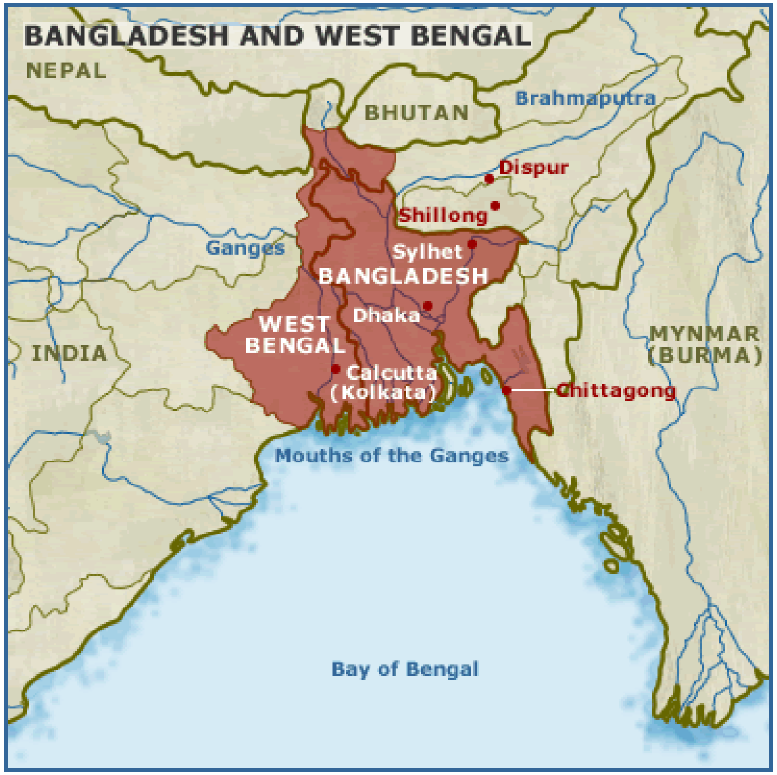
বাংলা

Bangla, a.k.a. Bengali

“Language in 10”

Andrew Wilkinson

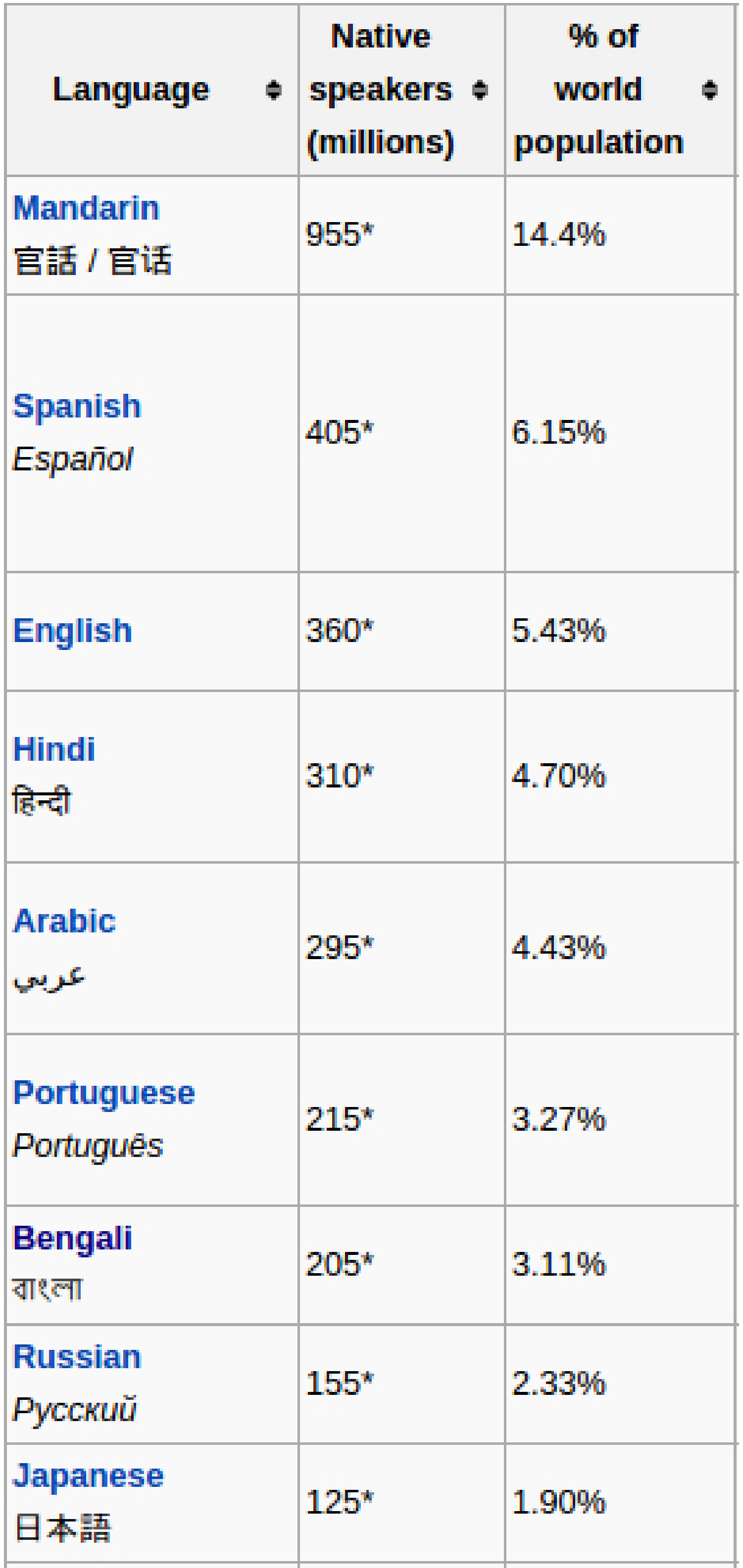


Demographics

- Bangla (IPA: /'baŋ.la/) is spoken primarily in Bangladesh and in the Indian state of West Bengal, which borders Bangladesh on the west
- Population of Bangladesh ≈ 157 million
- Population of West Bengal ≈ 92 million

Number of speakers

- Wikipedia's source, the *Nationalencyklopedin* of Sweden, gives the number of native speakers as 205 million
  - This puts Bangla at the seventh most widely spoken language in the world by number of native speakers
- Ethnologue is slightly more conservative, at 193 million
- There are a number of closely related languages spoken in Bangladesh, of varying degrees of mutual intelligibility with Bangla, such as Chittagonian and Sylheti
  - Often considered dialects



Language	Native speakers (millions)	% of world population
Mandarin 官話 / 官话	955*	14.4%
Spanish *Español*	405*	6.15%
English	360*	5.43%
Hindi हिन्दी	310*	4.70%
Arabic عربي	295*	4.43%
Portuguese *Português*	215*	3.27%
Bengali বাংলা	205*	3.11%
Russian *Русский*	155*	2.33%
Japanese 日本語	125*	1.90%

Writing system

Bangla is written with the Bengali script/alphabet (technically, an abugida). There is no capital/lowercase distinction.

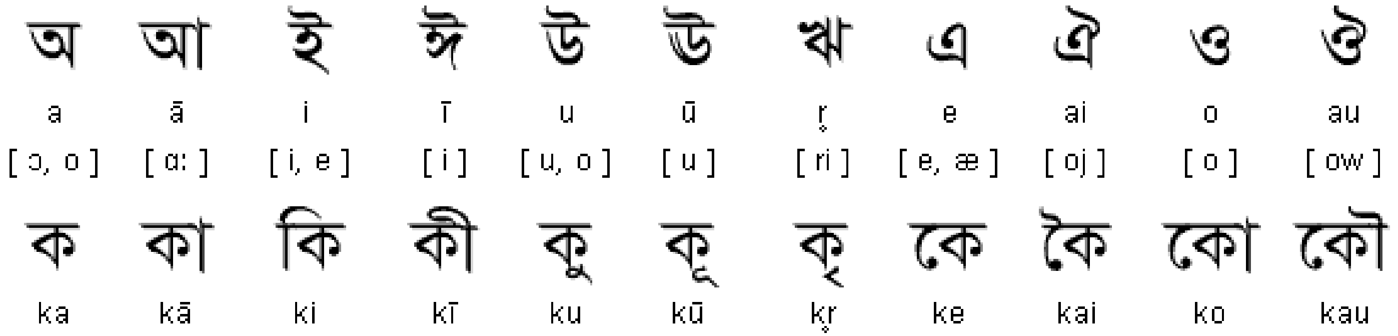
Vowels that occur syllable-initally are written with a vowel character; vowels that follow consonants in a syllable are written with a vowel diacritic or in certain cases are unwritten (“inherent vowels,” /ɔ/)

ক কা কি কী কু কূ কৃ কে কৈ কো কৌ

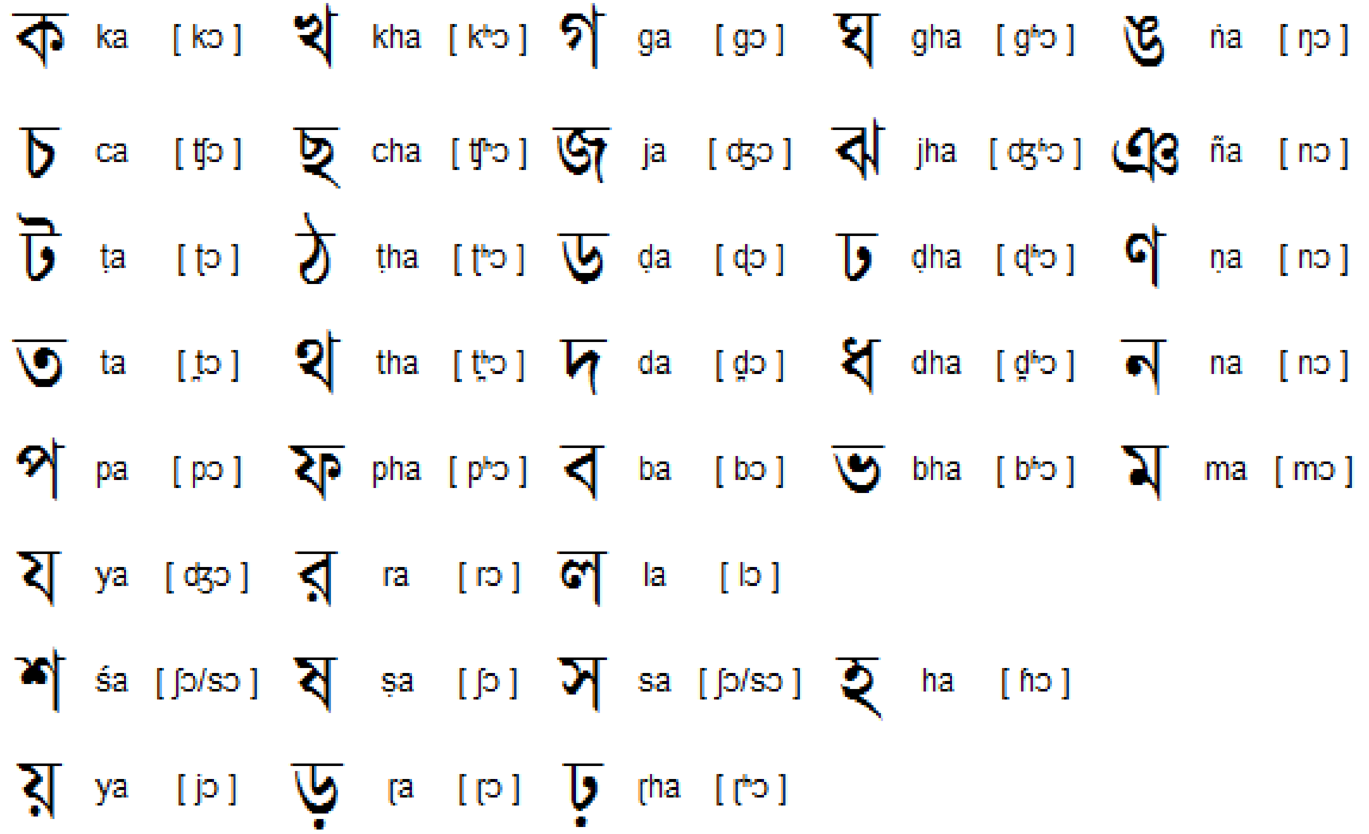
The consonant ক (kô) along with the diacritic form of the vowels অ, আ, ই, ঈ, উ, ঊ, ঋ, এ, ঐ, ও and ঔ.

The alphabet is reasonably phonemic, but without a 1-to-1 correspondence. Some sounds can be represented by more than one letter, and some letters can represent more than one sound, depending on environment. This is due largely to historical development of the language and language change.

Writing system (cont'd.)



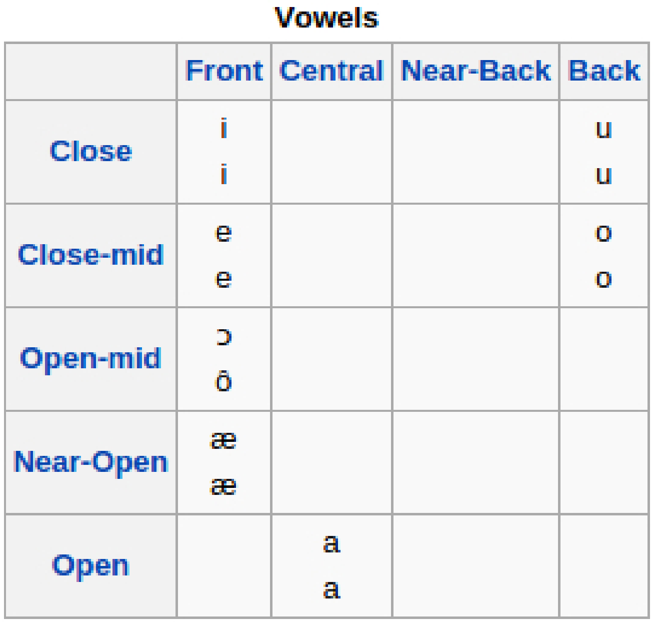
অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
a	ā	i	ī	u	ū	ṛ	e	ai	o	au
[ɔ, o]	[ɑː]	[i, e]	[i]	[u, o]	[u]	[ri]	[e, æ]	[oj]	[o]	[ow]
ক	কা	কি	কী	কু	কূ	কৃ	কে	কৈ	কো	কৌ
ka	kā	ki	kī	ku	kū	kṛ	ke	kai	ko	kau



ক	ka	[kɔ]	খ	kha	[kʰɔ]	গ	ga	[gɔ]	ঘ	gha	[gʰɔ]	ঙ	ṅa	[ŋɔ]
চ	ca	[ʧɔ]	ছ	cha	[ʧʰɔ]	জ	ja	[ʤɔ]	ঝ	jha	[ʤʰɔ]	ঞ	ña	[nɔ]
ট	ṭa	[ʈɔ]	ঠ	ṭha	[ʈʰɔ]	ড	ḍa	[ɖɔ]	ঢ	ḍha	[ɖʰɔ]	ণ	ṇa	[nɔ]
ত	ta	[t̪ɔ]	থ	tha	[t̪ʰɔ]	দ	da	[d̪ɔ]	ধ	dha	[d̪ʰɔ]	ন	na	[nɔ]
প	pa	[pɔ]	ফ	pha	[pʰɔ]	ব	ba	[bɔ]	ভ	bha	[bʰɔ]	ম	ma	[mɔ]
য	ya	[ʤɔ]	র	ra	[rɔ]	ল	la	[lɔ]						
শ	śa	[ʃɔ/sɔ]	ষ	ṣa	[ʃɔ]	স	sa	[ʃɔ/sɔ]	হ	ha	[ɦɔ]			
য়	ya	[jɔ]	ড়	ṛa	[ɽɔ]	ঢ়	ṛha	[ɽʰɔ]						

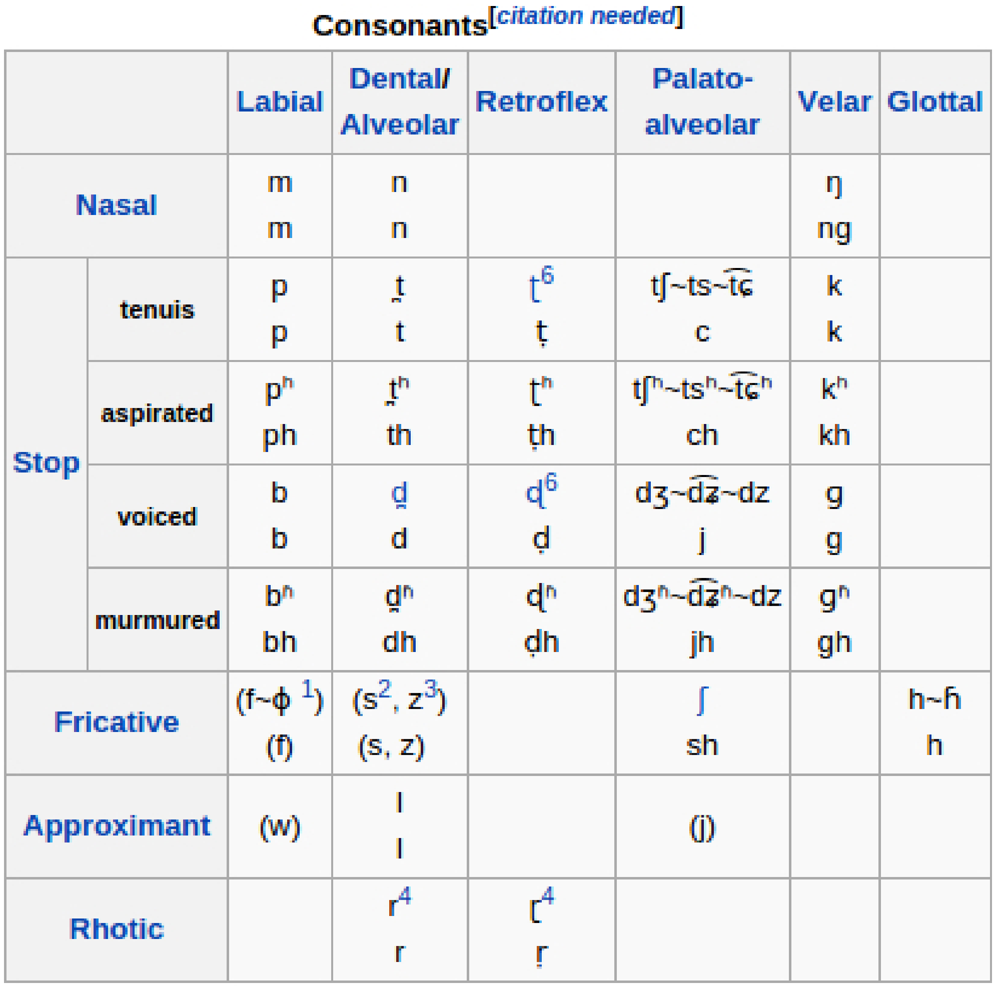
Phonology

- 29 consonant phonemes, 7-8 vowel phonemes, 15+ diphthongs



Vowels

	Front	Central	Near-Back	Back
Close	i i			u u
Close-mid	e e			o o
Open-mid	ɔ ô			
Near-Open	æ æ			
Open		a a		



Consonants[*citation needed*]

		Labial	Dental/ Alveolar	Retroflex	Palato-alveolar	Velar	Glottal
Nasal		m m	n n			ŋ ng	
Stop	tenuis	p p	t̪ t	ʈ[6] ṭ	tʃ~ts~t͡ɕ c	k k	
	aspirated	pʰ ph	t̪ʰ th	ʈʰ ṭh	tʃʰ~tsʰ~t͡ɕʰ ch	kʰ kh	
	voiced	b b	d̪ d	ɖ[6] ḍ	dʒ~d͡ʑ~dz j	g g	
	murmured	bʱ bh	d̪ʱ dh	ɖʱ ḍh	dʒʱ~d͡ʑʱ~dz jh	gʱ gh	
Fricative		(f~ɸ [1]) (f)	(s[2], z[3]) (s, z)		ʃ sh		h~ɦ h
Approximant		(w)	l l		(j)		
Rhotic			ɾ[4] r	ɽ[4] ṛ			

In keeping with a trend...

- It's the beginning of the Universal Declaration of Human Rights, read aloud!
- সমস্ত মানুষ স্বাধীনভাবে সমান মর্যাদা এবং অধিকার নিয়ে জ্নমগ্রহণ করে | তাঁদের বিবেক এবং বুদ্ধি আছে সুতরাং সকলেরই একে অপরের প্রতি ভ্রাতৃত্বসুলভ মনোভাব নিয়ে আচরণ করা উচিত্ |
- Shomosto manush shadhinbhabe shoman morjada ebong odhikar niye jonmogrohon kore. Tader bibek ebong buddhi achhe; shutorang shokoleri ekey oporer proti bhratrittoshulobh monobhab niye achoron kora uchit.
- All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

  (Article 1 of the Universal Declaration of Human Rights)

Typology/Grammar

- Bangla is an Indo-European language
  (→Indo-Iranian→Indo-Aryan→Eastern Zone→Bengali-Assamese)
  - Related to Hindi, Punjabi, Marathi, Gujarathi, etc. Descendent of Sanskrit.

- Word order is SOV, in general

- Pronouns are marked for number (singular and plural), person (1st, 2nd, 3rd), formality
  (very familiar, familiar, polite), proximity (here, there, elsewhere),
  and case (nominative, objective, possessive); but NOT gender

- Nouns are marked for case (nominative, objective, possessive, locative)

- Uses postpositions; definite articles also come after the nouns they modify
  - Number is not marked on nouns but is present in the definite article, when used

  - Ex. জুতা juta 'shoe/s' → জুতাটা juta-ṭa 'the shoe' → জুতোগুলো juta-gulo 'the shoes'
  - জুতার আগে juta-r age 'before a/the shoe' জুতার থেকে juta-r theke 'from a/the shoe' etc.

Typology/Grammar Cont'd.

- Verbs have moderately complex, mostly regular conjugations involving:
  - Four simple tenses: present, past, habitual past, future
  - Two moods: indicative, imperative
  - Three aspects: simple, progessive, perfect
  - Person and formality, but NOT number
    - 2p polite and 3p polite are the same
- Zero copula; i.e., no one word for “to be”
  - তিনি শিক্ষক tini shikkhok 'he/she is a teacher', lit. 'he/she teacher'

State of Bangla MT

- Despite being spoken by hundreds of millions of people, Bangla is “low-density,” or “resource-scarce”
  - Very little Bangla-English bitext exists in corpus form; monolingual corpora are relatively small compared to other widely-spoken languages
  - This is the biggest hurdle when it comes to developing SMT systems
  - The LDC offers 11,226 Bangla-English sentence pairs
  - The Center for Research on Bangla Language Processing, BRAC University, Bangladesh, offers a one million-plus–sentence monolingual dataset taken from the Prothom Alo newspaper, consisting of all their output from 2005
  - Post et al. (2012) use Amazon Turk to build a bitext corpus of about 20,000 sentences
- Not much useful, non-trivial research seems to have been done
  - I came across a number of rudimentary papers by computer scientists in Bangladesh and India
  - Most of this work focuses on very basic rule-based transformations only barely capable of translating simple sentences, reliant on dictionaries and verb conjugation tables
  - Evaluation scores often unreported

Google Translate

- GT came out with Bangla-English MT in 2011. As far as I can tell, it's the only decent publicly available system. Achieved BLEU score of 20.01 on Post et al.'s dataset
  - (Anubadok, a free MT system that comes up when you search for Bangla MT, achieved a BLEU score of 1.60, NIST score of 1.46, and TER score of 1.03 when evaluated by Islam, Tiedemann & Eisele (2010))

    In preparation for the storm, Amtrak said it plans to adjust its schedule based on the weather.

    ঝড় জন্য প্রস্তুতি, কিছুই এটি আবহাওয়ার উপর ভিত্তি করে তার সময়সূচি সমন্বয় করার পরিকল্পনা করছে.

    In preparation for the storm, it is based on the weather, plans to adjust its schedule.

    The debilitating long-term effects of heartworm pills on puppies who abused them in their playing days are unfortunately only beginning to be understood.

    তাদের খেলার দিন তাদের নির্যাতিত যারা একপাল উপর heartworm ঔষধ debilitating দীর্ঘমেয়াদী প্রভাব দুর্ভাগ্যবশত শুধুমাত্র বুঝতে হবে শুরু হয়

    Heartworm medication on those who abused them in their playing days puppies debilitating long-term effects are, unfortunately, only to be understood

Useful papers

- Most useful publications:

  Approaches to handle scarce resources for Bengali Statistical Machine Translation, Maxim Roy (2010) (Ph.D. thesis)

  http://summit.sfu.ca/system/files/iritems1/10031/etd5938.pdf

  - Uses hybrid of rules and SMT, including prepositions module, word reordering, semi-supervised learning

  English to Bangla Phrase-Based Machine Translation, Islam & Eisele (2009) (Master's thesis)

  http://www.lct-master.org/getfile.php?id=124&n=1&dt=TH&ft=pdf&type=TH

  A Hybrid Approach for Bengali to Hindi Machine Translation, Chatterji et al. (2009)

  https://www.academia.edu/attachments/31119595/download_file?st=MTQyMjkyNjIzOCwyNC4xMzEuMjU1LjI0NCwyNTU2MDkzOQ%3D%3D&s=swp-preview-selector-dropdown

Useful papers cont'd.

Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing, Post, Callison-Burch & Osborne (2012)

- Compare two different types of grammar, SAMT and Hiero

http://www.aclweb.org/anthology/W12-3152

Combining Bilingual and Comparable Corpora for Low Resource Machine Translation, Irvine & Callison-Burch (2013)

http://www.cs.jhu.edu/~anni/papers/irvineCCB_WMT13.pdf

- Focus on Bangla in the context of useful approaches across low-resource languages (e.g. bilingual lexicon induction); use corpus from above source; achieve BLEU score of 12.7 using Moses

Word Alignment-Based Reordering of Source Chunks in PB-SMT, Pal, Naskar & Bandyopadhyay (2014)

http://www.mt-archive.info/10/LREC-2014-Pal.pdf

Improving MT System Using Extracted Parallel Fragments of Text from Comparable Corpora, Gupta, Pal & Bandyopadhyay (2013)

http://www.mt-archive.info/10/BUCC-2013-Gupta.pdf