Mongolian Language Resource Assessment

Yiru

May 18, 2016
Overview

1 Introduction

2 Traditional Mongolian Script
Mongolian

- **Region:** All of Mongolia and Inner Mongolia; parts of Liaoning, Jilin, Heilongjiang and Gansu provinces in China
- **Native speakers:** 5.2 million (2005) (2.8m in Mongolia, half of 5.8m ethnic Mongols in China)
- **Standard forms:** Khalkha (Mongolia); Chakhar (Inner Mongolia)
- **Dialects:** Khalkha, Chakhar, Kharchin, Baarin, Ordos, and so on
- **Writing system:** Traditional Mongolian script (in Inner Mongolia), Cyrillic Mongolian script (in Mongolia), Todo script and so on
Red area includes all of Mongolia, most of Inner Mongolia and Kalmykia, three enclaves in Xinjiang, multiple tiny enclaves round Lake Baikal, part of Manchuria, Gansu, Qinghai, and one place that is west of Nanjing and in the south-south-west of Zhengzhou.
Issues

- **Dialect or Language:** Khalkha, Chakhar, Ordos; Buryat and Oirat (including the Kalmyk variety); Kharchin, Khorchin...
  
  According to UNESCO, Kalmyk is "Definitely endangered"

The delimitation of the Mongolian language within Mongolic is a much disputed theoretical problem.

- **Scripts:** Traditional Mongolian, Cyrillic, Todo, Square, KebtegeDorbeljin, Galik, Soyombo script...
Dialact
Classic Mongolian Script

(a) Traditional
(b) Todo
(c) Chinese
(d) Soyombo
(e) Square
(f) Cyrillic
(g) KebtegeDorbeljin
The traditional Mongolian script character code set has been placed in Unicode at the range of 1800-18AF. But it is not enough to solve problems in processing information in Mongolian.

- Written vertically from top to bottom in columns advancing from left to right. This directional pattern is unique among existing scripts. Thus, general operating systems fail to correctly display traditional Mongolian script.

- Characters are written in succession, meaning that depending on where a letter is placed in a word, it may have different forms. There are at least three different forms for each letter and some letters have a dozen different forms.
The Unicode standard includes only the basic character sets, special punctuation symbols and numerals, but does not explicitly encode the variant forms or the ligatures.

There was no standardized IME which supports Traditional Mongolian Script before the IME included in Windows Vista. In China a 8 bit encoding standard GB 8045-87 was established but not used in Mongolia.

Because of the unique characteristics, procedure to process the information such as inputting, displaying, encoding, typing, typesetting and recognizing have become more complicated.
There are more and more researches about Mongolian language:

- creating digital libraries
- comparison and conversion between Traditional Mongolian and Cyrillic Mongolian
- part of speech tagging, rendering

Problems: Not enough experience in NLP research and development; shortage of NLP trained human resource; lack of professional computational linguist
Garmaabazar Khaltarkhuu, Akira Maeda (2008)
Developing a Traditional Mongolian Script Digital Library
*Digital Libraries: Universal and Ubiquitous Access to Information* 41-50
Thank You!