

Out-of-the-box Universal Romanization Tool *uroman*

Ulf Hermjakob, Jonathan May, Kevin Knight • USC Information Sciences Institute • ulf@isi.edu

1. What is *uroman*?

A tool for converting text in myriads of scripts such as Chinese, Arabic and Cyrillic into a common Latin-script representation.

Romanization enables the application of string-similarity metrics across scripts.

| | Hindi | Urdu | English |
|--------------|--------|-------|---------|
| Original | नेपाल | نیپال | Nepal |
| Romanization | nepaal | nipal | Nepal |

| | Original | Romanization |
|----------|---|---|
| Amharic | በርሊን የጀርመን ዋና ከተማ ነው። | bareline yajaremane waanaa katamaa nawe. |
| Arabic | المملكة العربية السعودية | almmilka al'rbya als'wdya |
| Greek | Γερόυν Ντάισελμπλουμ | Geroun Daiselbloum |
| Hebrew | עזרת תורה בירושלים | 'zrt tvrh vyrvshlym |
| Japanese | アメリカ | amerika |
| Korean | 세계에서 6번째로 면적이 넓은 나라이다. | segyeeseo 6beonjjaero myeonjeogi neolbeun naraida. |
| Mandarin | 北卡罗来纳 | beikaluolaina |
| Nepali | तिब्बती भाषामा यसको नाम चोमोलुङ्गमा हो। | tibbatii bhaassaamaa yasako naam comolunggamaa ho . |
| Tamil | இதன் தலைநகராகச் சென்னை உள்ளது. | itan talainakaraakac cennai ullatu. |
| Tibetan | ལཱ་སྐ་གྲོང་མཉམ་པའི་རྒྱུ་ | lha'sa'grong'khyer |

2. How does *uroman* work?

- It uses *Unicode* tables to predict the romanization of a character: CYRILLIC CAPITAL LETTER TE WITH MIDDLE HOOK → TE → т

One set of heuristics identifies the pronunciation token (“TE”). A second set of heuristics identifies the core pronunciation (“т”).

- As the Unicode table heuristics often don’t work, we manually built 1,088 rules to deal with exceptions, especially for m→n character mappings. For examples, see upper table to the right.

- Pinyin table for Chinese characters.

Standard romanization algorithm for Korean Hangeul characters.

- Special module to map non-Western digital numbers to Western Arabic numerals. For examples, see lower table to the right.

| | | | | |
|-----|------|-----|------|-------------------------------------|
| ::s | μπ | ::t | b | ::use-only-at-start-of-word |
| ::s | μπ | ::t | mb | ::t-alt b, mp |
| ::s | چ | ::t | ch | ::t-alt q ::lcode uig |
| ::s | ئو | ::t | o | ::lcode uig |
| ::s | ちよ | ::t | cho | |
| ::s | フエ | ::t | fe | |
| ::s | eaux | ::t | eaux | ::t-alt o ::example Bordeaux |
| ::s | gh | ::t | gh | ::t-alt f, "" ::ex. laugh, daughter |

Romanization rules with two examples each for Greek, Uyghur, Japanese, and English, with a variety of n-to-m mappings. (::s = source; ::t = target; ::lcode = language code)

| | Original | Romanization |
|---------|----------|--------------|
| Amharic | ፲፱፻፲፰ | 1998 |
| Bengali | ১৯৪৯ | 1949 |
| Chinese | 二十五万六千 | 256000 |
| | 25.6万 | 256000 |

3. Features of *uroman*

- Input: UTF8-encoded text and an optional ISO-639-3 language code.
- Output: Romanized text (default) or lattice of romanization alternatives in JSON format.
- N-to-m mapping for groups of characters that are non-decomposable with respect to romanization.
- Nearly universal.

Current limitations: Japanese kanji interpreted as Mandarin Chinese; limited coverage for ancient extinct scripts (hieroglyphics, cuneiform).

- Context-sensitive and source language-specific romanization rules.
- Romanization includes (digital) numbers.
- Romanization includes punctuation.
- Preserves capitalization.
- Interactive demo URL: bit.ly/uroman
- Freely and publicly available** (data, Perl script) at bit.ly/isi-nlp-software

To the best of our knowledge, *uroman* is the first publicly available (near) universal romanizer that handles n-to-m character mappings.

4. Applications using *uroman*

- Named entity recognition (Ji et al., 2017; Mayfield et al., 2017)
- End-to-end transliteration (Mayhew et al., 2016)
- Machine translation of low-resource languages (Cheung et al., 2017)
- Chinese Room* tool (Hermjakob et al., 2018, see demo at ACL 2018)

This work has been published as *Out-of-the-box Universal Romanization Tool uroman* in the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Demo track. Melbourne. July 2018. ACL-2018 Best Demo Paper Award.