Apertium: A free/open-source platform for machine translation and basic language technology

www.apertium.org

Mikel L. Forcada[1] **Francis M. Tyers**[2,3]

[1]Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03690 Sant Vicent del Raspeig (Spain)
[2]Department of Linguistics, Indiana University, IN 47405, U.S.A.
[3] School of Linguistics, Higher School of Economics, Moscow, Russia
mlf@ua.es, ftyers@iu.edu

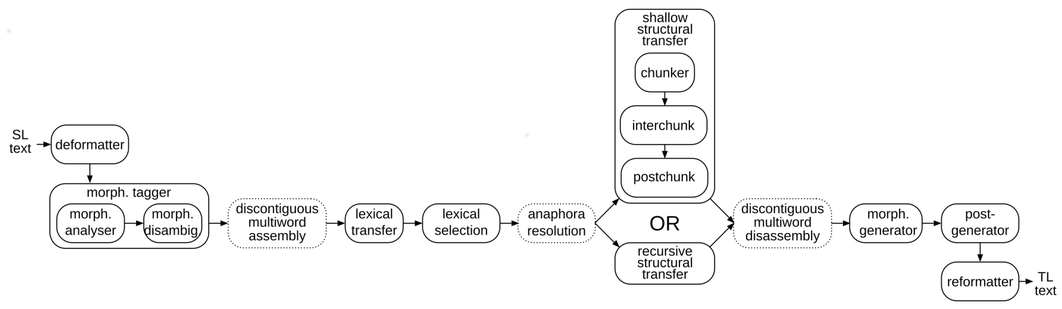
Apertium components

Since 2005, Apertium provides:

1. A free/open-source, modular, shallow/deep transfer, language-independent machine translation *engine* with: text format management, finite-state lexical processing, statistical and constraint-based lexical disambiguation, discontiguous multi-word assembly and disassembly, anaphora resolution, and shallow structural transfer based on finite-state pattern matching.
2. Free/open-source **linguistic data** in well-specified XML formats for a variety of languages and language pairs
3. Free/open-source tools: **compilers** to turn linguistic data into an efficient form used by the engine and software to learn disambiguation or translation rules from corpora.

Interfaces with external technologies: HFST (for some morphologically-rich languages), VISL CG-3 (for rule-based lexical disambiguation).

Machine translation but not only!



- Apertium is a *rule-based machine translation system* but the pipeline contains many **monolingual** modules that can be used for other human-language technology tasks.
- Most modules are based on finite-state technology, with HMMs used for part-of-speech tagging and interpreted in the structural transfer.in the structural transfer.

Licensing: free/open-source

Apertium language data and code are both licensed under the GNU GPL:

- a free/open-source license allowing free distribution of unmodified and modified versions
- a *copylefted* license: it avoids private appropriation and encourages giving improvements back to the project (a *commons*) → **community**

The Apertium community

- Very active group of hundreds of developers
- Wiki documentation (`wiki.apertium.org`)
- Easy entry: Apertium linguistic modelling is simple, no need to program.
- IRC channel `#apertium` in `freenode.net`
- Mailing lists: `apertium-stuff@lists.sf.net` and other lists

Research and business with Apertium

Apertium is already an active research and business platform:

- **Research:** 40+ publications, 2 PhD thesis, 4 master's theses
- **Business:** companies (Prompsit, Eleka, Imaxin Software, etc.) offering services to customers such as Autodesk, the Government of Catalonia, one of the main Basque banks, the daily newspapers *Levante* and *La Voz de Galicia*, etc.)

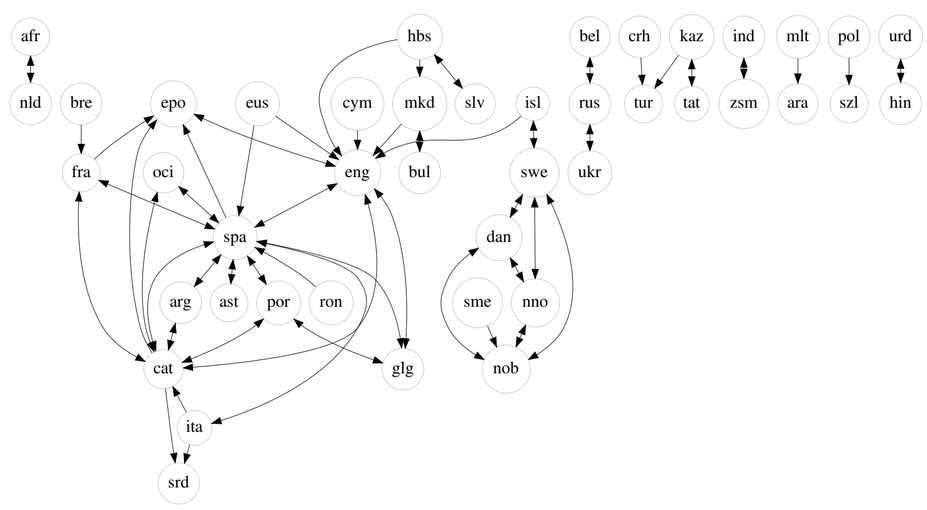
The free/open-source model creates a **community** which effectively connects **researchers**, **developers**, **vendors** and **users**.

Ready-to-use Apertium products

- Available as a PPA repository for GNU/Linux users.
- Stand-alone applications for the desktop: `apertium-caffeine` (Java), *Apertium Simpleton* (Windows, MacOS).
- an Android version for handhelds
- a plug-in for the OmegaT CAT platform `apertium-omegat`

Languages and language pairs

Language data is encoded mostly in XML, but some language pairs contain data encoded in other text-based formats.
Stable language pairs (bilingual data) include:



For many of these languages, there are separate *monolingual* packages.

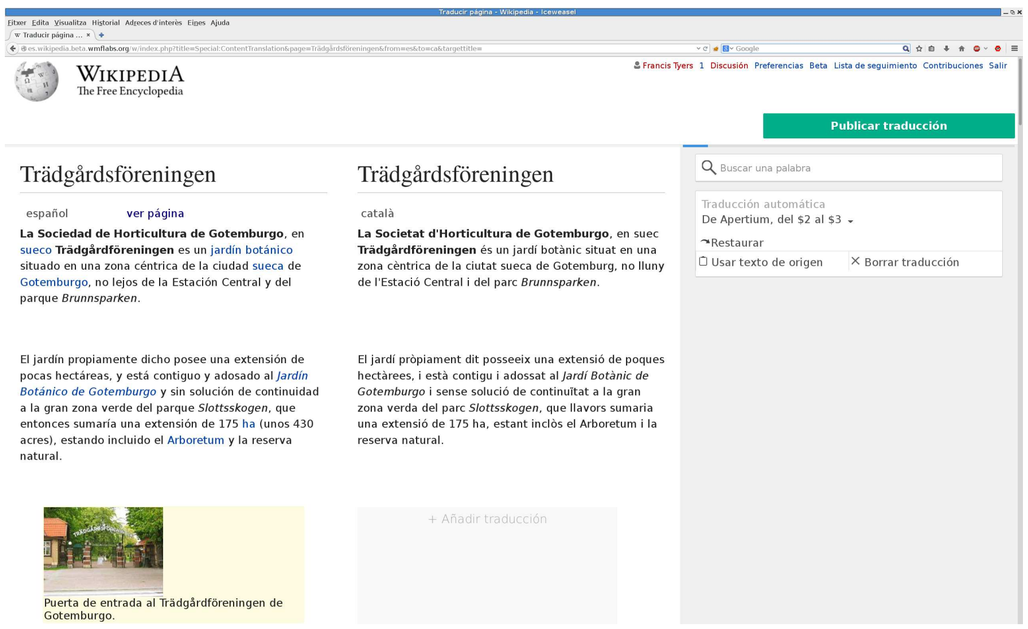
Apertium loves small languages

Some unique MT systems for small languages:

Breton→French
Aragonese↔Spanish
Occitan↔Catalan
Aragonese↔Catalan
Occitan↔Spanish
North Sámi→Norwegian
Crimean Tatar → Turkish

Success cases

Apertium is used:



- in Wikimedia Content Translation to generate Wikipedia content,
- to produce a Catalan edition of Valencia daily newspaper *Levante-EMV*,
- by Universities in the Catalan speaking area to help in the generation of courseware and academic information,
- in PLATA, the Spanish government platform for webpage machine translation.

Funding

- The Ministry of Industry, Tourism and Commerce of Spain (also, the Ministries of Education and Science and of Science and Technology of Spain)
- The Secretariat for Technology and the Information Society of the Government of Catalonia
- The European Commission (Abu-Matran project)
- The Ministry of Foreign Affairs of Romania
- Universitat d'Alacant and Universitat Oberta de Catalunya
- Ofis Publik ar Brezhoneg (Breton Language Board)
- Google Summer of Code scholarships (2009–2014, 2016) and Google Code-In donations.
- Companies: Prompsit Language Engineering, ABC Enciklopedioj, Eleka Ingeniartiza Linguistikoa, imaxin—software, Eolaistriu Technologies, Kaldera, etc.