
 - *Expert Knowledge*: consult two senior linguists
- For each sense identified assign the most appropriate sense id from the Princeton English WordNet
- Senses, sense ids, sense relations have been extracted from the lexicographic files of the Princeton WordNet

Work flow





Issues

- Morphological Forms
 - All the morphological forms should be mapped to a lemma:
needs morphological analyzer
- Compound Nouns and Verbs
 - Compound nouns and verbs have meanings different from their constituents
- Language and Culture Specific Senses
 - For language and culture specific senses don't have sense ids in Princeton WordNet
- Sense Assignment
 - Some words have senses which are not present in the current usage of Sinhala



Thank You!