apertium-cy: A collaboratively-developed RBMT system for Welsh to English

Francis M. Tyers\textsuperscript{1,2} and Kevin Donnelly\textsuperscript{3}

\textsuperscript{1}Departament de Llenguatges i Sistemes Informàtics, Universitat d’Alacant, E-03071 Alacant (Spain)
\textsuperscript{2}Prompsit Language Engineering, Avinguda Sant Francesc 74, 1-L. E-03195 L’Altet - Elx (Spain)
\textsuperscript{3}Llanfairpwll, Ynys Môn, LL61 6UX (Wales)

16th January 2009
So what is Apertium?

- GPL-licensed platform for machine translation
- Modular – made up of stand-alone programs which communicate using Unix pipes
- Developed by universities, companies and independent developers
- 17 available “stable” language pairs
- More in development
  - And language data in development for many other languages... Breton, Icelandic, Hindi, ...
Apertium is a shallow-transfer MT system meaning development consists of:

- 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 ($\text{nom adj} \rightarrow \text{adj nom}$)
  - Chunking ($\text{adj adj nom} \rightarrow \text{SN[adj adj nom]}$)
  - Insertions, deletions and substitutions of lexical units and chunks
Existing data (Welsh$\rightarrow$English)

We were able to directly use:

- English morphological dictionary from the apertium-en-ca language pair

And the following after conversion:

- English–Welsh bilingual dictionary from Eurfa\(^1\)
- Monolingual Welsh dictionary from Eurfa and Konjugator\(^2\)
- Some rules from the English–Spanish pair

\(^1\)http://www.eurfa.org.uk
\(^2\)http://www.konjugator.org.uk
Collaboration

Development was performed collaboratively over the internet.

- Wiki and E-mail – For describing transfer and disambiguation rules
- SVN – For version control
- IRC and E-mail – For technical support
(1.3.39) Periphrastic tenses with "bod"

Re comments in 1.3.32 above:

For Welsh pattern "VBGER.pres + subject + [qualifiers] + yn + verb.infin"
output English "subject + [qualifiers] + is/are + verb + -ing"

mae'r bachgen yn mynd -> The boy is going

Fine - the equivalent of 1.3.8 above, and it seems to be catered for already.

For Welsh pattern "VBGER.pii + subject + [qualifiers] + yn + verb.infin"
output English "subject + [qualifiers] + was/were + verb + -ing"

roedd y bachgen yn mynd -> The boy was going

Fine again.

For Welsh pattern "VBGER.pres + subject + [qualifiers] + wedi + verb.infin"
output English "subject + [qualifiers] + has/have + verb.pp"

mae'r bachgen wedi mynd -> *The boy is after go - the boy has gone

Done. - Francis Tyers
Transfer rules

“In Welsh, if a noun is followed by the definite article followed by another noun, then output in English definite article, then the first noun, then the preposition ‘of’ then the definite article followed by the second noun.”
### Status

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon</td>
<td>10,994</td>
<td>90.1%</td>
</tr>
<tr>
<td>Disambiguation</td>
<td>56</td>
<td>-</td>
</tr>
<tr>
<td>Chunk</td>
<td>84</td>
<td>-</td>
</tr>
<tr>
<td>Inter-chunk</td>
<td>36</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table:** Statistics from current SVN revision #8140
This page allows you to test the pre-alpha (from SVN) versions of Apertium data. These data will probably not work more often than not. Often the data are woefully incomplete and they are not supported in any way.

Roedd y cwningod yn hapus

Roedd/Bod\textltproc\p{i}{p}{v}{s}{e}{r}{p}{i}{i}{p}{3}{s}{g}{y}{y}{r}{v}{p}{a}{r}{t}\textaff/\y{r}\d{e}{t}\d{e}{f}\s{p}{p}\ltproc\cwningod/cwningen\n{c}{f}{p}{l}\yn\yn\pr\hapus/hapus\adj\cgproc\hapus/hapus\adj\apertium-tagger\apertium-pretransfer\apertium-transfer\apertium-interchunk\apertium-postchunk\ltproc\ltproc\ltproc

The rabbits were happy
The rabbits were happy
The rabbits were happy
We took two main approaches to evaluation.

- **Quantitative** – To be comparable with other systems, and provide a useful “at a glance” measure of quality.

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

We also made some tests as to how the output compared with other available systems.
Two corpora were used for quantitative evaluation.

- **Wikipedia** – 318 sentences (5,492 words) selected at random from the Welsh Wikipedia, translated, post-edited then WER, PER, BLEU calculated against post-edited translations.

- **PNAW** – 50,000 sentences selected at random from a bilingual parallel corpus translated and then WER, PER, BLEU calculated against reference translations.

---

3 Proceedings of the National Assembly for Wales
<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>PER</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia true-case</td>
<td>55.78</td>
<td>30.59</td>
<td>30.70</td>
</tr>
<tr>
<td>Wikipedia lowercased</td>
<td>53.40</td>
<td>27.22</td>
<td>32.21</td>
</tr>
<tr>
<td>PNAW true-case</td>
<td>65.99</td>
<td>35.44</td>
<td>15.12</td>
</tr>
<tr>
<td>PNAW lowercased</td>
<td>64.94</td>
<td>34.35</td>
<td>15.68</td>
</tr>
</tbody>
</table>

**Table:** WER, PER and BLEU metrics for the two corpora
In general terms:

- Short sentences or long sentences made up of sequential parts work reasonably well

- Sentences with marked formations, long multiword units or subordinate clauses often come across “mangled”
  - Pam mae’r bocs yn wag? → *Why the box is empty?
  - Aeth y dyn i ffonio’r heddlu pan welodd y ddamwain. → *The man went to phone the police when the accident saw.
  - Gwasanaeth Tân ac Achub Canolbarth a Gorllewin Cymru → *Fire service and Save Central region and Wales West

- Word-choice is often “unusual”, but rarely “ridiculous”
  - Llai na’r cyfradd chwyddiant → *Smaller than the rate inflation
We also did some brief comparative tests with InterTran\textsuperscript{4} and default Moses configuration trained on the PNAW corpus.\textsuperscript{5}

or \textit{Roedd y Comisiwn yn ymchwilio i’r honiadau bod yr AS wedi methu datgan £103,000 o roddion.}

in He was the Commission crookedly ymchwiliad I ’ group claims be he drives ACE has failed declare he gifts.

mo The comission to investigate the allegation that the MP has failed to declare £103,000 of roddion.

ap \textit{The Commission was investigating the allegations that the MP has failed to declare £103,000 gifts.}

Moses and Apertium can both be tested side-by-side online.\textsuperscript{6}

\begin{itemize}
\item \textsuperscript{4}http://www.tranexp.com/
\item \textsuperscript{5}http://xixona.dlsi.ua.es/corpora/
\item \textsuperscript{6}http://elx.dlsi.ua.es/~fran/welsh/
\end{itemize}
Cynllun i greu cymuned Aeleg

Mae ymgyrch wedi dechrau i geisio creu cymuned Aeleg yn Ucheldiroedd Yr Alban.

Dyweddodd ymgrych wyr y byddai cymuned o fwy na 1000 o siaradwyr Gaelig yn "hwb enfawr" i’r iathi.

Mae nifer y siaradwyr wedi lleihau’n fawr yn y blynyddoedd diwethaf, cyfanswm o 58,652 yn ôl Cyfrifiad 2001, a’r mwyaf daf yn byw yn yr Ucheldiroedd a’r ymwysoedd.

Eisoes mae 30 o bobl wedi ymrywmo i fyw yn y pentref ac mae’r ymgrych wyr am gynnal cyfarfod yn Inverness ym mis Rhagfyr.

Dim ‘cyfle cymdeithasol’

Mae dysgu’r Aeleg ar gynnydd mewn ysgolion.
Plan to create Scots Gaelic community

campaign has begun to try create Scots Gaelic community in Scotland Highlands.

campaigners Said community would be he bigger than 1000 speakers Scots Gaelic in "enormous push" to the language.

the number of the speakers have diminished big in the last years, total of 58,652 according to Census 2001, and the majority in life in the Highlands and the islands.

Already 30 peoples have committed oneself to a life in the village and the campaigners are going to hold a meeting in *Inverness in December.

Not 'a social opportunity'
Why not corpus-based MT?

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

- No wide-coverage freely available corpus of Welsh–English
- Little chance of finding one – most text is not free
- In this case existing GPL linguistic data was available
- Interested linguist – use available talent

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.
Cynllun i greu cymuned Aeleg

Mae ymgyrch wedi dechrau i geisio creu cymuned Aeleg yn Ucheldiroedd Yr Alban.

Dywedodd ymgyrchwyr y byddai cymuned o fwy na 1000 o siaradwyr Gaeg yn hwb enfawr i'r iathi.

Mae nifer y siaradwyr wedi lleihau'n fawr yn y blynyddoedd diwethaf, cyfanswm o 58,652 yn ôl Cyfrifiad 2001, a'r mwyaf o'r byw yn yr Ucheldiroedd a'r ynysoedd.

Eisoes mae 30 o bobl wedi ymrwymo i fyw yn y pentref ac mae 'r ymgyrchwyr am gynnal cyfarfod yn Inverness ym mis Rhagfyr.

Dim cyfle cymdeithasol

Mae dysgu 'r Aeleg ar gynnydd mewn ysgolion.
Future work

For the Apertium platform in general:

- Implementation of a full parser
- Improvement of lexical selection
- Expansion of number of language pairs ... working on es-ast, es-it, br-fr
- Increase ease of contribution

For apertium-cy in particular:

- Reverse the direction (cy→en) – the pitfalls of Scymraeg
- Increase vocabulary coverage
- Other translators with Welsh – e.g. Spanish (Patagonia), Breton (related)
Diolch / Thanks!