

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Free/open-source machine translation: the Apertium platform

Mikel L. Forcada^{1,2,3}

¹Centre for Next Generation Localisation, School of Computing, Dublin City University, Dublin 9 (Ireland)

²Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, E-03071 Alacant (Spain)

³Prompsit Language Engineering, S.L., E-03195 Elx (Spain)

June 7, 2010: Translingual Europe 2010, Berlin



Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Contents

- 1 Free/open-source software
- 2 Free/open-source machine translation
- 3 The Apertium platform
- 4 Lots of free/open-source MT out there!

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Free/open-source software
Copyleft
Free/open-source software: open for business

Free/open-source software

Software is **free** (Free Software Foundation, www.fsf.org) when

- 0 anyone can use it for any purpose
- 1 anyone can examine it to see how it works and modify it for any new purpose
- 2 anyone can freely distribute it
- 3 anyone may release an improved version so that everyone benefits

Conditions 1 and 3 require access to the source code, hence the name **open-source** (Open Source Initiative, www.opensource.org).

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Free/open-source software
Copyleft
Free/open-source software: open for business

Copyleft

- **Copyleft** (pun on *copyright*, but still copyright) when added to a free license, means that modifications have to be distributed with the same (copylefted) license.
- **Non-copylefted free/open-source licenses:** The *3-clause* or “Simplified” BSD license, the MIT License, the Apache Software License, the Creative Commons Attribution license, v. 3.0.
- **Copylefted licenses:** *GNU GPL* (General Public License), creative Commons Attribution-Sharealike license, v. 3.0

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Free/open-source software
Copyleft
Free/open-source software: open for business

Copyleft, commons, platforms

- Copyleft secures the existence of a software or knowledge *commons*.
- It protects that commons from private appropriation (incorporation into non-free software): a level field.
- It enables communities of programmers to build shared bodies of free/open-source resources. . .
- . . .by requiring that all derivative work is always distributed under the same license.
- Copyleft may be used to promote the growth of a software *platform*.

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Free/open-source software
Copyleft
Free/open-source software: open for business

Free/open-source software: open for business/1

Free/open-source software opens new business models:

- emphasizes *service-centered* models over traditional *license-centered* models
- customers avoid *vendor lock-in* and may move into *technological partnership* with the provider of their choice

Free/open-source software: open for business/2

Note that third parties may engage in business without permission from the original authors (limited vulnerability from competition):

- Business on copylefted software:
 - Installing, configuring, etc. for a customer.
 - Adapting the code (a price break may be offered in exchange for the right to distribute modifications).
 - Distributing the software and charging for the distribution cost.
- Business on non-copylefted software: all of the above, plus
 - Creating closed-source software (back to license-centered models)

Machine translation software

- Machine translation is special: it strongly depends on data
 - *rule-based* MT (RBMT): dictionaries, rules
 - *corpus-based* MT (CBMT): sentence-aligned parallel text, monolingual corpora
- Three components in every MT system:
 - *Engine* (also *decoder*, *recombinator*...)
 - *Data* (linguistic data, corpora)
 - *Tools* to maintain these data and convert them to the format used by the engine
- For MT to be free/open-source, the **engine**, the **data** and the **tools** must all be free/open-source (NB, in corpus-based MT this includes **corpora**!)

Background: it all started in 1999

Apertium (first release, 2005) is based on the technologies developed by the Transducens group at the Universitat d'Alacant during the development of two existing closed-source systems:

- **interNOSTRUM** (`interNOSTRUM.com`, Spanish–Catalan)
- **Tradutor Universia** (`tradutor.universia.net`, Spanish–Portuguese)

These technologies, initially designed for related-language pairs, have been extended to handle language pairs which are not so related.

The Apertium platform

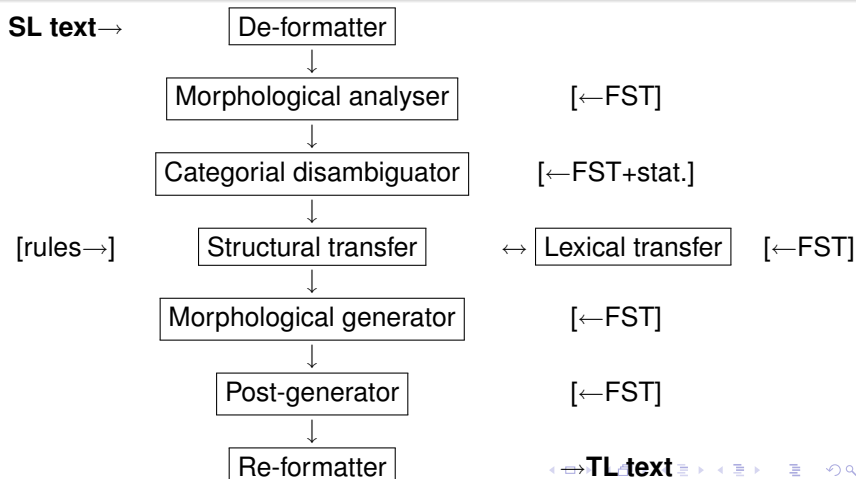
Apertium is a free/open-source machine translation platform (<http://www.apertium.org>) providing:

- 1 A free/open-source, modular, shallow-transfer, language-independent machine translation **engine** with:
 - text format management
 - finite-state lexical processing
 - statistical lexical disambiguation
 - shallow transfer based on finite-state pattern matching
- 2 Free/open-source **linguistic data** in well-specified XML formats for a variety of language pairs
- 3 Free/open-source tools: **compilers** to turn linguistic data into a fast and compact form used by the engine and software to learn disambiguation or translation rules.

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Background
The Apertium platform
The Apertium engine
Language-pair data
Funding
The Apertium community
Research and business with Apertium

The Apertium engine



Language-pair data

The Apertium project hosts the development of a large number of language pairs:

- *Stable* language pairs include: $br \rightarrow fr$, $ca \rightarrow eo$, $ca \leftrightarrow oc$, $cy \rightarrow en$, $es \rightarrow ast$, $en \leftrightarrow ca$, $en \leftrightarrow es$, $en \leftrightarrow gl$, $es \leftrightarrow ca$, $es \rightarrow eo$, $es \leftrightarrow fr$, $es \leftrightarrow gl$, $es \leftrightarrow pt$, $es \leftrightarrow oc$, $eu \rightarrow es$, $is \rightarrow en$, $nn \leftrightarrow nb$, $pt \leftrightarrow ca$, $pt \leftrightarrow gl$, $ro \rightarrow es$
- There is also a growing number of language pairs under development.

Project funding

Funded by

- 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 Ministry of Foreign Affairs of Romania
- Universitat d'Alacant and Universitat Oberta de Catalunya
- The Ofis ar Brezhoneg (Breton Language Board)
- Google Summer of Code scholarships (2009, 2010)
- Companies: Prompsit Language Engineering, ABC Enciklopedioj, Eleka Ingeniartiza Linguistikoa, imaxin|software, etc.

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Background
The Apertium platform
The Apertium engine
Language-pair data
Funding
The Apertium community
Research and business with Apertium

The Apertium community

Not the ideal community development situation, but close.

- Very active group of more than 100 developers in `sf.net/projects/apertium`
- Wiki documentation (`wiki.apertium.org`)
- IRC channel `#apertium` in `freenode.net`
- Mailing list `apertium-stuff@lists.sf.net`
- Stable packages ported to the Debian and Ubuntu GNU/Linux distributions

Research and business with Apertium

Apertium is already an active research and business platform:

- **Research:** 30+ publications, 1 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 newspaper *La Voz de Galicia*, etc.)

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

Free/open-source software
Free/open-source machine translation
The Apertium platform
Lots of free/open-source MT out there!

Lots of free/open-source MT out there!

- For a comprehensive list of free/open-source machine translation (FOSMT) systems, visit: `www.fosmt.info`
- In addition to Apertium, I am involved in OPENMATREX (`www.openmatrex.org`): a free/open-source release of the basic components of Dublin City University's marker-driven example-based machine translation system MATREX

These slides are free/open-source

This work may be distributed under the terms of

- the Creative Commons Attribution–Share Alike license:

`http:`

`//creativecommons.org/licenses/by-sa/3.0/`

- the GNU GPL v. 3.0 License:

`http://www.gnu.org/licenses/gpl.html`

Dual license! E-mail me to get the sources: `mlf@ua.es`