APERTIUM'S WEB TOOLCHAIN FOR LOW-RESOURCE LANGUAGE TECHNOLOGY

Sushain Cherivirala
Apertium Community

Shardul Chiplunkar
Apertium Community

Jonathan North Washington
Swarthmore College, Apertium Community

Kevin Brubeck Unhammer
Apertium Community

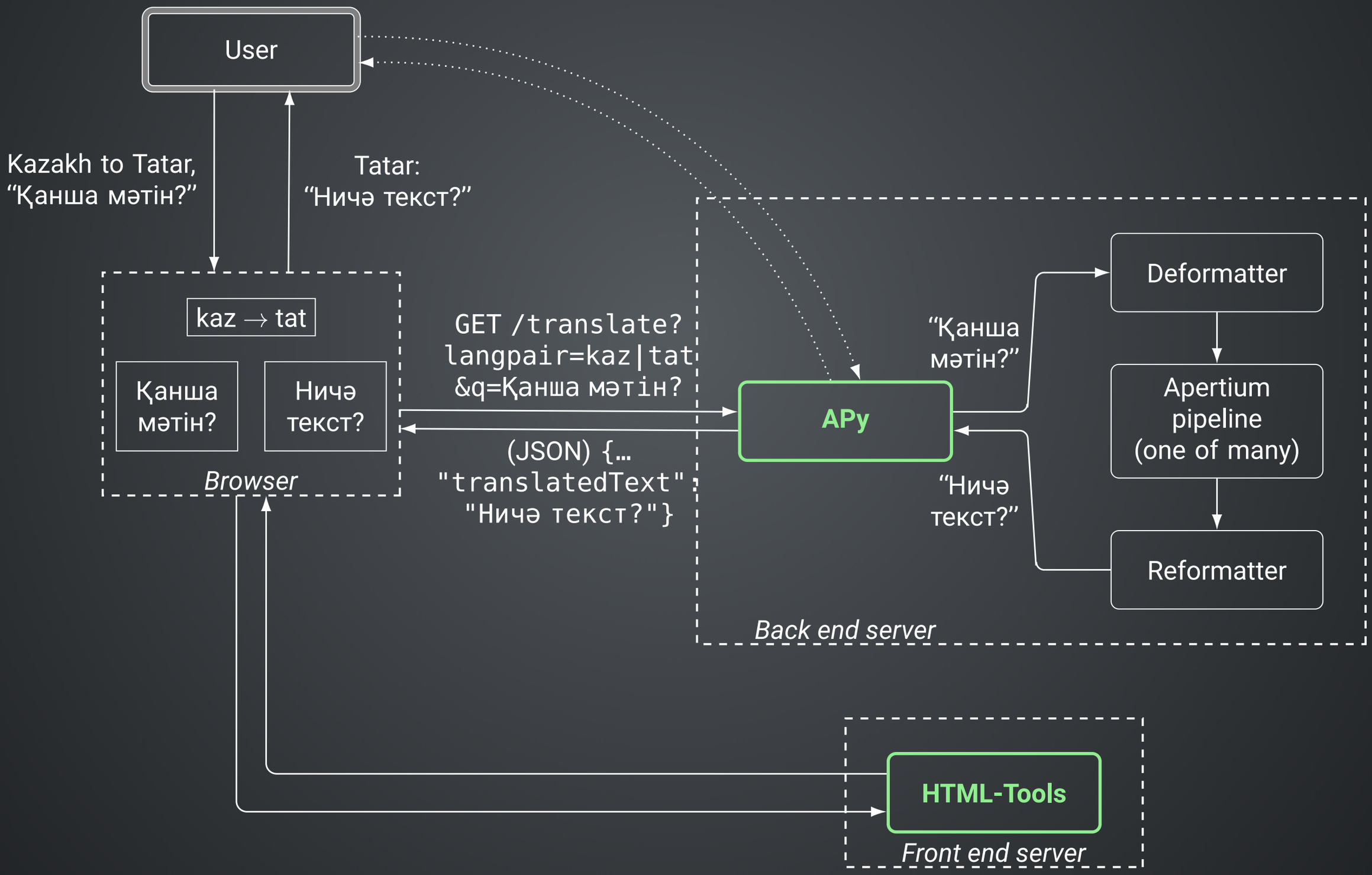
OVERVIEW

- Components:

	HTML-Tools (*web front end*)
+	**APy** (*API in Python*)
=	Apertium's web toolchain

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

OVERVIEW


User
Kazakh to Tatar,
“Қанша мәтін?”
Tatar:
“Ничә текст?”
kaz → tat
Қанша мәтін?
Ничә текст?
Browser
GET /translate?
langpair=kaz|tat
&q=Қанша мәтін?
(JSON) {…
"translatedText"
"Ничә текст?"}
APy
“Қанша мәтін?”
“Ничә текст?”
Deformatter
Apertium pipeline (one of many)
Reformatter
Back end server
HTML-Tools
Front end server

HTML-TOOLS

Apertium Turkic

Una prataforma lìbera a còdighe abertu pro sa tradutzione automàtica

sardu

Tradutzione | Anàlisi morfològica | Generatzione morfològica

қазақша | татарча | къумукъча | Rileva sa limba in manera automàtica

⇄

татарча | кыргызча | къумукъча

Borta

Қазақшадан татаршаға оңай мәтіндер қатесіз аудара алады.

Казакъчадан татарчага уңай текстлар хатасыз тәрҗемә итә ала.

Borta unu documentu | Jorgal neahttasiiddu

Sinnala sas paràulas disconnotas

Tradutzione istantànea

Traducció en múltiples passes (experimental)

Informatziones | Iscàrriga | Documentatzione | Cuntatade·nos

Ais agatadu una faddina? Agiuade·nos a megiorare Apertium!

v3.2.2-27-g126c9c7

FULL INTERNATIONALISATION

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

OTHER FEATURES

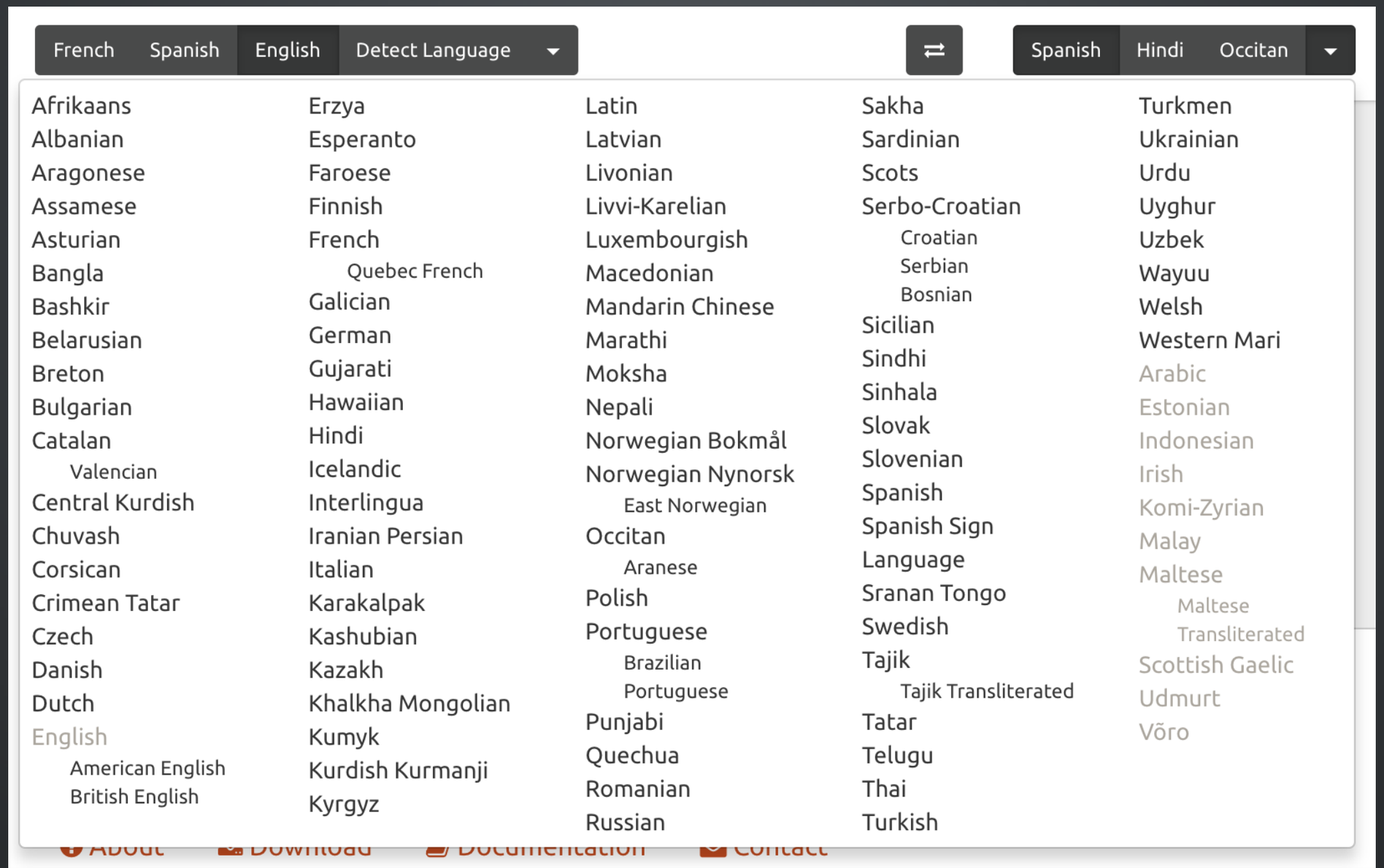
- Instant translation
- Remembers text and selections, updated in URL
- Simple build, can be built and used offline
- Matomo for analytics

APY

- 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

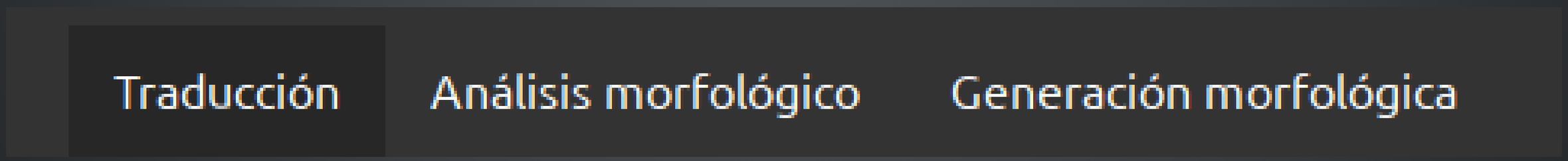


WEBPAGE AND DOCUMENT TRANSLATION

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

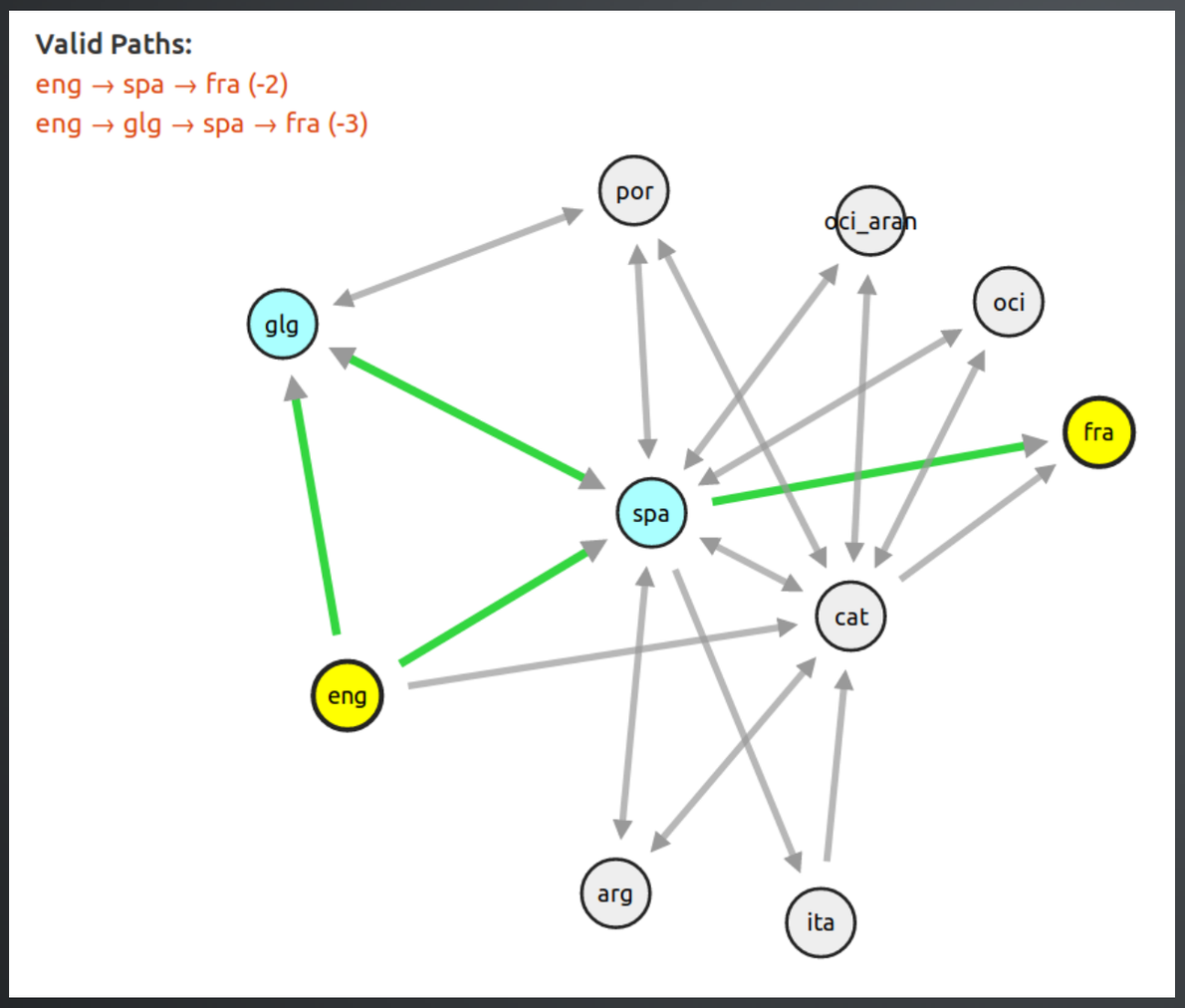
MODES

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



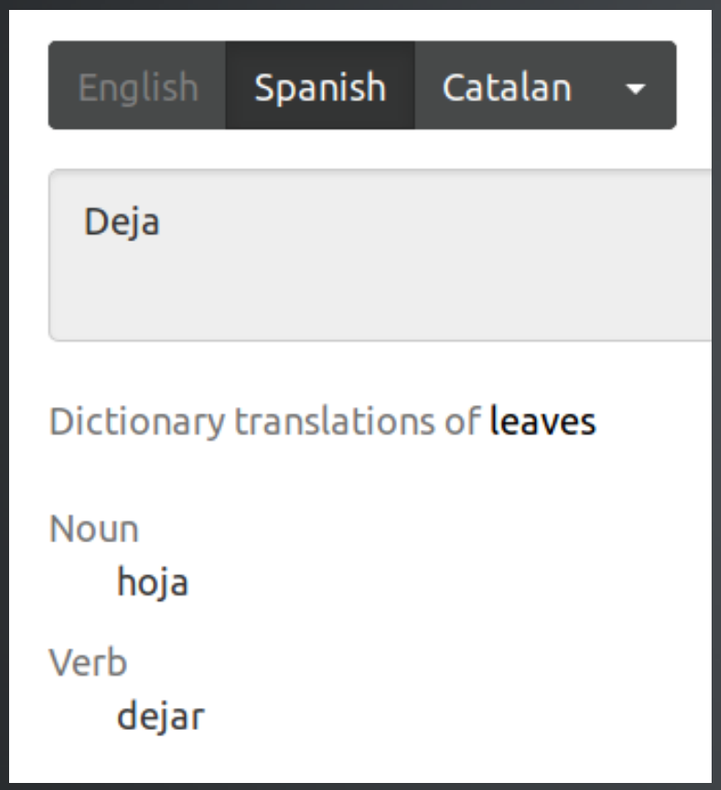
ONGOING WORK

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

DICTIONARY LOOKUP

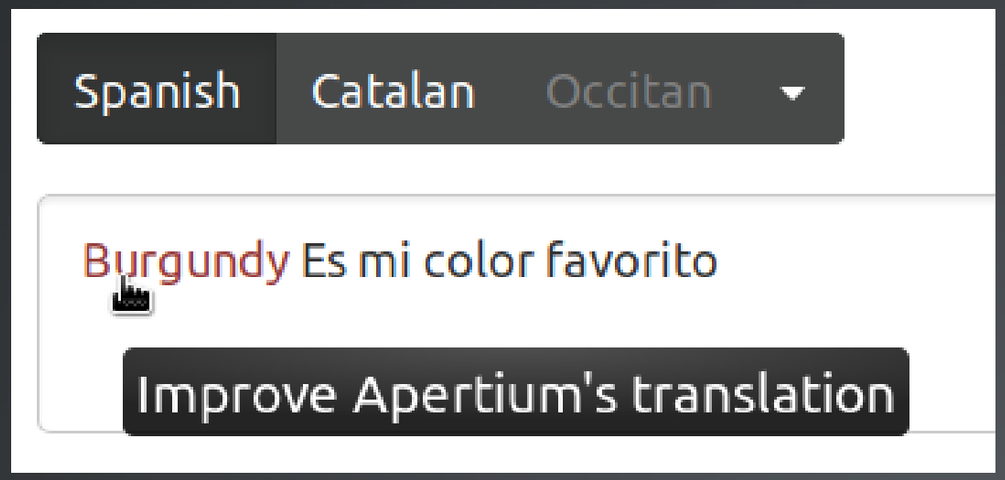


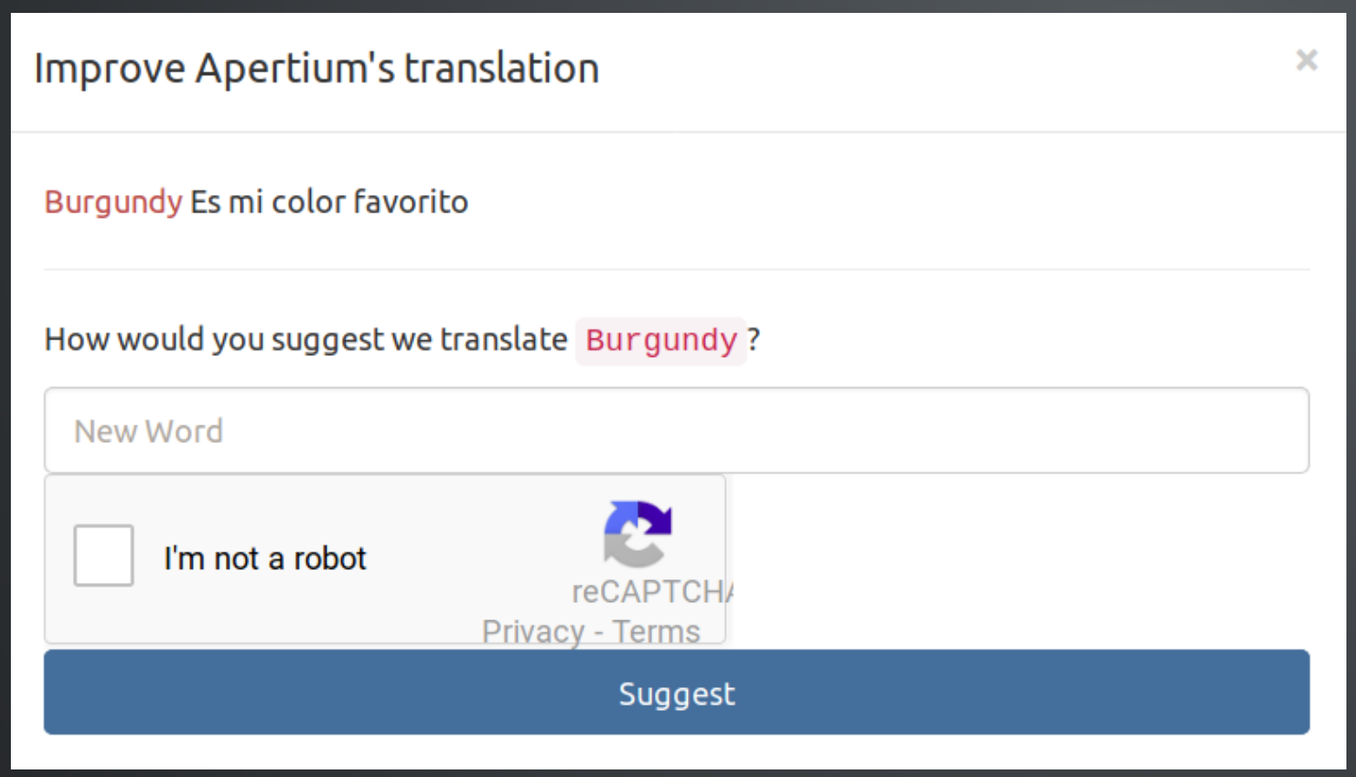
- All possible translations with part of speech
- Future work:
  - Reverse translations for context
  - Morphological subcategory information (Gender, number, conjugation, etc.)
  - Multi-word units

SPELL-CHECKING

- `libvoikko` or `hfst-ospell` create spellers from language modules
- Separate tab in HTML-Tools

SUGGESTIONS





- User can suggest translations for unknown words in context
- Logged by APy, can seed quality improvements
- Rating translations?

USAGE

- From `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`
  - `github.com/apertium/apertium-apy`
- Non-technical contributions: localizations, translation improvements
- **Mailing list:** `apertium-stuff` on SourceForge
- **IRC:** `#apertium` on Freenode
- **Give it a try!**
  `apertium.org`, `turkic.apertium.org`, `beta.apertium.org`

THANKS!