

Development of a free Basque to Spanish machine translation system

Mireia Ginestí-Rosell,¹ Gema Ramírez-Sánchez,¹
Sergio Ortiz-Rojas,¹ Francis M. Tyers,² and Mikel L. Forcada²

¹Prompsit Language Engineering,
Avinguda Sant Francesc 74, 1-L. E-03195 L'Altet - Elx (Spain)

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

XXV Congreso de la SEPLN, Donostia, 8th September 2009

Contents

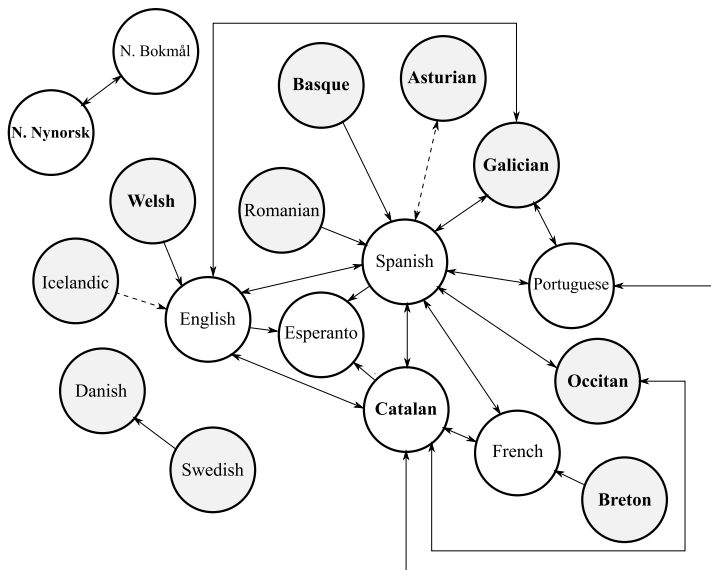
- 1 Introduction
- 2 Development
- 3 Evaluation
- 4 Discussion

What is Apertium

So what is Apertium?

- GPL-licensed platform for machine translation
- Modular – made up of stand-alone programs which communicate in plain text using Unix pipes
- Developed by universities, companies and independent developers
- 21 available “stable” language pairs
- More in development
- Language data available and in development for many other languages. . . Icelandic, Asturian, Bengali, . . .

Stable language pairs

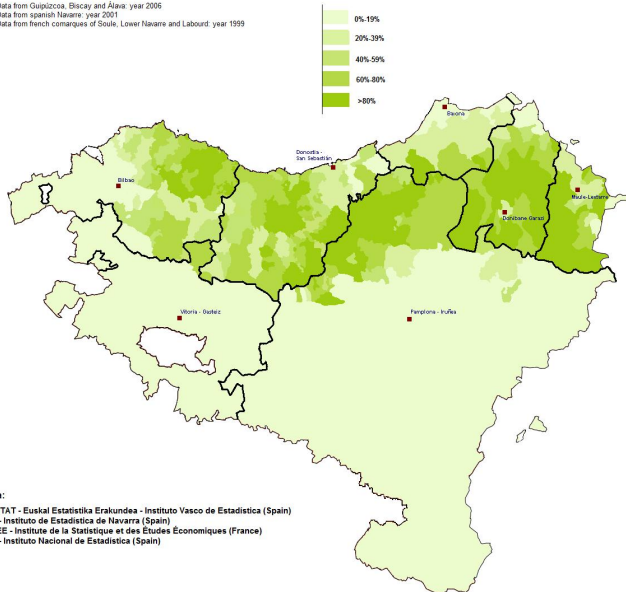


Percentage of basque speakers as initial language by municipalities in the Autonomous Community of the Basque Country, Foral Autonomous Community of Navarre (Upper Spanish Navarre) and french *comarques* of Soule valleys, Lower Navarre and Labourd

Data from Guipúzcoa, Biscay and Álava: year 2006

Data from spanish Navarre: year 2001

Data from french *comarques* of Soule, Lower Navarre and Labourd: year 1999



Data:

EUSTAT - Euskal Estatistika Erakundea - Instituto Vasco de Estadística (Spain)

IEN - Instituto de Estadística de Navarra (Spain)

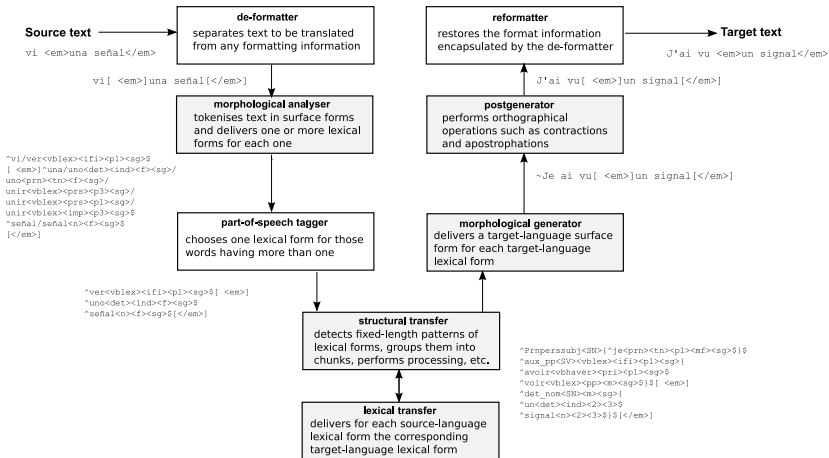
INSEE - Institut de la Statistique et des Etudes Économiques (France)

INE - Instituto Nacional de Estadística (Spain)

The Apertium MT platform is of the shallow-transfer type. Development consists of creating the following language data:

- Morphological dictionaries (analysis / generation)
- Disambiguation rules and training statistical tagger (including optional target-language training)
- Bilingual dictionary (lexical transfer)
- Shallow syntactic transfer rules
 - Local re-ordering (nom adj \rightarrow adj nom)
 - Chunking (adj adj nom \rightarrow SN[adj adj nom])
 - Insertions, deletions and substitutions of lexical units and chunks

Translation model



Existing data (Basque→Spanish)

We were able to directly use:

- Spanish morphological dictionary from the `apertium-es-ca` language pair (GPL licensed)

And the following after conversion:

- Basque–Spanish bilingual dictionary from Matxin¹ (GPL licensed)
- Monolingual Basque dictionary from Matxin (GPL licensed)

The conversion process involved changing tags and trimming down the morphological dictionary of Basque.

¹<http://matxin.sourceforge.net>

Transfer rules

CHUNKER

```
<rule comment="NOM + DET + POST (liburuko)">
  <pattern>
    <pattern-item n="nom"/>
    <pattern-item n="det"/>
    <pattern-item n="post"/>
  </pattern>
  <action>
    <call-macro n="SPgenitivo">
      <with-param pos="4"/>
    </call-macro>
    (...)
  <out>
    <chunk name="pr_det_nom">
      <tags>
        <tag><var n="chunk"/></tag>
        <tag><var n="tipusdet"/></tag>
        <tag><var n="gen_chunk"/></tag>
        <tag><var n="nbr_chunk"/></tag>
      </tag>
      <clip pos="3" side="t1" part="whole"/>
      <b/>
      <clip pos="2" side="t1" part="whole"/>
      <b/>
      <clip pos="1" side="t1" part="whole"/>
    </chunk>
  </out>
</action>
</rule>
```

INTERCHUNK

```
<rule comment="SPGEN + SN">
  <!-- [liburuko] [lehen ipuina] -->
  <pattern>
    <pattern-item n="SPGEN"/>
    <pattern-item n="SN"/>
  </pattern>
  <action>
    (...)
  <out>
    <chunk>
      <clip pos="2" part="whole"/>
    </chunk>
    <b pos="1"/>
    <chunk>
      <clip pos="1" part="whole"/>
    </chunk>
  </out>
</action>
</rule>
```

Module	Number
Lexicon	6,300 lemmas
Disambiguation	56 rules
Chunk	175 rules
Inter-chunk	54 rules

Table: Statistics from current SVN revision #14997

Corpus	Running tokens	Tokens found	Coverage
Wikipedia ²	2,531,313	1,958,836	77.38%
Berria ³	3,665,880	3,335,363	90.98%

Table: Coverage statistics from current SVN revision #14997

²<http://eu.wikipedia.org>

³<http://www.berria.info>

We took two main approaches to evaluation.

- **Quantitative** – To be comparable with other systems, and provide a useful “at a glance” measure of quality. This was done for both
 - **adequacy** (*assimilation* setting)
 - and **post-editing word error rate** (*dissemination* setting)

- **Qualitative** – To give a better idea of where the strengths and weaknesses of the system are

Two corpora were used for quantitative evaluation in *assimilation* and *dissemination* settings.

- **Berria** – 50 sentences (703 words) selected at random (*assimilation*)
 - Translated, along with a sentence of context either side
 - Translation and original given to monolingual reviewer
 - Monolingual reviewer attempted **blind** post-editing
 - Bilingual speaker scores (0%–100%) how adequate the post-edited sentence is.
- **Berria*** – separate set of 100 sentences (*dissemination*)
 - machine-translated
 - post-edited
 - WER, PER of raw translations calculated against post-edited translations

20.

- = EHUK eta Eusko Jaurlaritzak finantziario akordioa lortu dute
- + EHUK euskal presoei ikasketak ematen segitu nahi duela adierazi du Montero errektoreak
- = EHUK martxoaren 24an egingo ditu errektore berria aukeratzeko hauteskundeak

20.

= UPV y el Gobierno vasco el acuerdo de financiación han conseguido

+ UPV a los presos vascos los estudios dando seguir quiere , ha expresado *Montero el rector

= UPV en 24 del marzo hará las elecciones para elegir el rector nuevo

20.

- = UPV y el Gobierno vasco el acuerdo de financiación han conseguido
- + UPV a los presos vascos los estudios dando seguir quiere , ha expresado *Montero el rector
- + UPV quiere seguir dando a los presos vascos estudios, según ha expresado el rector Montero
- = UPV en 24 del marzo hará las elecciones para elegir el rector nuevo

20.

- = UPV y el Gobierno vasco el acuerdo de financiación han conseguido
- + UPV a los presos vascos los estudios dando seguir quiere , ha expresado *Montero el rector
- + UPV quiere seguir dando a los presos vascos estudios, según ha expresado el rector Montero
- ++ La UPV quiere seguir ofreciendo a los presos vascos estudios, según ha expresado el rector Montero
- = UPV en 24 del marzo hará las elecciones para elegir el rector nuevo
- == Pretty good (90%)

100%

- Src. Errenteriko alkatearen kontrako pintadak egin dituzte
Trg. *Errenteriko a la residencia de PSE-EE fuego le han dado
Post. **Le han prendido fuego a la sede del PSE-EE en Errenteria**
Corr. Le han prendido fuego a la sede del PSE-EE en Errenteria

90%

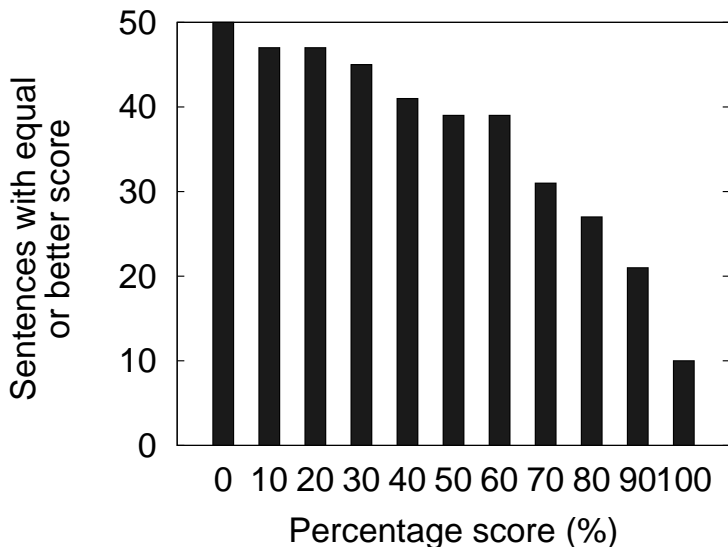
- Src. EHUK euskal presoek ikasketak ematen segitu nahi duela adierazi du Montero errektoreak
- Trg. UPV a los presos vascos los estudios dando seguir quiere , ha expresado *Montero el rector
- Post. **UPV quiere seguir dando a los presos vascos estudios, según ha expresado el rector Montero**
- Corr. La UPV quiere seguir ofreciendo a los presos vascos estudios, según ha expresado el rector Montero

80%

- Src. EAJ eta PSOE bilduta egon dira Madrilen
- Trg. PNV y PSOE reunido han estado en Madrid
- Post. **PNV y PSOE se han reunido en Madrid**
- Corr. PNV y PSOE han estado reunidos en Madrid

70%

- Src. Hori leporatuta, Espainiako Auzitegi Nazionalak zazpi urte eta sei hilabeteko zigorra jarria dio, eta Ertzaintza lehengo astean haren bila joan zen.
- Trg. Ese responsabilizado, el Tribunal Nacional de España siete años y la sanción de los seis meses puesto dice, y la Ertzaintza en la semana anterior en busca de aquel fue.
- Post. **Este fue hallado culpable, el Tribunal Nacional de España le ha puesto una sanción de siete años y seis meses, y la Ertzaintza fue en su búsqueda la semana pasada.**
- Corr. Acusado de ello, la Audiencia Nacional de España le ha impuesto una pena de siete años y seis meses, y la Ertzaintza fue en su búsqueda la semana pasada.



The results of the 100 sentence post-editing task:

	WER	PER
Berria*	72.41%	39.86%

Table: WER and PER scores for post-editing task

Obviously not suitable for post-editing ($\leq 15\%$ WER required), but over 60% of correct words in word-for-word translation is a good start.

Study of the main sources of errors:


- Incomplete lexical coverage (Basque words not translated)
- In particular, absence of frequent multiword units: *Audiencia Nacional* (*Auzitegi Nazionala* 'National Court') or *quedar en suspenso* (*bertan behera gelditu* 'to be suspended')
- Lack of a mechanism to deal generally with inflected proper nouns.
- Errors in lexical selection: *elkarrizketa* → *diálogo* 'dialogue' instead of → *entrevista* 'interview'
- Incomplete handling of compound ("periphrastic") verbs by the system.
- Phrases with co-ordination sometimes become "mangled":
 - 'Zazpi urte eta sei hilabeteko zigorra'
 - Siete años y seis meses-GEN sanción
 - Siete años y la sanción de seis meses
 - Una pena de siete años y seis meses

Berria - Iceweasel

Fitxer Editatu Bisualizatu Historiala Adresez d'interès Eines Ajuda

http://xixona.dlsi.ua.es/~fran/basque/berria/berria.php

Google



Thu, 03 Sep 2009 11:50:41 +0200 ([link](#))

Peruko armadako bi kide hil dituzte, helikopteroan zihozela

Atzo Perun hiru militar zauritu egin zituzten atentatuan Acobamban herrialderan erdialdean dago-, eta gaur haiek laguntzera helikopteroan zihozzen militarrek ere eraso jaso dute eta bi hil egin dira, Rafael Rey Peruko Gobernuako Defentsa ministroak jakinarazi duenez. Sendero Luminoso erakundeari egotzi diote atentatua.

Helikopteroako pilotoa da hildako haietako bat eta haren laguntzailea bestea. Eddie Pasquel Alfaro armadako komandantea ere bazioan haiekin eta lepauztaia hautsi du.

Lurreratzear zirela izan da atentatua. Reyk adierazi du egileek ahalmen luzeko armamentua erabili dutela.

Astelehenean beste militar zauritu zituzten.

dos miembros del ejército del Perú han muerto, en el helicóptero iban ,

Ayer *Perun tres militares hirieron en el atentado *Acobamban *herrialderan en el centro está-, y hoy ellos a acompañar en el helicóptero iban los militares también atacado han recogido y dos han muerto, *Rafael *Rey el ministro de Defensa del Gobierno del Perú según ha dado a conocer. *Sendero *Luminoso Al organismo le han acusado el atentado.

Del helicóptero *pilotoa es un de aquellas víctimas y su ayudante el otro. *Eddie *Pasquel *Alfaro El comandante del ejército también si iba con aquellos y *lepauztaia ha roto.

A punto de tomar tierra eran , ha sido el atentado. *Reyk Ha expresado los autores el armamento de la facultad larga han utilizado ,.

En el lunes otro militar hirieron.

Fet

Why not corpus-based MT?

But wouldn't it be quicker to use corpus-based MT?

- No wide-coverage freely available corpus of Basque–Spanish
- Little chance of finding one – most text is not free
- In this case existing GPL linguistic data was available
- No existing systems – opportunity to do something new and interesting

Creating an RBMT system also involves creating useful linguistic tools which can be used by other approaches to MT (e.g. SMT) and other linguistic software.

For the Apertium platform in general:

- Implementation of a full parser
- Improvement of lexical selection
- More language pairs ... working on es-ast, es-it, is-en
- Increase ease of contribution

For Basque in particular:

- Increase vocabulary coverage
- Improve rule coverage (possibly add Constraint Grammar)
- Other translators with other languages – e.g.
 - English
(widely spoken language)
 - French, Catalan, Occitan, Galician, Breton
(languages of Spain and France)

Try it out!

The system can be tested online for text at:

```
http://www.erdaratu.eu
```

The source code can be downloaded from:

```
http://www.apertium.org
```

And we welcome your comments at:

```
apertium-stuff@lists.sourceforge.net
```

Thanks

```
$ echo "Eskerrik asko" | apertium eu-es
```

Muchas gracias