Domovyk:

Multilingual Transliteration for Cyrillic Text

Ian Goodale
UT-Austin
ian.goodale@austin.utexas.edu

Introduction:

Domovyk is a package for the computational transliteration between Cyrillic alphabets and Latin script, with functionality and support for languages not seen in other Python transliteration packages.

Key Points:

- The package supports eight languages, each with a unique alphabet, and can transliterate to and from each language while supporting unique aspects of transliteration in each language.
- It follows the American Library Association - Library of Congress transliteration tables, providing a standardized and vetted system of transliteration for each language.
- It also supports composite Unicode characters, allowing for greater accuracy when performing transliteration to and from Cyrillic.

Conclusion:

Domovyk aims to empower researchers, developers working with NLP and text processing, and other users who are interested in furthering work on multilingual text within the Python ecosystem.

Domovyk provides robust multilingual, standardized transliteration for Cyrillic languages with custom algorithm and alphabets, support for composite Unicode characters, and support for seamlessly transliterating back and forth between Cyrillic and Latin alphabets.

Domovyk is named after household spirits from Slavic mythology (pictured to the right).

https://github.com/ian-nai/domovyk

Background image source: https://gallica.bnf.fr/ark:/12148/btv1b53258790f

The languages Domovyk provides support for range from the commonly used, such as Russian and Ukrainian, to the less common, like Church Slavonic. Each language has a customized transliteration algorithm tailored to its individual needs, allowing for accuracy and faithfulness to each language's transliteration rules.

Languages supported are:

- Belarusian
- Bulgarian
- Carpatho-Rusyn
- Church Slavonic
- Macedonian
- Russian
- Serbian
- Ukrainian

Available under the GNU General Public License

Four functions can be called for each language:

- transliterate(var, lang) - Transliterates a Cyrillic string or list of strings (var) from a specified language (lang) into the Latin alphabet. Returns a string.
- translatinate(var, lang) - Transliterates a string or list of strings in the Latin alphabet (var) to a specified Cyrillic script (lang). Returns a string.
- transliterateSents(var, lang) - Tokenizes a given Cyrillic (var) into a list of sentences, then transliterates those sentences from a specified language (lang) into the Latin alphabet. Returns a list.
- translatinateSents(var, lang) - Tokenizes a given string in the Latin alphabet (var) into a list of sentences, then transliterates those sentences to a specified Cyrillic script (lang). Returns a list.

The package is simple to use; here are some simple examples:

```
from domovyk import translit

belarusian_to_latin =
translit.transliterate('Як справы?', 'bel')
latin_to_macedonian =
translit.translatinate('Hi, how are you?',
'mac')
ukrainian_to_latin_sents =
translit.transliterateSents('Єхидна, ґава,
їжак ще й шиплячі плазуни бігцем
форсують Янцзи.', 'ukr')
latin_to_russian_sents =
translit.translatinateSents('Hello, how are
you doing?', 'rus')
```