APERTIUM’S WEB TOOLCHAIN FOR LOW-RESOURCE LANGUAGE TECHNOLOGY

Sushain Cherivirala
sushain@skc.name
Apertium Community

Shardul Chiplunkar
shardul.chiplunkar@gmail.com
Apertium Community

Jonathan North Washington
jonathan.washington@swarthmore.edu
Swarthmore College,
Apertium Community

Kevin Brubeck
Unhammer
unhammer+apertium@mm.st
Apertium Community
OVERVIEW

- Components:
  - HTML-Tools \textit{(web front end)}
  - \textbf{APy} \textit{(API in Python)}
  - \textbf{Apertium's web toolchain}

- Provides online access to Apertium language technology (MT, etc.)
- Roadmap of talk:
  - How \textbf{HTML-Tools}, \textbf{APy}, and \textbf{Apertium} fit together
  - Key features and technology
  - Advanced features and ongoing work
  - Usage statistics
  - Relevance to low-resource languages
HTML-TOOLS
HTML-TOOLS

FULL INTERNATIONALISATION

- Default to browser locale
- RTL text and layout
- Autoglossonyms from APy: SIL, Unicode CLDR, manual curation
- Easy to contribute/extend localisations
HTML-TOOLs

- Instant translation
- Remembers text and selections, updated in URL
- Simple build, can be built and used offline
- Matomo for analytics
Example of use
- query (URL): `/translate?langpair=kaz|tat&q=Қанша мәтін?`
- response (JSON): `{... "translatedText": "Ничә текст?"}`

Written entirely in python3 (tornado)
Query endpoints modelled after ScaleMT
- In turn modelled after Google Translate's API
- Drop-in replacement for major APIs

Pipelines for each pair and morphological analysis/generation
- Leverages existing Apertium mode files
- Pipeline kept open and flushed between requests
- Multiple pipelines on-demand, split large requests
- APy can make a web service from any Unix pipeline: e.g. Ávvir
# ADVANCED FEATURES

## LANGUAGE VARIANTS

<table>
<thead>
<tr>
<th>French</th>
<th>Spanish</th>
<th>English</th>
<th>Detect Language</th>
<th>French</th>
<th>Spanish</th>
<th>Hindi</th>
<th>Occitan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans</td>
<td>Erzya</td>
<td>Latin</td>
<td>Sakha</td>
<td>Turkmen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albanian</td>
<td>Esperanto</td>
<td>Latvian</td>
<td>Sardinian</td>
<td>Ukrainian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aragonese</td>
<td>Faroese</td>
<td>Livonian</td>
<td>Scots</td>
<td>Urdu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assamese</td>
<td>Finnish</td>
<td>Livvi-Karelian</td>
<td>Serbo-Croatian</td>
<td>Uyghur</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asturian</td>
<td>French</td>
<td>Luxembourgish</td>
<td>Croatian</td>
<td>Uzbek</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangla</td>
<td>Quebec French</td>
<td>Macedonian</td>
<td>Serbian</td>
<td>Wayuu</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bashkir</td>
<td>Galician</td>
<td>Mandarin Chinese</td>
<td>Bosnian</td>
<td>Welsh</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belarusian</td>
<td>German</td>
<td>Marathi</td>
<td>Sicilian</td>
<td>Western Mari</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breton</td>
<td>Gujarati</td>
<td>Moksha</td>
<td>Sindhi</td>
<td>Arabic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>Hawaiian</td>
<td>Nepali</td>
<td>Sinhala</td>
<td>Estonian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalan</td>
<td>Hindi</td>
<td>Norwegian Bokmål</td>
<td>Slovak</td>
<td>Indonesian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valencian</td>
<td>Icelandic</td>
<td>Norwegian Nynorsk</td>
<td>Slovenian</td>
<td>Irish</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central Kurdish</td>
<td>Interlingua</td>
<td>East Norwegian</td>
<td>Spanish</td>
<td>Komi-Zyrian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chuvash</td>
<td>Iranian Persian</td>
<td>Occitan</td>
<td>Spanish Sign</td>
<td>Malay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corsican</td>
<td>Italian</td>
<td>Aranese</td>
<td>Language</td>
<td>Maltese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crimean Tatar</td>
<td>Karakalpak</td>
<td>Polish</td>
<td>Sranan Tongo</td>
<td>Translitered</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>Kashubian</td>
<td>Portuguese</td>
<td>Swedish</td>
<td>Scottish Gaelic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danish</td>
<td>Kazakh</td>
<td>Brazilian</td>
<td>Tajik</td>
<td>Udmurt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>Khalkha Mongolian</td>
<td>Portuguese</td>
<td>Tajik Transliterated</td>
<td>Võro</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>American English</td>
<td>Punjabi</td>
<td>Tatar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
WEBPAGE AND DOCUMENT TRANSLATION

- **Webpage translation**: output displayed in iframe
  - Links within webpages trigger translation too
- **Document translation**: RTF, LibreOffice/OpenOffice, MS Office 2003+
ADVANCED FEATURES

MODES

- Tabs in HTML-Tools interface
- Currently deployed: morphological analysis and generation
- Other developed modes: APy sandbox, spell-checker
- Customizable and extendable
MULTI-STEP TRANSLATION

- Translation through intermediate (pivot) languages
- Defaults to shortest path—future work may incorporate pair quality
- User can choose a path by selecting languages and viewing valid paths
ONGOING WORK

DICTIONARY LOOKUP

- All possible translations with part of speech
- Future work:
  - Reverse translations for context
  - Morphological subcategory information (Gender, number, conjugation, etc.)
  - Multi-word units
ONGOING WORK

SPELL-CHECKING

- libvoikko or hfst-ospell create spellers from language modules
- Separate tab in HTML-Tools
User can suggest translations for unknown words in context
Logged by APy, can seed quality improvements
Rating translations?
USAGE

• From [apertium.org](http://apertium.org) since April 2014:
  □ 2.1M visits
  □ 183 countries
  □ 19.8M translation requests
  □ 11 pairs with over 100k requests each
  □ most popular translation directions:
    - nob ↔ nno (Norwegian Bokmål ↔ Nynorsk)
    - spa → cat (Spanish → Catalan)
    - por → spa (Portuguese → Spanish)
• HTML-Tools and APy analytics tools
RELEVANCE

- Provides quick and easy access to Apertium language technology (able to be used by anyone, regardless of ability to use command line)
- Public access, low cost, low maintenance
  - Low disk, memory, processing requirements, and offline usage → low infrastructure costs
  - Free and open source (GPLv3)
  - Focus funding and time on the specific language technology
- Customizability and sub-sites (e.g. turkic.apertium.org)
- Modularity: APy as a general service
- Multi-step translation: new moderate-quality pairs with little effort, extending range as a temporary solution
- Suggestions: an additional path for low-resource language users to directly help in development
COMMUNITY

- **Code:** bugs, suggestions, pull requests welcome!
  - [github.com/apertium/apertium-html-tools](https://github.com/apertium/apertium-html-tools)
  - [github.com/apertium/apertium-apy](https://github.com/apertium/apertium-apy)
- Non-technical contributions: localizations, translation improvements
- **Mailing list:** [apertium-stuff](https://sourceforge.net/lists/list.jsp?group_id=56129) on SourceForge
- **IRC:** [#apertium](https://freenode.net) on Freenode
- **Give it a try!**
  - [apertium.org](http://apertium.org), [turkic.apertium.org](http://turkic.apertium.org), [beta.apertium.org](http://beta.apertium.org)
THANKS!