

# A Widely Used Machine Translation Service and its Migration to a Free/Open-Source Solution: the Case of Softcatalà

Xavier Ivars-Ribes  
Victor M. Sánchez-Cartagena

II FreeRBMT (Barcelona) January 21, 2011



# Table of Contents

---

- ▶ **Brief History of Softcatalà**
- ▶ **New Machine Translation Service**
- ▶ **Translation Service Usage Analysis**
- ▶ **Using the Crowd to Improve the Data**
- ▶ **Conclusions and Future Work**

# Table of Contents

---

- ▶ **Brief History of Softcatalà**
  - ▶ **The Association**
  - ▶ **The Machine Translation Service**
- ▶ **New Machine Translation Service**
- ▶ **Translation Service Usage Analysis**
- ▶ **Using the Crowd to Improve the Data**
- ▶ **Conclusions and Future Work**

# **Brief History of Softcatalà: the Association**

- ▶ In the 90s, Catalan was missing in ICT context
- ▶ Non-profit association was created in 1998
- ▶ Netscape Navigator was the first translated software
- ▶ Other translations
  - ▶ OpenOffice.org, Mozilla (Firefox & Thunderbird), GIMP, Fedora, Ubuntu, Gnome...
- ▶ Linguistic tools
  - ▶ Term glossary, style guide, translation memory and spell-checker

## **Brief History of Softcatalà: the MT Service**

---

- ▶ Machine translation service available since 2000
- ▶ InterNOSTRUM translation engine
  - ▶ Non-free, funded by *Caja Mediterráneo*
- ▶ Most used service of Softcatalà's website
  - ▶ 70% of 1.2M visits
  - ▶ Translator ↔ Softcatalà
  - ▶ Main source of income (advertisement)
- ▶ Web service physically located at UA

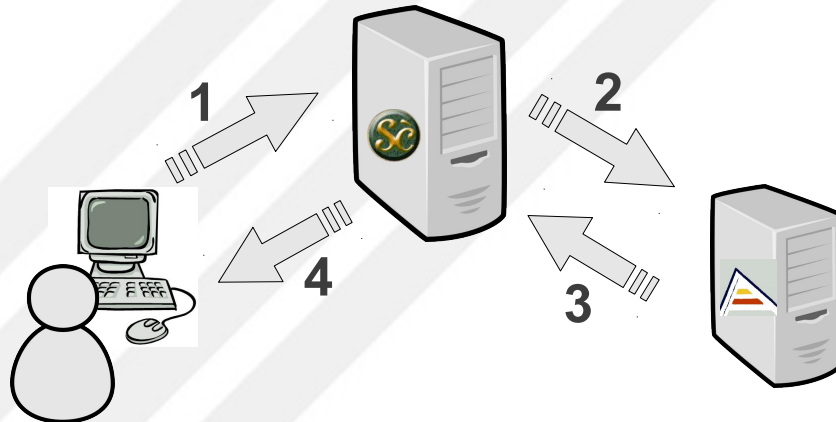
# Table of Contents

---

- ▶ Brief History of Softcatalà
- ▶ **New Machine Translation Service**
  - ▶ Apertium
  - ▶ **ScaleMT**
- ▶ Translation Service Usage Analysis
- ▶ Using the Crowd to Improve the Data
- ▶ Conclusions and Future Work

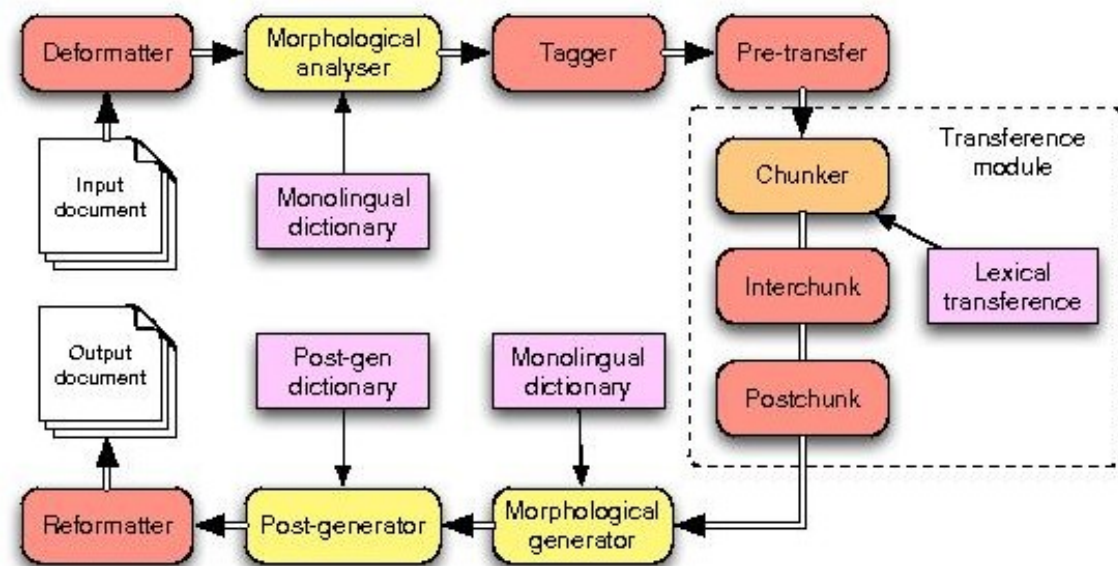
# New Machine Translation Service: Why?

- ▶ **Problems with the previous service**
  - ▶ **Difficult customization and improvement**
  - ▶ **Inability to manage the infrastructure where the service is deployed**



# New MT Service: Apertium

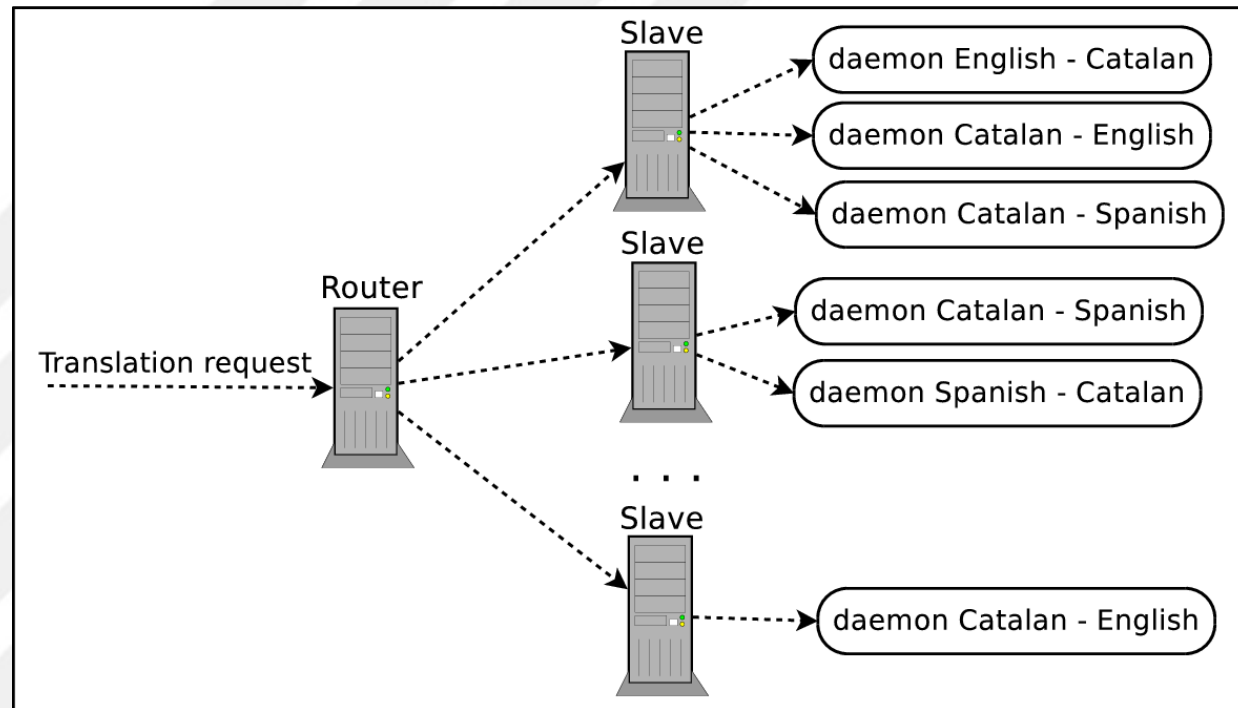
- ▶ interNOSTRUM is Apertium's ancestor
- ▶ Rule-Based Machine Translation Platform
  - ▶ Multiple language pairs supported
- ▶ Language-independent engine
- ▶ Data in XML
- ▶ F/OSS – GPL
- ▶ Pipeline architecture
- ▶ Frequent update






# New MT Service: ScaleMT

- ▶ Framework for building scalable MT services
- ▶ Initially developed through a GSoC grant
- ▶ Translation resources are kept in memory
- ▶ More computers can be added seamlessly
- ▶ F/OSS – AGPL
- ▶ API is compatible with Google Translate



# New MT Service: server status

---

- ▶ Router and a single Slave in the same machine
- ▶ Language pairs installed
  - ▶ Catalan\* ↔ Spanish
  - ▶ Catalan ↔ English 
  - ▶ Catalan ↔ French 
  - ▶ Catalan ↔ Portuguese 

\* Spanish → Catalan can also generate Valencian variant

# Table of Contents

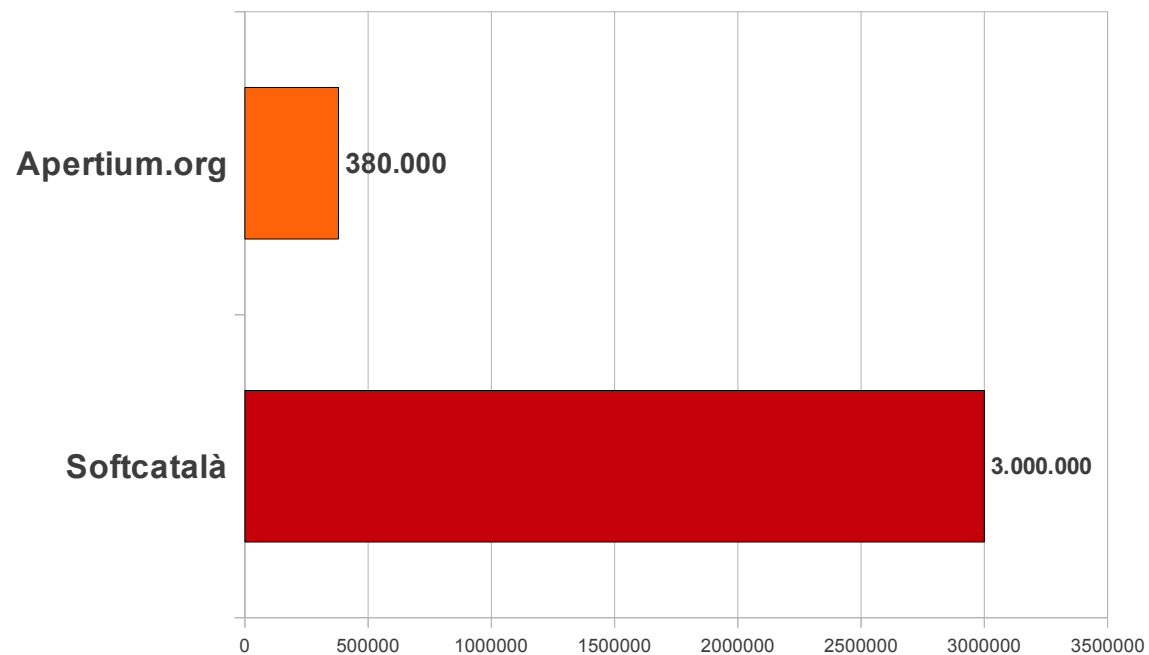
---

- ▶ Brief History of Softcatalà
- ▶ New Machine Translation Service
- ▶ **Translation Service Usage Analysis**
  - ▶ Hourly and Daily Distribution
  - ▶ Impact of the Platform Switch
  - ▶ Language pair distribution
- ▶ Using the Crowd to Improve the Data
- ▶ Conclusions and Future Work

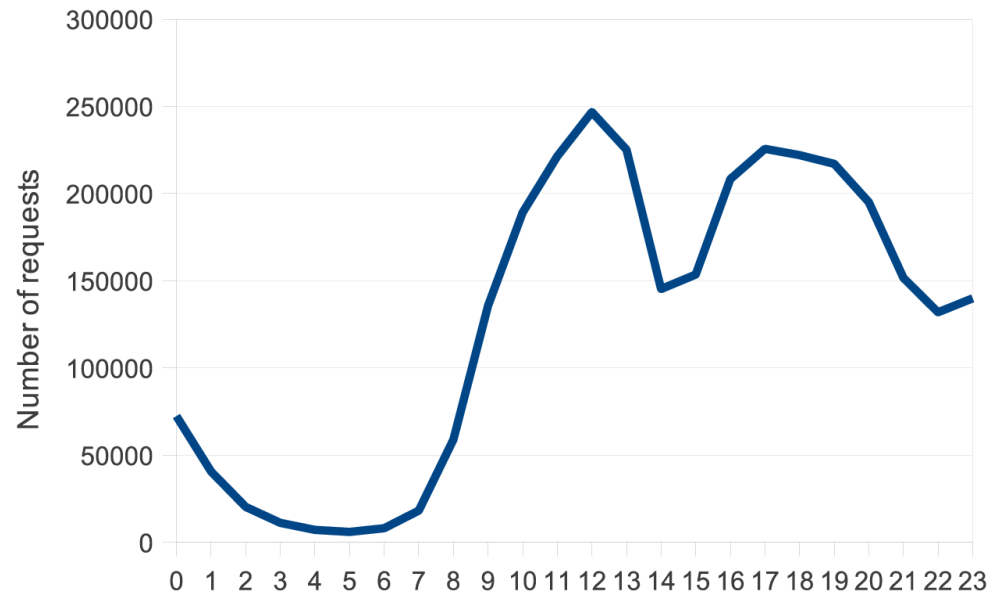
# TS Usage Analysis

---

- ▶ More than 850k monthly visits to the webpage
- ▶ More than 3M monthly translations (9 lang. pairs)
- ▶ *Apertium.org: 380k monthly translations (40 lang. pairs)*



# TS Usage Analysis: Time Distribution



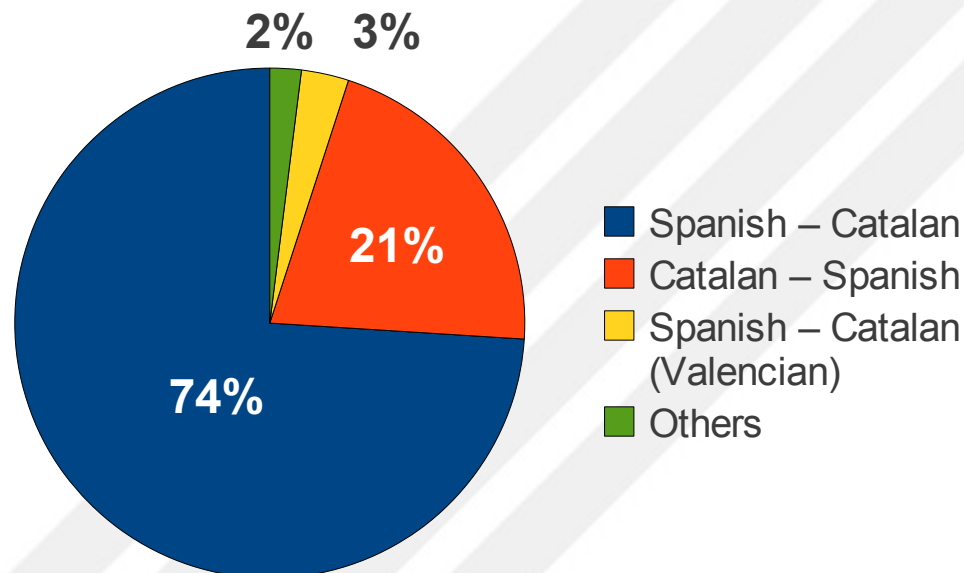
Hourly distribution



Daily distribution

# TS Usage Analysis: Language Pair Distribution

- ▶ Most used pair “Spanish ⇒ Catalan”
  - ▶ TS used for dissemination



Language Pair distribution

# Table of Contents

---

- ▶ Brief History of Softcatalà
- ▶ New Machine Translation Service
- ▶ Translation Service Usage Analysis
- ▶ **Using the Crowd to Improve the Data**
  - ▶ Automatic Unknown Word Extraction
  - ▶ Alternative Translation Suggestions
- ▶ Conclusions and Future Work

# Improvements: Unknown Word Extraction

- ▶ Apertium *pipeline* modification
- ▶ Easy extraction of the most frequent unknown words
- ▶ Examples of extracted unknown words:

<u>es-ca</u>	<u>ca-es</u>	<u>en-ca</u>	<u>ca-en</u>
cortadora	AMPA	nursery	penitenciaris
Sócrates	Moodle	trinity	comanda
Freud	Martini	summertime	incompliment
pH	burret	default	enganxines
estiramiento	perdigot	anymore	Acta



# Improvements: User Suggestions

- ▶ New suggestion form appears after translation is performed
- ▶ Users can send better translations

- ▶ Parallell sentences are saved
- ▶ Web interface to check suggestions

Traductor

Escriu un text en aquest formulari (màxim 4.000 caràcters) i trieu un sentit de la traducció

Texto en castellano.

Sentit de la traducció: castellà » català

Marca les paraules desconegudes  Utilitza les formes valencianes

---

**Traducció**  
selecciona tot el text

Text en castellà.

selecciona tot el text

Teniu alguna traducció millor?

Traductor català ↔ {castellà, anglès, portuguès, francès} basat en la tecnologia d'Apertium i ScaleMT, dissenyats pel Grup Transducens del Departament de Llenguatges i Sistemes Informàtics de la Universitat d'Alacant, i desenvolupats sota les llicències lliures GPL i AGPL, respectivament.

Per a qualsevol suggeriment relacionat amb el traductor, podeu posar-vos en contacte amb nosaltres a [traductor@softcatala.org](mailto:traductor@softcatala.org)

168	es-ca_valencia	Bolsa	Borsa	Bossa	PENDING	<input type="checkbox"/>
169	es-ca_valencia	cañizo	*cañizo	canyís	PENDING	<input type="checkbox"/>
170	es-ca_valencia	solera	*solera	solera	PENDING	<input type="checkbox"/>
171	en-ca	They adopted the Chinese writing system and created excellent bronze swords.	Van adoptar el sistema d'escriptura xinès i espases de bronze excel·lents creades.	Van adoptar el sistema d'escriptura xinès i crearen excel·lents espases de bronze.	PENDING	<input type="checkbox"/>
175	es-ca_valencia	su labor al frente de esta entidad	la seua labor al capdavant d'aquesta entitat	la seua llabor al capdavant d'aquesta entitat	PENDING	<input type="checkbox"/>

# Improvements: User Suggestions

---

- ▶ Some useful feedback
  - ▶ Dictionary improvements with new words
  - ▶ Tagger bug when working with ScaleMT
    - ▶ “Durant molt de temps...” ⇒ “Durando mucho tiempo...”
  - ▶ PoS disambiguation bug
    - ▶ “La sal provoca sed” ⇒ “La sal provoca sigueu”
    - ▶ *Forbid* rules added to the tagger solved the problem

```
<label-sequence>
  <label-item label="VLEXIMP"/>
  <label-item label="VSERIMP"/>
</label-sequence>
[...]
<label-sequence>
  <label-item label="VLEXPFCI"/><!-- provoca sed-->
  <label-item label="VSERIMP"/>
</label-sequence>
```

# Table of Contents

---

- ▶ Brief History of Softcatalà
- ▶ New Machine Translation Service
- ▶ Translation Service Usage Analysis
- ▶ Using the Crowd to Improve the Data
- ▶ **Conclusions and Future Work**

# Conclusions

---

- ▶ **Up-to-date and more stable MT system**
- ▶ **Control over its deployment**
- ▶ **System improves after user suggestions**
  - ▶ **Updated MT data is available to the community**
  - ▶ **Active users will notice a stronger improvement**

# Future Work

---

- ▶ **Improve suggestion web interface**
  - ▶ Show MT pipeline to make debug easier
  - ▶ Combine unknown-words extractor, remove repeated suggestions, email pair maintainers, etc.
- ▶ **Create mobile applications using the web service API**
  - ▶ iPhone and Meego apps developed, being tested
  - ▶ Android app in development

**Moltes gràcies!**

**Thank you very much!**

[xavier.ivars@ua.es](mailto:xavier.ivars@ua.es)



# License and Contact

---

- ▶ This presentation may be distributed under the terms of any of the following licenses
  - ▶ GNU GPL v. 3.0
    - ▶ <http://www.gnu.org/licenses/gpl.html>
  - ▶ GNU FDL v. 1.2
    - ▶ <http://www.gnu.org/licenses/gfdl.html>
  - ▶ CC-BY-SA v. 3.0
    - ▶ <http://creativecommons.org/licenses/by-sa/3.0/>
- ▶ You can contact us
  - ▶ Xavier Ivars-Ribes: [xavier.ivars@ua.es](mailto:xavier.ivars@ua.es)
  - ▶ Víctor M. Sánchez-Cartagena: [vmsanchez@dlsi.ua.es](mailto:vmsanchez@dlsi.ua.es)