

DEVELOPING BANGLA TEXT, SPEECH AND SCRIPT RESOURCES: CURRENT STATUS AND CHALLENGES AHEAD

S. M. Murtoza Habib

Center for Research on Bangla Language Processing

BRAC University

Dhaka, Bangladesh

MISSIONS

- Develop
 - Text resources
 - speech resources
 - script language resources
- Freely available
- Enhance these resources
- Develop NLP applications



TEXT RESOURCES: **CRBLP PROTHOM-ALO CORPUS**

- First challenge
 - Find digital data
- News corpus (Prothom-Alo)
- Technical challenges
 - converting the non-Unicode data
 - cleaning the corpus
- Corpus contains
 - 18,067,470 tokens
 - 386,639 tokens types



TEXT RESOURCES: ENGLISH-BANGLA PARALLEL CORPUS

- Brown Corpus
 - 100k words
- Primary challenge
 - Finding competent translators
- Current status
 - 10k words are translated



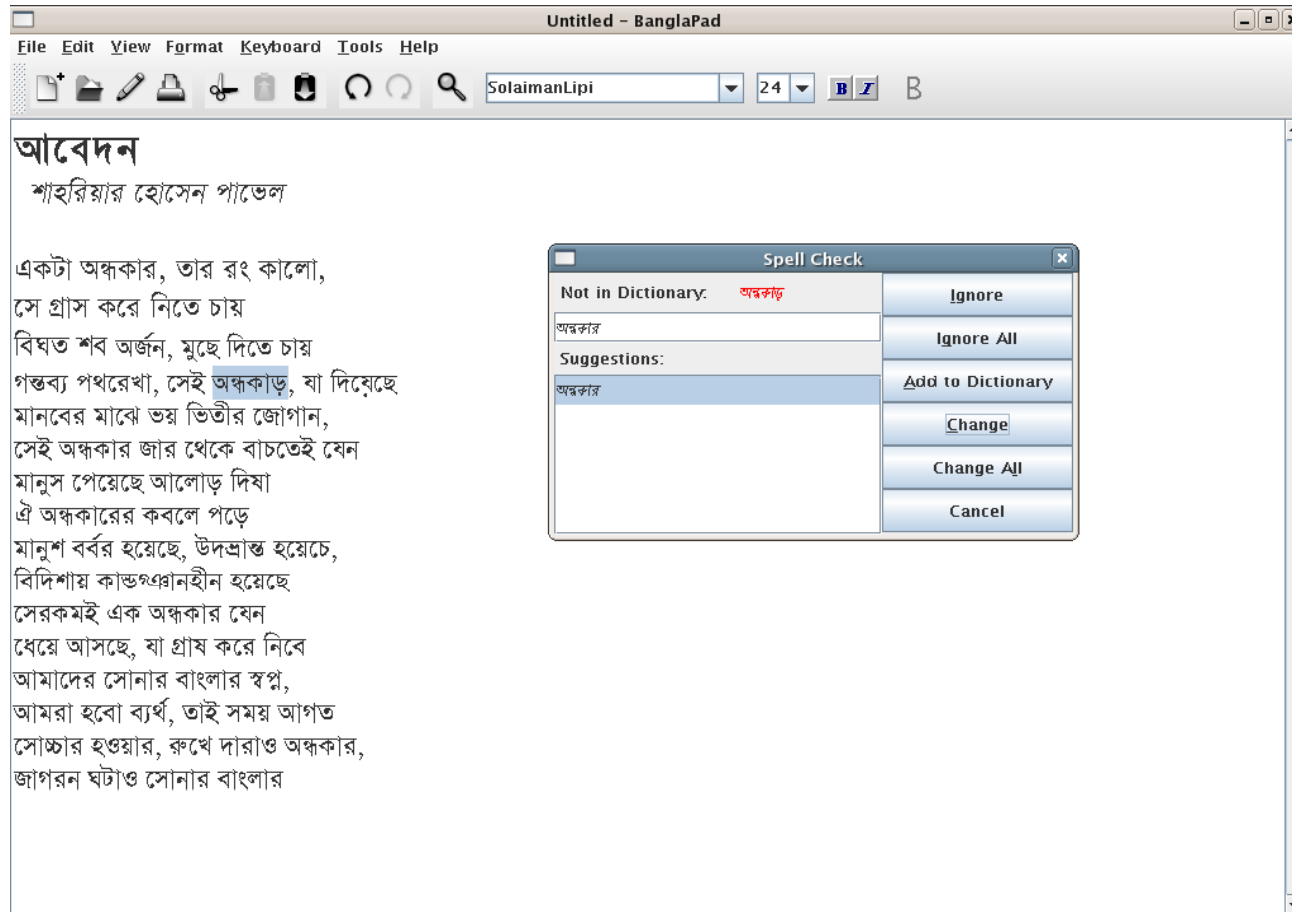
TEXT RESOURCES: DIGITAL LEXICON

- Based on the CRBLP Prothom-Alo corpus
- Annotated with pronunciation (IPA)
- Uses
 - CRBLP Bangla speller
 - Pronunciation for TTS
 - Digital dictionary with pronunciation



TEXT RESOURCES: DIGITAL LEXICON

○ Spellchecker on BanglaPad



TEXT RESOURCES: BANGLA WORDNET

- Just started
 - Noun category
 - Hypernymy relation
- Status
 - Word (synonyms): 1650
 - Synset (unique): 850
- Pushing the limits of our linguistic capacity



TEXT RESOURCES: BANGLA WORDNET

WordNet 1.0 Browser

কাইল সহায়িকা

শব্দ অনুসন্ধান:

পাখি অনুসন্ধান:

There are 5 senses of পাখি

- গগনগতি, ঋণ, ঋণম, খেচর, চিড়িয়া, নতুকা, পখি, পখী, পক্ষধর, পক্ষ্যু, পক্ষী, পতঙ্গ, পততি, পাক, পত্নরথ, পত্নী, পাখি, বিহগ, বিহস, বিহসম, পাখি (পাখি) -- (পালকানুত এবং অগ্র-পদাঙ্গ পাখ্যে নৃপাধারিত হযেছে এমন উচ্চরক্ত বিশিষ্ট ডিম প্রসবকারী মেনুদণ্ডী প্রাণী।)
- জমির একক বিশেষ, ৩০ কানি ভূমি, ২৬/৩৩/৩৫ শতাংশ, পাখি, অঞ্চল একক -- (জমির পরিমাপ পদ্ধতির জন্য ব্যবহৃত একক।)
- চক্র-অবলম্বক, স্পোক (spoke), পাখি (চক্র) -- (চক্রকেন্দ্র এবং চক্রের পরিধির মধ্যবর্তী অক্ষপ্রসারী সংযোজক দ্বারা সৃষ্ট অবলম্বক।)
- পাখি (আড়াকার্ত) -- (মইয়ের ধাপ হিসাবে ব্যবহৃত আড়াকার্ত।)
- পাখি (তনুগী) -- (অববসনী নারী (নৃপকার্শ)।)

WordNet 1.0 Browser

কাইল সহায়িকা

শব্দ অনুসন্ধান:

পাখি অনুসন্ধান:

5 senses of পাখি

Sense 1
 গগনগতি, ঋণ, ঋণম, খেচর, চিড়িয়া, নতুকা, পখি, পখী, পক্ষধর, পক্ষ্যু, পক্ষী, পতঙ্গ, পততি, পাক, পত্নরথ, পত্নী, পাখি, বিহগ, বিহস, বিহসম, পাখি
 => মেনুদণ্ডী -- (বাদের মেনুদণ্ড- খণ্ড অংশ বা তনুগাখি দিয়ে তৈরি এবং যা কলোটিতে যুক্ত থাকে।)
 => কটেট -- (দেহকাঠামোতে নটোকর্ক বা মেনুদণ্ড আছে, এমন কর্ভাটা পর্বের যে কোন প্রাণী।)
 => প্রাণী -- (জীবজগতের সদস্য, যারা শ্বাস-প্রশ্বাস চলাচল করতে পারে।)
 => জীবসত্তা -- (আল-উন্নানের সামার্থ্য আছে বা স্বাধীনভাবে কার্যক্রম সম্পন্ন করতে পারে এমন সত্তা।)
 => জীবন্ত বস্তু -- (জীবন্ত বা এক সমস্ত জীবিত ছিল।)
 => শৈবিক পক্ষ্যবস্তু -- (যা স্পর্শ করা যাবে এবং দেখা যাবে এবং ছোয়া প্রদান করে।)
 => ইন্দ্রিয়গোচ্য সত্তা, শৈবিক সত্তা -- (শৈবিক অস্তিত্ব আছে এমন সত্তা।)
 => অস্তিত্ব, সত্তা -- (স্বতন্ত্র অস্তিত্ব আছে এমন ইন্দ্রিয়গোচ্য বা জ্ঞাত বা সিদ্ধান্তকৃত এমন সত্তা কিছু (জীবিত বা জড়)।)

Sense 2
 জমির একক বিশেষ, ৩০ কানি ভূমি, ২৬/৩৩/৩৫ শতাংশ, পাখি, অঞ্চল একক -- (জমির পরিমাপ পদ্ধতির জন্য ব্যবহৃত একক।)
 => পরিমাপ একক (নির্মাপ) -- (প্রমিত পরিমাপ বা বিনিময় মান অনুসারে গৃহীত যে কোনো পরিমাপের বিভাজন একক (নির্মাপ)।)
 => সুনির্দিষ্ট পরিমাপ -- (কোনো কিছু পরিমাপের জন্য সুনির্দিষ্ট পরিমাপ।)
 => পরিমাপ -- (কোন কিছু কি পরিমাপ আছে, তা পরিমাপ করতে পারে।)
 => বিমূর্তন -- (সুনির্দিষ্ট উদাহরণাদি থেকে গৃহীত সাধারণ নির্দেশাদি দ্বারা সৃষ্ট সাধারণ ধারণা।)
 => অমূর্ত-সত্তা, অনূর্ণ-সত্তা, নিরূর্ণ-সত্তা, বিমূর্ত-সত্তা -- (যুধুমতে বিমূর্ত (শৈবিক নৃণহীন) অস্তিত্ব আছে এমন সত্তা।)
 => অস্তিত্ব, সত্তা -- (স্বতন্ত্র অস্তিত্ব আছে এমন ইন্দ্রিয়গোচ্য বা জ্ঞাত বা সিদ্ধান্তকৃত এমন সত্তা কিছু (জীবিত বা জড়)।)

Sense 3
 চক্র-অবলম্বক, স্পোক (spoke), পাখি (চক্র) -- (চক্রকেন্দ্র এবং চক্রের পরিধির মধ্যবর্তী অক্ষপ্রসারী সংযোজক দ্বারা সৃষ্ট অবলম্বক।)
 => অবলম্বক -- (অন্য কিছুর ভার ধারণ করে এমন কোনো উদ্ভাগ।)
 => ডিভাইস -- (একটি সুনির্দিষ্ট উদ্দেশ্য সাধনের জন্য যান্ত্রিক উপায়ে উদ্ভাবিত।)
 => যান্ত্রিক-উপায়ে কৃত -- (একটি মানবসৃষ্ট (অথবা মানবসৃষ্টির পদ্ধতি), যা যান্ত্রিক উপায়ে সম্পন্ন হয়েছে।)
 => মানবসৃষ্টি -- (মানুষের সৃষ্ট একক (নির্মাপ) সময় পক্ষ্যবস্তু।)
 => সময় -- (উপকরণাদি দ্বারা সংযোজিত অবস্থায় যা একক (নির্মাপ) সত্তা হিসাবে পরিচিত।)
 => শৈবিক পক্ষ্যবস্তু -- (যা স্পর্শ করা যাবে এবং দেখা যাবে এবং ছোয়া প্রদান করে।)
 => ইন্দ্রিয়গোচ্য সত্তা, শৈবিক সত্তা -- (শৈবিক অস্তিত্ব আছে এমন সত্তা।)

SPEECH RESOURCES: CONTINUOUS SPEECH CORPUS

- Continuous Read Speech Corpus
 - 13 hours 33 minutes
 - 106,860 words
 - Wide range of domains



SPEECH RESOURCES

- Speech Corpus for Acoustic Analysis for Bangla phoneme inventory
 - 30 consonants
 - 14 vowels (monophthong including nasal)
 - 21 diphthongs
- Speech Corpus for Developing Diphone
 - 4355 diphones



SCRIPT CORPUS

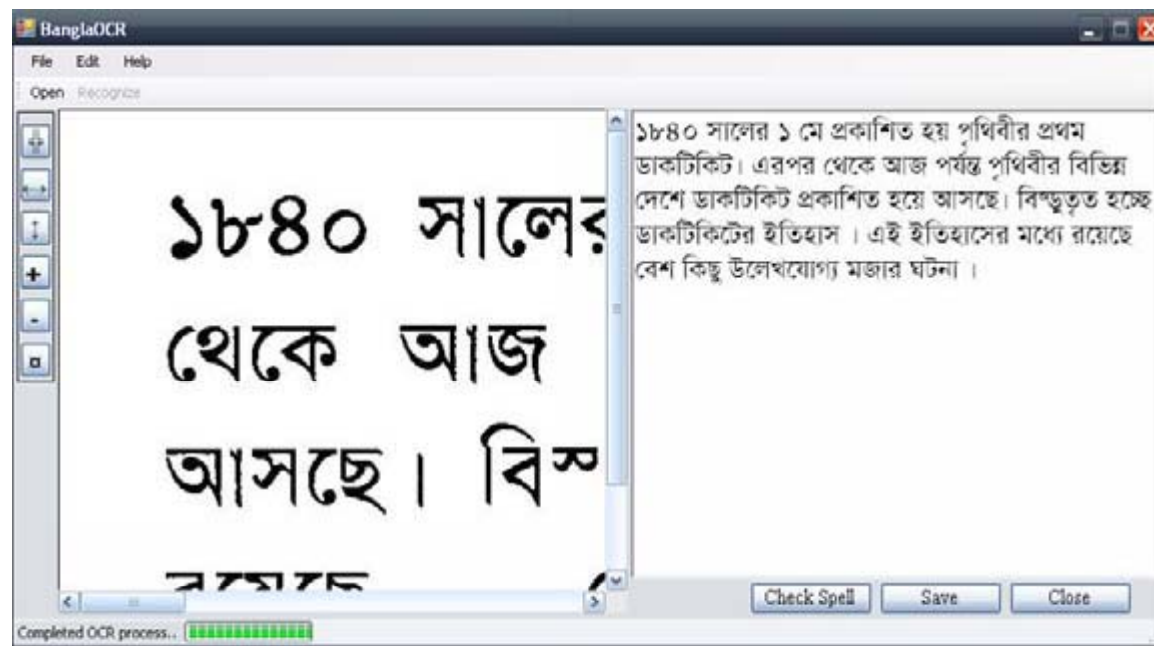
- Framework for our OCR
 - Training
 - 3200 data units
 - Testing

ছ ছা ছি ছী চু চু চ্ছে ছো
চ্ছ চ্ছা চ্ছি চ্ছী চ্ছু চ্ছু চ্ছে চ্ছো
চ্ছ ছা চ্ছি চ্ছী চ্ছু চ্ছু চ্ছে চ্ছো
চঞ চঞা চিঞ চিঞী চঞু চঞু চেঞ চেঞা
চ্য চ্যা চি চী চ্য চ্য চে চ্যো
ছ ছা ছি ছী ছু ছু ছে ছো
জ জা জি জী জু জু জে জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো
জ্জ জ্জা জ্জি জ্জী জ্জু জ্জু জ্জে জ্জো



SCRIPT CORPUS

- BanglaOCR released
 - Based on Google's Tesseract



THANK YOU

