Homework1
Name: Chakrya Ros

My native language is Khmer, (Cambodia language). Khmer contents have 1,110,809 characters and about 160,000 Khmer words. I know 26,000 Khmer words.
Here is an example of Khmer text:

ប្រជាជនគ្រប់ជនជាតិទាំងអស់នៅលើពិភពលោកតែងតែរៀបចំ ប្រារព្ធពិធីប្រុណ្យចូលឆ្នាំតាមប្រពៃណីជាតិ របស់ខ្លួនរៀងរាល់ឆ្នាំ។

ពិធីនេះរៀបចំឡើងតាមទំនៀមទម្លាប់នៃជនជាតិរបស់គេ។ ទន្ទឹមគ្នានេះ ដែរ ប្រជាជនខ្មែរប្រារព្វ ពិធីបុណ្យចូលឆ្នាំតាម ជំនឿ និង ទំនៀមទំលាប់របស់ខ្លួនដែរ។ [1]

Translated in English : People of all nations of the world are always organized celebrate its national traditions every year. The festival is organized according to the traditions of their people. At the same time, also, khmer people celebrate this year's festival according to their beliefs and customs its habit too.

In this text, there are 46 tokens and 31 vocabulary words. In Khmer text, there are no spaces between words, but the spaces indicate the end of a clause or sentence.

1. Khmer is primarily an analytic language with no inflection. Syntactic relations are mainly determined by word order. Khmer script is written from left to right with multiple level of character stacking possible and uses diacritics that improve the pronunciation of words. There are 35 consonants (see Table II), 14 independent vowels (see Table III), 23 dependent vowels (see Table IV) and some diacritic signs and special characters (see Table VI). [2]

To consider the word, each Khmer words contain one based character that include consonants, independent vowels, consonant subscripts and some other signs. A word could also be a join of two or more Khmer words together. Some examples of Khmer words are កម្ពុជា (Cambodia) pronoun "kampouchea" it has 3 consonants, 1 subscript consonants and 2 dependent vowel, វចនានុក្រម (dictionary) pronoun "vochneanoukram" it has 6 consonants, 3 dependent vowels. These combinations are considered a word in Khmer. An example of a word that joins two words together is ចូលចិត្ត (like) pronoun "chaulchett" that have two words joined, ចូល (in) pronoun "chaul' and *ចិត្ត* (heart) pronoun "chett".

[1] "new_year_set - SEAsite." http://www.seasite.niu.edu/khmer/reading/intermediate/culture/new_year/new_year_set.htm. Accessed 22 Jan. 2020.
[2] "Khmer Word Segmentation based on Bi-Directional Maximal ...." http://www.apsipa.org/proceedings_2014/Data/paper/1406.pdf. Accessed 22 Jan. 2020.

2. I know it’s the word by listening to people talk, and I learn to say those words, write those words in the sentence. When I was born, I didn’t understand any words that my parents or other people talk to me. However, when I started to grow up from a baby to an adult. I know those are the words because they spoke to me everyday and I spoke those words and wrote those words in the sentences.

Another argument that I know the word, when people say those words to me, I understand the meaning of those words and I can explain those words to other people from what I understand each word.

3.   In Khmer, there are about 160,000 words and I argue that I know 26,000 words. The reasons that I don’t know all Khmer words because I don’t read or write in Khmer everyday. I just speak Khmer to my family. I use maybe 200 or 300 frequency words in Khmer in my family.

However, I argue that I know 26,000 words because I read some articles in Khmer for this research. I found out many words I didn’t know what them means. For an example, I read one article “ The fishermen warm about extinction in Tonle Sap Lake”[3], it’s about 1420 tokens, and 19 words that I didn’t understand what it means. Another article is “Innovative culture and business innovation in Ghana are growing”[4], it’s about 630 tokens and 8 words that I didn’t understand.

4. The method I used to come up with these numbers is I read several Khmer articles online.[5] I know that many words in Khmer texts that I don’t understand what they mean. Some words I even don’t know how to pronounce. In this research, I could not find any tools to count the words in Khmer text like English using NLTK. So I decided to pick two articles that I described in questions 3 to count manually. I used website: “https://kheng.infor”[6] to help me to count each word. I just copied Khmer text into the box and submitted. Then when I put the cursor on each word, it highlights each word for me and if I double click one each word, it pronounced and explained the meaning of each word. That’s a useful website for understand Khmer words.

Overall, I don’t all all the Khmer words. If I don’t keep reading or writing them, I would forget all. It's more difficult than English because I have more consonants, subscript consonants and vowels to form a word.

[3] "ប្រជានេសាទព្រមានពីមូលហេតុ ... - VOA Khmer." 22 Jan. 2020, https://khmer.voanews.com/a/fishermen-warn-about-fish-extinction-in-tonle-sap-lake/5256254.html. Accessed 23 Jan. 2020.

[4] "វប្បធម៌នវានុវត្តន៍និងការ ... - VOA Khmer." 4 Dec. 2019, https://khmer.voanews.com/a/ghanas-innovation-and-startup-culture-thriving/5192136.html. Accessed 23 Jan. 2020.

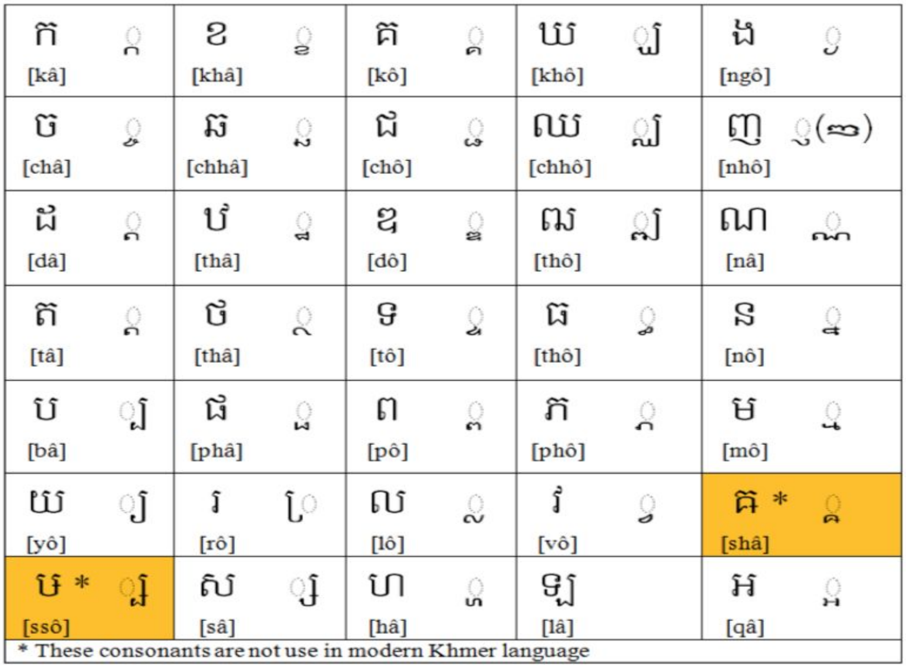
[5] "Reading - SEAsite." http://seasite.niu.edu/khmer/Reading/reading_set.htm. Accessed 23 Jan. 2020.

[6] "kheng.info / read." https://kheng.info/. Accessed 23 Jan. 2020.

Reference:

1. "new_year_set - SEAsite." http://www.seasite.niu.edu/khmer/reading/intermediate/culture/new_year/new_year_set.htm. Accessed 22
2. "Khmer Word Segmentation based on Bi-Directional Maximal ...." http://www.apsipa.org/proceedings_2014/Data/paper/1406.pdf. Accessed 22 Jan. 2020.
3. Wikipedia, Khmer Language.
   Available at http://en.wikipedia.org/wiki/Khmer_Language
4. https://polymath.org/khmer_body.php
5. https://kheng.info/
6. https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Khmer
7. https://khmer.voanews.com/a/fishermen-warn-about-fish-extinction-in-tonle-sap-lake/5256254.html
8. https://khmer.voanews.com/a/fishermen-warn-about-fish-extinction-in-tonle-sap-lake/5256254.html

TABLE II
KHMER CONSONANTS AND CONSONANT SUBSCRIPTS



ក ្ក [kâ]	ខ ្ខ [khâ]	គ ្គ [kô]	ឃ ្ឃ [khô]	ង ្ង [ngô]
ច ្ច [châ]	ឆ ្ឆ [chhâ]	ជ ្ជ [chô]	ឈ ្ឈ [chhô]	ញ ្ញ(្ញ) [nhô]
ដ ្ដ [dâ]	ឋ ្ឋ [thâ]	ឌ ្ឌ [dô]	ឍ ្ឍ [thô]	ណ ្ណ [nâ]
ត ្ត [tâ]	ថ ្ថ [thâ]	ទ ្ទ [tô]	ធ ្ធ [thô]	ន ្ន [nô]
ប ្ប [bâ]	ផ ្ផ [phâ]	ព ្ព [pô]	ភ ្ភ [phô]	ម ្ម [mô]
យ ្យ [yô]	រ ្រ [rô]	ល ្ល [lô]	វ ្វ [vô]	ឝ * ្ឝ [shâ]
ឞ * ្ឞ [ssô]	ស ្ស [sâ]	ហ ្ហ [hâ]	ឡ [lâ]	អ ្អ [qâ]

* These consonants are not use in modern Khmer language

TABLE III
KHMER DEPENDENT VOWELS

◌ា [a] / [éa]	◌ិ [ĕ] / [ĭ]	◌ី [ei] / [i]	◌ឹ [œ̆]	◌ឺ [œ]	◌ុ [ŏ] / [ŭ]	◌ូ [o] / [u]	◌ួ [uŏ]
◌ើ [aeu]/[eu]	◌ឿ [eua]	◌ៀ [iĕ]	◌េ [é]	◌ែ [ê]	◌ៃ [ai]/[ey]	◌ោ [aô]/[oŭ]	◌ៅ [au]/[ŏu]
◌ុំ [om]/[ŭm]	◌ំ [âm]/[um]	◌ាំ [ăm]/[ŏâm]	◌ះ [ăh]/[eăh]	◌ុះ [ŏh]/[uh]	◌េះ [éh]	◌ោះ [aŏh]/[uŏh]	

TABLE IV
KHMER INDEPENDENT VOWELS

ឥ [ĕ]	ឦ [ei]	ឧ [ŏ]	ឨ *	ឩ [ŭ]	ឪ [ŏu]	ឫ [rœ̆]	ឬ [rœ]
ឭ [lœ̆]	ឮ [lœ]	ឯ [é]	ឰ [ai]	ឱ , ឲ [aô] , [aôy]	ឳ [âu]		

* These Independent Vowels is not use in modern Khmer language

TABLE V
KHMER NUMERALS

Khmer Numerals	០	១	២	៣	៤	៥	៦	៧	៨	៩
Arabic Numerals	0	1	2	3	4	5	6	7	8	9

TABLE VI
KHMER DIACRITIC SIGNS AND OTHER SPECIAL CHARACTERS

◌ំ	◌ះ	◌ៈ	◌៉	◌៊	◌្	◌់	◌៌
◌៍	◌៎	◌៏	◌័	◌៑	◌̟	៛	ៗ
។	៕	៖	?	!	,	.	"
«	»	៙	៚				