



# Lexical Resources and Wordnet

Kamrul Hayder, Farhana Faruqe and  
Mumit Khan

BRAC University  
Bangladesh

# Introduction

- Two major linguistic resources under development – lexicon and Wordnet
- Annotated lexicon – 80k tagged entries, level-1 POS tags + IPA pronunciation
- Wordnet currently has around 2000 entries, with translated upper ontology and synsets derived from high frequency words

# Lexicon

- Started in Phase-I with just a wordlist used for spelling checker
- Corpus aided effort in Phase-II, resulting in significant changes (many entries removed in fact)
- Completed POS tagging (level-1 and some level-2) for 80k all entries.
- Just completed IPA pronunciation

# Use and status of CRBLP Lexicon

- Used in:
  - *Spelling checker*
  - *TTS*
  - *OCR Post-processor*
  - *POS tagger; this one requires more work as we are currently using a different tagset*
- Serves as a source for the Wordnet effort
- Future lexica would be derived from the Wordnet

# Wordnet design methodology

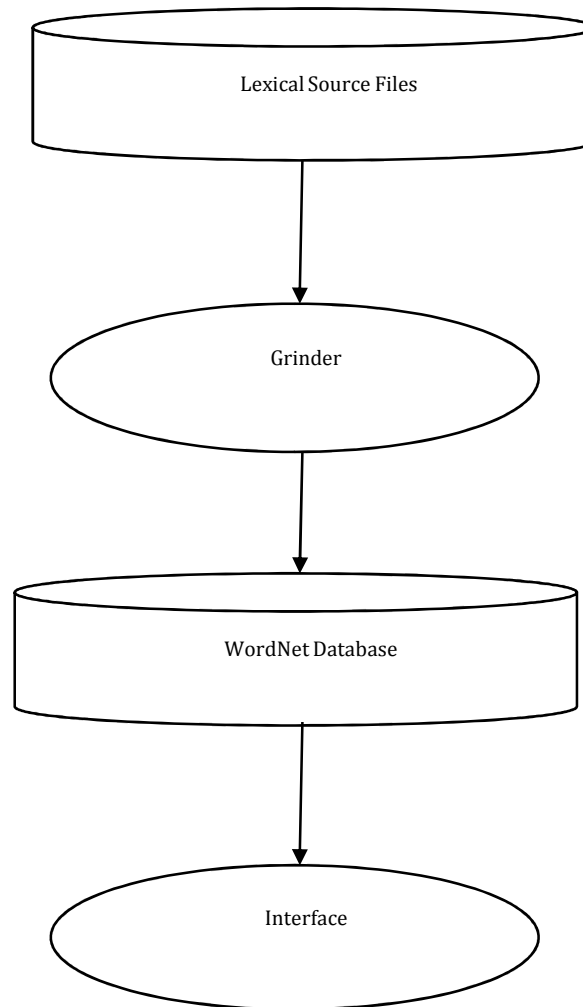
- **Wordnets are even harder than tagsets!**
- Two pronged “bootstrap” approach:
  - Translate the upper level ontology in English Wordnet 3.0, but keeping in mind linguistic differences
  - Start with high frequency words, and incrementally build the synsets
- Bangla Wordnet (BWN) software designed to aid incremental development

# Status of BWN

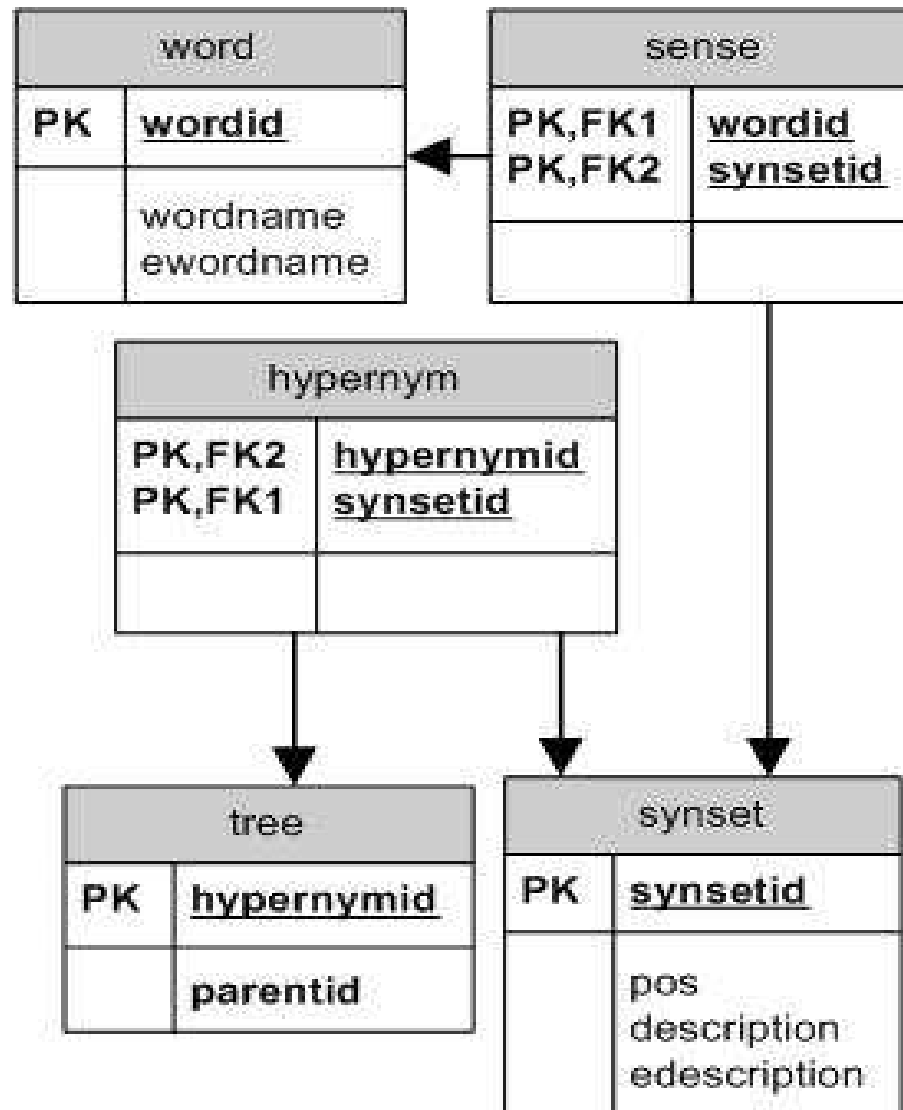
- Current size is approximately 2000 entries, but another 1000 or so projected by end February
- BWN software developed (and published), allows for incremental development (edit and update of existing entries)

[Ref: *BWN - A Software Platform for Developing Bengali WordNet*, Farhana Faruqe and Mumit Khan, *Proc. ICS<sup>2</sup>E*, (2008).]

# BWN Software architecture



# BWN database schema



# Using BWN – sense list

শব্দ অনুসন্ধান:

অংশ

অংশ অনুসন্ধান:

বিশেষ্য

There are 2 senses of অংশ

1. অংশ (সম্পর্ক), অবশিষ্ট, বাকি অংশ, শেষ অংশ -- (কোনোকিছুর সাথে সুদূতভাবে সম্পর্কিত কোনো কিছুর অন্তর্ভুক্ত ভাগ।)
2. অংশ (অবস্থান), একাংশ, একপার্শ্ব -- (কোনো কিছুর বর্ধিত বিস্তৃত-অবস্থান।)

# Using BWN - hypernyms

শব্দ অনুসন্ধান:

অংশ

অংশ অনুসন্ধান:

বিশেষ্য

2 senses of অংশ

Sense 1

অংশ (সম্পর্ক), অবশিষ্ট, বাকি অংশ, শেষ অংশ, -- (কোনো কিছুর সাথে সুদৃঢ়ভাবে সম্পর্কিত কোনো কিছুর অন্তর্ভুক্ত ভাগ।)

=> সম্বন্ধ, সম্পর্ক -- (দুটি সত্তার বা অংশের গুণাবলী দ্বারা সম্পর্কযুক্ত বা অংশভাগী এমন বিমূর্তন সত্তা।)

=> বিমূর্তন -- (সুনির্দিষ্ট উদাহরণাদি থেকে গৃহীত সাধারণ নিদর্শনাদি দ্বারা সৃষ্ট সাধারণ ধারণা।)

=> অমূর্ত-সত্তা, অরূপ-সত্তা, নিরূপ-সত্তা, বিমূর্ত-সত্তা -- (শুদ্ধমাত্র বিমূর্ত (দৈহিক রূপহীন) অস্তিত্ব আছে এমন সত্তা।)

=> সত্তা -- (স্বতন্ত্র অস্তিত্ব আছে এমন ইন্দ্রিয়গ্ধ্য বা জাত বা সিদ্ধান্তকৃত এমন সত্তা কিছু (জীবিত বা জড়)।)

Sense 2

অংশ (অবস্থান), একাংশ, একপার্শ্ব, -- (কোনো কিছুর বর্ধিত বিস্তৃত-অবস্থান।)

=> স্থিতি, অবস্থান -- (কোন সুনির্দিষ্ট পরিসরে অবস্থিত স্থান বা বিস্তৃত স্থান।)

=> দৈহিক লক্ষ্যবস্তু -- (যা স্পর্শ করা যায় এবং দেখা যায় এবং ছায়া প্রদান করে।)

=> দৈহিক সত্তা -- (দৈহিক অস্তিত্ব আছে এমন সত্তা।)

=> সত্তা -- (স্বতন্ত্র অস্তিত্ব আছে এমন ইন্দ্রিয়গ্ধ্য বা জাত বা সিদ্ধান্তকৃত এমন সত্তা কিছু (জীবিত বা জড়)।)

# Using BWN – editing entries

The screenshot shows a software window titled "Edit & Delete Operation". It contains several input fields and buttons. At the top, there is a label "Enter Bangla WordName for Edit:" followed by a text box containing "সত্তা". Below this is a "POS:" label with a dropdown menu showing "বিশেষ্য" and a "Search" button. A "jLabel7" label is positioned below the search section. The main editing area consists of five rows, each with a label and a text box: "Bangla WordName:" with "সত্তা", "English WordName:" with "hypernym", "Description:" with "সত্তা", "English Description:" with "hypernym", and "POS:" with "বিশেষ্য". At the bottom of the window are two buttons: "Save" and "Delete".

# Summary

- Corpus-derived lexicon annotated with POS and IPA pronunciation is a substantial resource
- Bootstrapping approach has proven effective
- BWN software has been a worthwhile effort
- Wordnet development should be done on target
- Linguistic review may be difficult because of the extensive lexical semantic knowledge required

# Future work

- Morphological tagging of lexicon
- Thorough linguistic review of Wordnet after 2500 entries
- Compare with Hindi Wordnet (IIT-Bombay)
- Collaborate with MultiWordnet teams
- *Wordnets may be difficult to develop, but certainly worth the effort!*