



KHMER COLLATION API TECHNICAL REPORT

Cambodia Country Component
PAN Localization project
PAN Localization Cambodia (PLC) of IDRC

June 12, 2009

Prepared by: Mr. Tith Sakal

1. Introduction.....	2
2. Khmer Collation API	2
2.1. Normalization.....	2
2.2. Previous version	3
2.3. Functionality	3
3. Khmer Collation API Data.....	5

1. Introduction

Collation is a process of comparing two words of string respecting to a specific order rule of a language. To provide an easy way for developer to collate Khmer words for the purpose of any application development, we need to build Khmer Collation Engine.

2. Khmer Collation API

Khmer Collation API has been developed by using Visual Studio Dot Net 2005 as the tool. Thus, we got a DLL (Dynamic Link Library) file. There are more than two classes contain in the API. Anyways, the main important classes in this are, NormalizationEngine, CollationEngine and SortingEngine.

NormalizationEngine is a class used to normalized Khmer word written into different way to one specific order; Khmer Unicode font provides different way to write some Khmer word. CollationEngine is a class to hold the process of collates two Khmer words. The purpose of SortingEngine is to sort Khmer Unicode string array given from outside application and return the list of word which already sorted.

2.1.Normalization

Khmer Unicode font makes it possible for the user to represent a word or syllable in different way.

For example, the word ស្រី can be typed in two different ways:

ស្រី	=	ស	+	្រ	+	្រ	+	ី
ស្រី	=	ស	+	្រ	+	្រ	+	ី

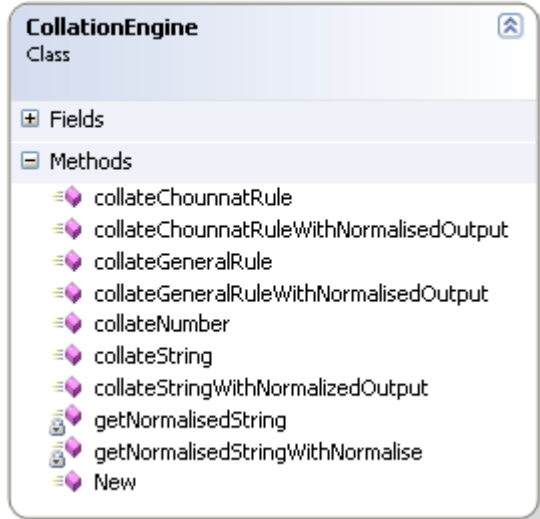
In user perspective, the two strings are considered the same according to its visibility. However, in Computer, the two strings are different because of the order.

2.2.Previous version

For the previous version of Khmer Collation API we got three Library, PLCNormalization.DLL, PLCCollation.DLL and PLCSorting.DLL. To use these library developers need to copy the data file and put them in the same directory for these library access to. According to the current version we combine those API and the data files into one API called KhmerCollation API.

2.3.Functionality

CollationEngine



Usage:

```
Dim Str1 as String = “ផ្អែ”  
Dim Str2 as String = “កកយ”  
Dim CollationObject as New CollationEngine()  
Dim Result as Byte  
  
Result = CollationObject.CollateString( Str1, Str2)
```

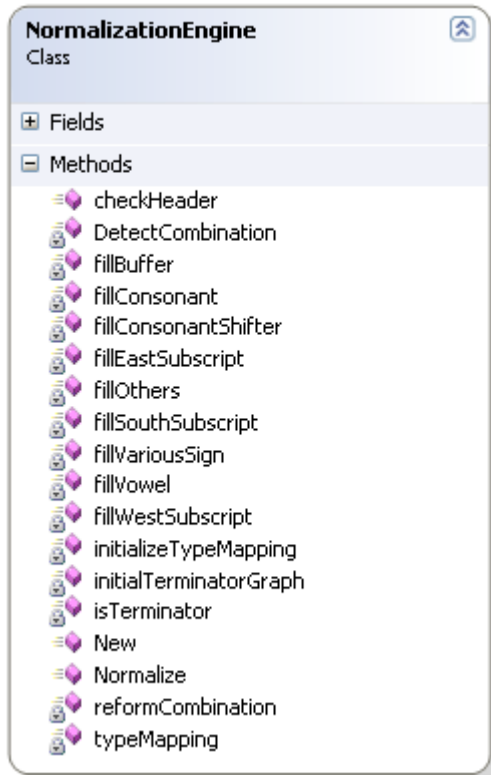
Result:

- If the Result = 0 means two sequences are equal
- If the Result = 1 means the first sequence is bigger
- If the Result = 2 means the second sequence is bigger

Note:

- The Result return 0 if String1 and String2 = Nothing
- The Result return 2 if String1 = Nothing
- The Result return 1 if String2 = Nothing

- **NormalizationEngine**



Usage:

```
Dim StringToNormalize as String = “ស្រី”
Dim AfterNormalized as String
Dim NormalizeObject as New NormalizationEngine()
AfterNormalized = NormalizeObject.Normalize(StringToNormalize)
```

Result:

StringToNormalize	=	ស	+	្រ	+	្រ	+	្រ
AfterNormalized	=	ស	+	្រ	+	្រ	+	្រ

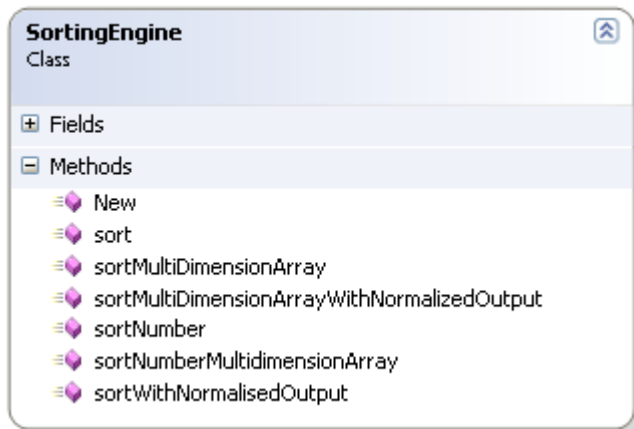
Note:

The visibly of the result is the same, in fact, the ordering of subscript in the string is different. After the string has been Normalized the ordering must be:
Base Consonant + South Subscript + West Subscript + Vowel

General Rule:

- 1- Base Consonant
- 2- First Subscript (South Subscript)
- 3- Second Subscript (West Subscript or East Subscript)
- 4- Consonant Shifter
- 5- Vowel
- 6- Various Sign

SortingEngine



Usage:

```
Dim StringCollection () As String = {"ខន្តី", "កណ្តាល", "ចាវ", "កាយ", "វិល", "ហិនហោច",  
"ឃុំ", "កុំ"}  
Dim SortingObject as New SortingEngine()  
  
SortingObject.Sort( StringCollection, RULE.CHOUNNAT_DICTIONARY)
```

Result:

```
StringCollection = {"កាយ", "កណ្តាល", "កុំ", "ខន្តី", "ឃុំ", "ចាវ", "វិល", "ហិនហោច"}
```

3. Khmer Collation API Data

KhmerCollation API not only contains functionality but also Data. Dot Net Technology provides us to store data files in Byte and Text file format. We embedded our data files in to the API project and we extract them to local disk for our API access to. After our API has been stop using the data files will delete automatically.

To extract data we have built a class named ExtractData. Its responsibility is to extract data from API to local disk ([SystemDrive]\Widows\System32). Some classes in KhmerCollation API will explore data from the path above.

Data files used in the API are **chounnatwordlist.dat**, **DataFile.dat**, **terminatorGraph.dat**, **CollationSequence.dat** and **CollationSequenceChounnat.dat**. These files are text file and store in UTF-8 encoding.