Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

Luis Villarejo*, Sergio Ortiz** and Mireia Ginestí**

*Learning Technologies Office - UOC
**Prompsit Language Engineering

Alacant, November 3d 2009
Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

Outline

1. Apertium for UOC: a use case
2. Developed features
3. Linguistic Data Improvements
4. Early conclusions and further work
Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

Apertium for UOC: a use case

Outline

1. Apertium for UOC: a use case
2. Developed features
3. Linguistic Data Improvements
4. Early conclusions and further work
UOC Linguistic Needs

- 1.167 courses
- 54.000 students
- Online university (written communication)
- 1994 (ca) → 2000 (ca, sp) → 2009 (ca, sp, en, fr)...
- Translation orders from more than 40 different groups inside UOC (Press, Marketing, Web site, Journals, administrative communication...) 
- Linguistic Service managed 4.5 Million words in 2008.

Translations for post-edition on a daily basis, enabling information re-use.
   - Terminology extraction
   - Translation memories
UOC previous MT solution

- Rule based system
  - ca-es
  - few formats: txt, rtf and html
  - stability was an issue
  - proprietary software: less freedom
  - domain adaptation by means of post-edition
  - quite difficult to evolve, we were just another small client

- We decided to look for another MT solution.
Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

Apertium for UOC: a use case

UOC goes Apertium

- Apertium
  - ca-es, ca-en, ca-fr
  - stable and fast
  - wide range of formats
  - translation memories integrated
  - free software: promoting free software, accessibility and interoperability is one of the institution’s core values.

- quality, predictability, traceability, scalability, potential for growth, strong developer and user community...
Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

---

Developed features

Outline

1. Apertium for UOC: a use case
2. Developed features
3. Linguistic Data Improvements
4. Early conclusions and further work
Developed features /1

In the Machine Translation Engine:

- Marking of unknown and ambiguous words.
- Display of more than one translation per word. *La dirección correcta* → \{ *La direcció correcta* — *L’adreça correcta*\}
In the User Interface:

- Text translation
- Translation-as-you-browse
- Document translation: HTML, RTF, ODT, ODS, ODP, SXW, DOC, XSL, PPT, DOCX, PPTX, XLSX, DOCXML, QUARKXPRESS TAGS and XML-UOC.
- First steps towards a filter to translate PDF files
- Advanced HTML translation:
  
  \[ \text{http://www.uoc.edu/masters/esp/web/index.html} \]

  ↓

  \[ \text{http://www.uoc.edu/masters/cat/web/index.html} \]

  Use of Tidy HTML to fix systematic errors in the format
Developed features /3

In the user interface /2:

- Compressed archive translation
  - Keeps internal folder structure.
  - Translates files of known document types.
  - The result is a compressed archive of the same compression format.

- TMX creation: The `apertium-tmxbuild` command creates a bilingual translation memory from each translation carried out inside the service

- Word count functionality
Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

The interface
Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

Linguistic Data Improvements

Outline

1. Apertium for UOC: a use case
2. Developed features
3. Linguistic Data Improvements
4. Early conclusions and further work
Linguistic Data Improvements

Three axes

- Dictionary
- Lexical Disambiguation module
- Structural changes

Sources

- Bilingual corpora: UOC’s web (domain words) and non-UOC sources.
- Pool of documents, originals and translated by linguists (translation errors and tmx repository)
Linguistic Data Improvements /2

- Dictionary improvement.
  - Web corpus: unknown words
  - Documents pool: mistranslations or wrongly disambiguated words

<table>
<thead>
<tr>
<th>Multiwords</th>
<th>Source lang</th>
<th>Previously translated as</th>
<th>New multiword</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ilegar tarde</strong></td>
<td><em>arribar tarda</em></td>
<td>arribar tard</td>
<td></td>
</tr>
<tr>
<td><strong>M’obre</strong></td>
<td><em>me obro</em></td>
<td>me abre</td>
<td></td>
</tr>
<tr>
<td><strong>Cuenta con</strong></td>
<td><em>explica amb</em></td>
<td>compta amb</td>
<td></td>
</tr>
</tbody>
</table>

- Postgeneration errors:
  
  *pares i fills → *padres y hijos → padres e hijos*
Linguistic Data Improvement /3

- Figures on the addition of vocabulary

<table>
<thead>
<tr>
<th>Dictionaries</th>
<th>es</th>
<th>ca</th>
<th>es-ca</th>
</tr>
</thead>
<tbody>
<tr>
<td>New lemmas</td>
<td>4,337</td>
<td>4,371</td>
<td>4,477</td>
</tr>
<tr>
<td>Total lemmas</td>
<td>24,735</td>
<td>24,660</td>
<td>26,662</td>
</tr>
</tbody>
</table>

- Results on the evaluation of the linguistic data improvement

<table>
<thead>
<tr>
<th></th>
<th>es-ca</th>
<th>ca-es</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>After</td>
<td>Before</td>
</tr>
<tr>
<td>Coverage</td>
<td>97.3%</td>
<td>97.5%</td>
</tr>
<tr>
<td>WER</td>
<td>4.86%</td>
<td>3.93%</td>
</tr>
</tbody>
</table>
Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

Early conclusions and further work

Outline

1. Apertium for UOC: a use case
2. Developed features
3. Linguistic Data Improvements
4. Early conclusions and further work
Early conclusions and further work

- Performed a user satisfaction test on the web interface (4.3/5)
- Developments and linguistic data contributed to the Apertium user community.

- Public presentation in December.
- PDF translation through OpenOffice.
- Image pseudo-translation in webs (alt attribute).
- Global campus: English and French improvements
Thanks

Luis Villarejo, Sergio Ortiz and Mireia Ginestí.