Common Bengali Handwritten Graphemes in Context

Introductory Booklet

Grapheme

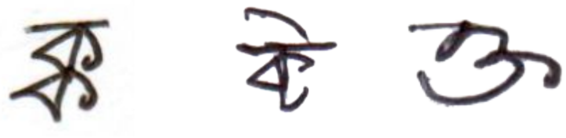
A **grapheme** is the smallest unit of a writing system for any language. The definition of grapheme for English is tied to how phonemes are defined/spoken (see Appendix A for basic alphabets in Bengali).

Bengali graphemes fall in the *Abugida* family which makes it different from english in three ways:

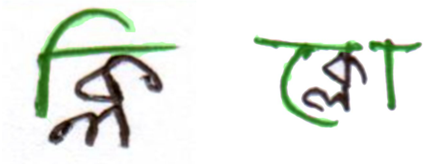
1. Commonality of consonant conjuncts (the same glyph containing two consonants হ্ম = হ (ha) + ম (ma))
2. Use of *diacritics to create consonant+vowel compounds.
3. Allographs (different ways of writing the same grapheme)

*Diacritics are glyphs added to a letter that changes its sound-values, i.e. **á** (greek), কি (bengali)

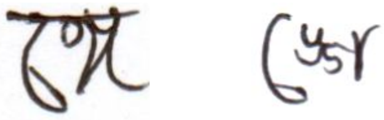
Consonant Conjuncts with ক (ka)



Vowel Diacritics with Root ক্ল (kla)



Allographs (dissimilar glyphs of the same grapheme)

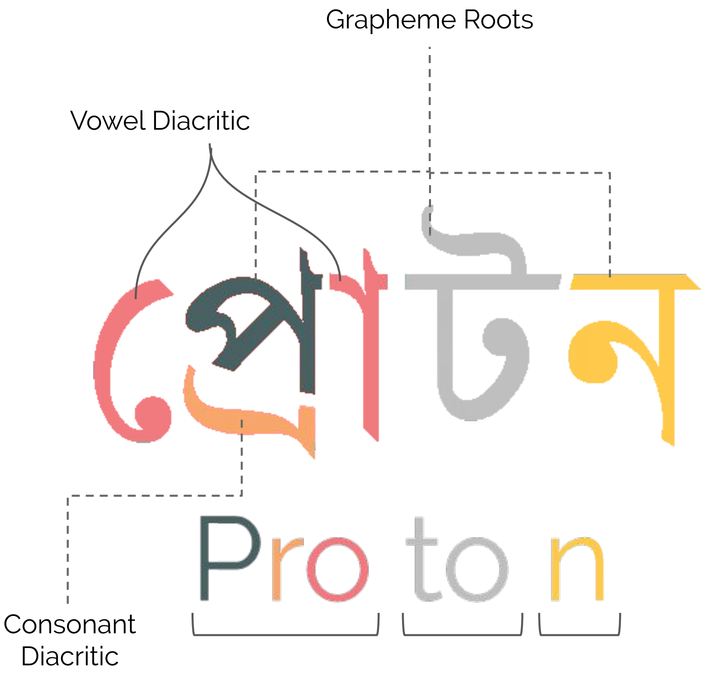


Why Grapheme?

Consonant/Vowels in bengali words are not **isolated** and **arranged horizontally** like English alphabets. Bangla orthographic words, are a combination of the following kinds of graphemes [1] :

1. Independent form of *vowel graphemes*, i.e. আ (a) , ই (e)
2. *Consonant graphemes* with or without a vowel diacritic attached to them, i.e. কো (k-o), ক (k-a)
3. *Consonant conjuncts* behaving as individual consonants i.e ক্লো (kl-o), ক্ল (kl-a)
4. *Consonant diacritic symbols* indicating nasalization of vowels, and suppression of the inherent vowels, i.e. চাঁ (ch'a)

A benchmarking dataset of isolated graphemes is therefore a stepping stone for Bengali OCR



* Orthographic components are isolated in the image for representation

[1] A study of readability of texts in Bangla through machine learning approaches (2016)

Why Common Graphemes in Context?

Bangla graphemes are a combination of the following units:

1. **Grapheme roots** which can be vowels, consonants or conjuncts
2. **Vowel Diacritics**, i.e. ◌া, ◌ি, ◌ী, ◌ু, ◌ূ, ◌ৃ, ◌ে, ◌ৈ, ◌ো, ◌ৌ or Nil
3. **Consonant Diacritics**, i.e. ◌্য, ◌্র, র্◌, ◌র্য, ◌্র্য or Nil

Commonality of consonant conjuncts in Bangla results in an expansive set of different grapheme roots.

Not all diacritics are used with every grapheme root. Naive combinations can lead upto **~13k** different graphemes.

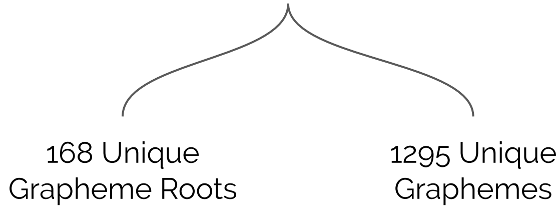
We use the text transcriptions for the Bengali ASR dataset by Google as our corpus and statistically select **1295 commonly used bengali graphemes** in everyday vocabulary., to reduce the number of classes to collect.

Google AI

Large Bengali ASR Dataset

Text Transcription: ~196k sentences

Vocabulary size: ~57k words



168 Unique Grapheme Roots

1295 Unique Graphemes

Bengali.AI Community

The Competition

- Given an image of a grapheme, predict the label of the **root, vowel diacritic** and **consonant diacritic** (see Appendix B for labels).

- Evaluation Criteria :
  Weighted recall = 0.50 * **root** recall + 0.25 **vowel diacritic** recall + 0.25 * **consonant diacritic** recall

- Timeline: 3 Months

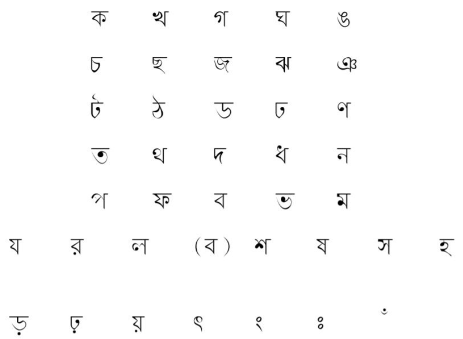
- Winning algorithms must be open-sourced.

- See competition for details (www.kaggle.com/c/bengaliai-cv19)

Appendix A: Bengali Orthography

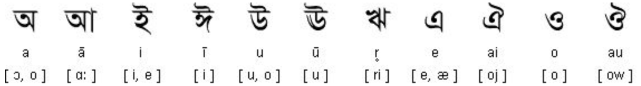
- The bengali alphabets have 40 consonants, and 10 vowels
- Some vowels have a composite phoneme associated with it, i.e. ঋ (ri), ঔ (ow)
- Apart from the 50 alphabets, 118 different consonant conjuncts are commonly used in bangla.

Consonants



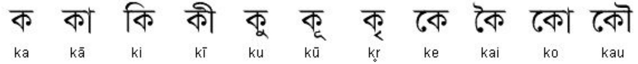
ক	খ	গ	ঘ	ঙ			
চ	ছ	জ	ঝ	ঞ			
ট	ঠ	ড	ঢ	ণ			
ত	থ	দ	ধ	ন			
প	ফ	ব	ভ	ম			
য	র	ল	(ব)	শ	ষ	স	হ
ড়	ঢ়	য়	ৎ	ং	ঃ	ঁ	

Vowels



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]

Vowel Diacritics with ক



ক	কা	কি	কী	কু	কূ	কৃ	কে	কৈ	কো	কৌ
ka	kā	ki	kī	ku	kū	kṛ	ke	kai	ko	kau

Appendix B: Labels (see competition data page for label mappings)

Label	Class
Grapheme roots (168)	অ, আ, ই, ঈ, উ, ঊ, ঋ, এ, ঐ, ও, ঔ, ক, ক্ক, ক্ট, ক্ত, ক্ল, ক্ষ, ক্ষ্ণ, ক্ষ্ম, ক্স, খ, গ, গ্ধ, গ্ন, গ্ব, গ্ম, গ্ল, ঘ, ঘ্ন, ঙ, ঙ্ক,ঙ্ক্ত, ঙ্ক্ষ, ঙ্খ, ঙ্গ, ঙ্ঘ, চ, চ্চ, চ্ছ, চ্ছ্ব, ছ, জ, জ্জ, জ্জ্ব, জ্ঞ, জ্ব, ঝ, ঞ, ঞ্চ, ঞ্ছ, ঞ্জ, ট, ট্ট, ঠ, ড, ড়, ড্ড, ঢ, ঢ়, ণ, ণ্ট, ণ্ঠ, ণ্ড, ণ্ণ, ত, ত্ত, ত্ত্ব, ত্থ, ত্ন, ত্ব, ত্ম, থ, দ, দ্ঘ, দ্দ, দ্ধ, দ্ব, দ্ভ, দ্ম, ধ, ধ্ব, ন, ন্জ, ন্ট, ন্ঠ, ন্ড, ন্ত, ন্ত্ব, ন্থ, ন্দ, ন্দ্ব, ন্ধ, ন্ন, ন্ব, ন্ম, ন্স, প, প্ট, প্ত, প্ন, প্প, প্ল, প্স, ফ, ফ্ট, ফ্ফ, ফ্ল, ব, ব্জ, ব্দ, ব্ধ, ব্ব, ব্ল, ভ, ভ্ল, ম, ম্ন, ম্প, ম্ব, ম্ভ, ম্ম, ম্ল, য, য়, র, ল, ল্ড, শ, শ্ব, ষ, স, স্ট, স্থ, হ, ল্ক, ল্গ, ল্ট, ল্প, ল্ব, ল্ম, ল্ল, শ্চ, শ্ন, শ্ম, শ্ল, ষ্ক, ষ্ট, ষ্ঠ, ষ্ণ, ষ্প, ষ্ফ, ষ্ম, স্ক, স্ত, স্ন, স্প, স্ফ, স্ব, স্ম, স্ল, স্স, হ্ন, হ্ব, হ্ম, হ্ল, ৎ, ◌ং, ◌ঃ
Vowel Diacritics (11)	Null, ◌া, ◌ি, ◌ী, ◌ু, ◌ূ, ◌ৃ, ◌ে, ◌ৈ, ◌ো, ◌ৌ
Consonant Diacritics (7)	Null, ◌্য (য-ফলা), ◌্র (র-ফলা), র্ (রেফ), ◌ঁ, ◌র্য (রেফ য-ফলা), ◌্র্য (র-ফলা য-ফলা)

Bengali.AI
Community