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

Francis M. Tyers[1,2] and Kevin Donnelly[3]

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

16th January 2009

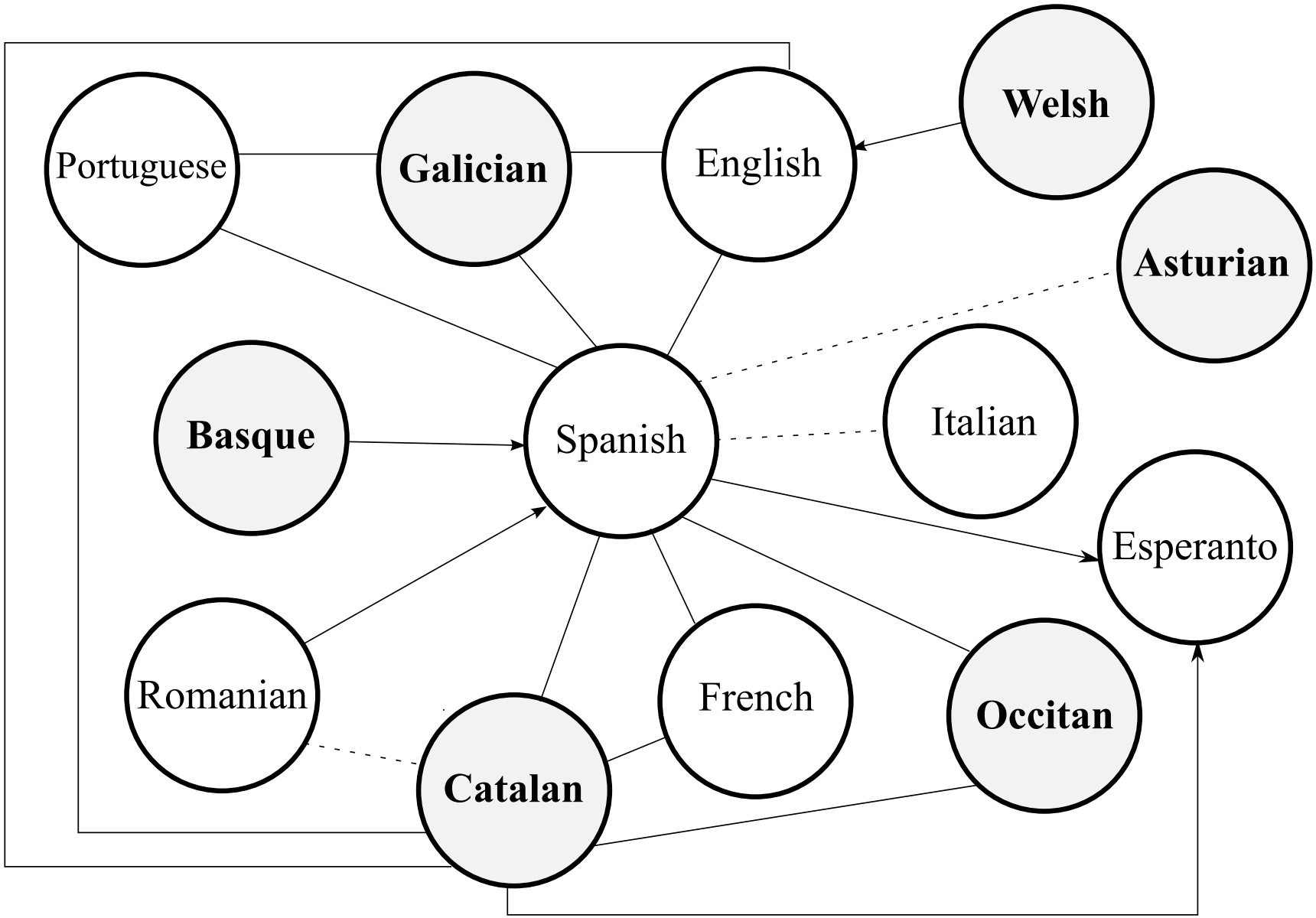
Contents

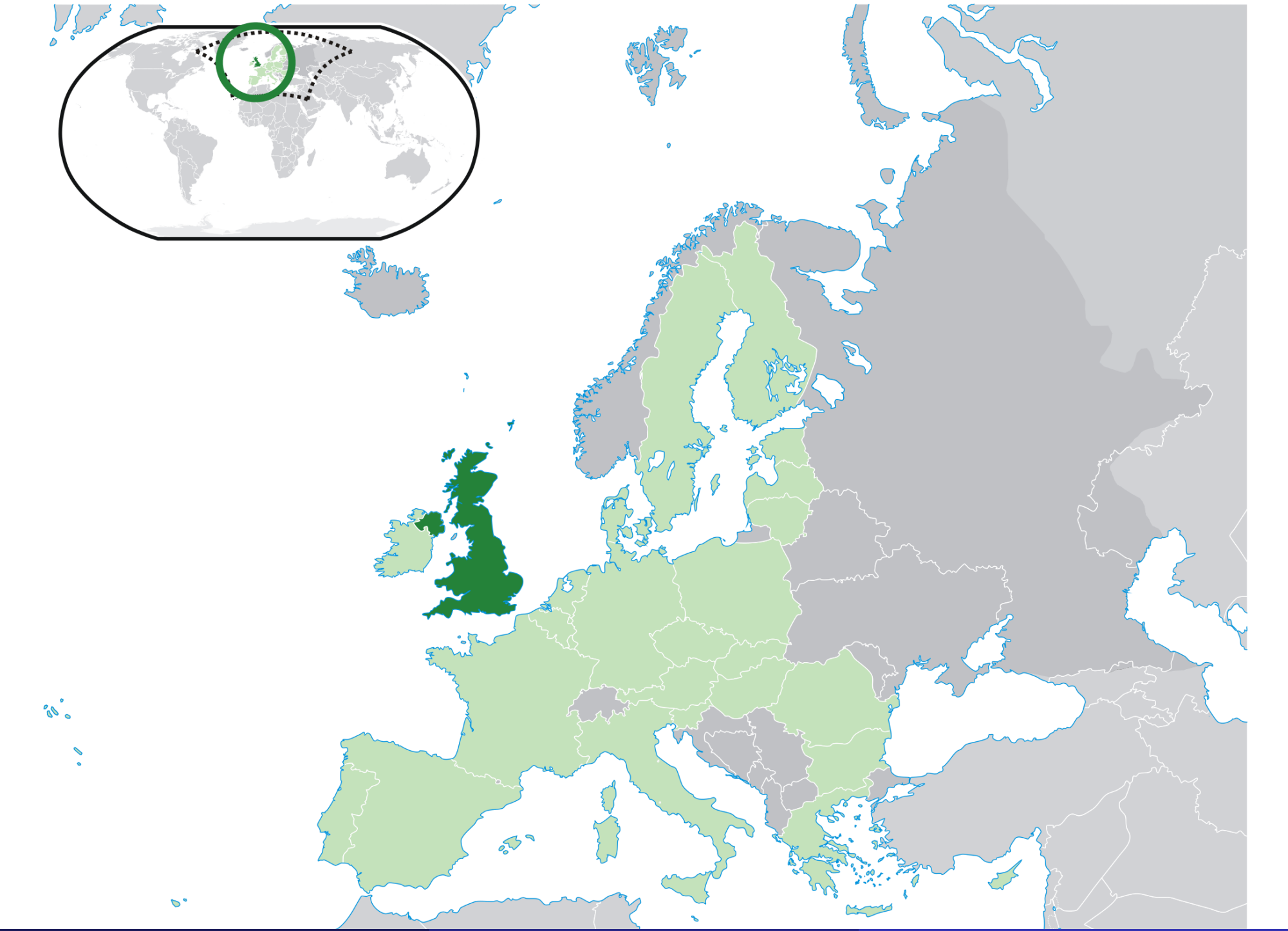
What is Apertium

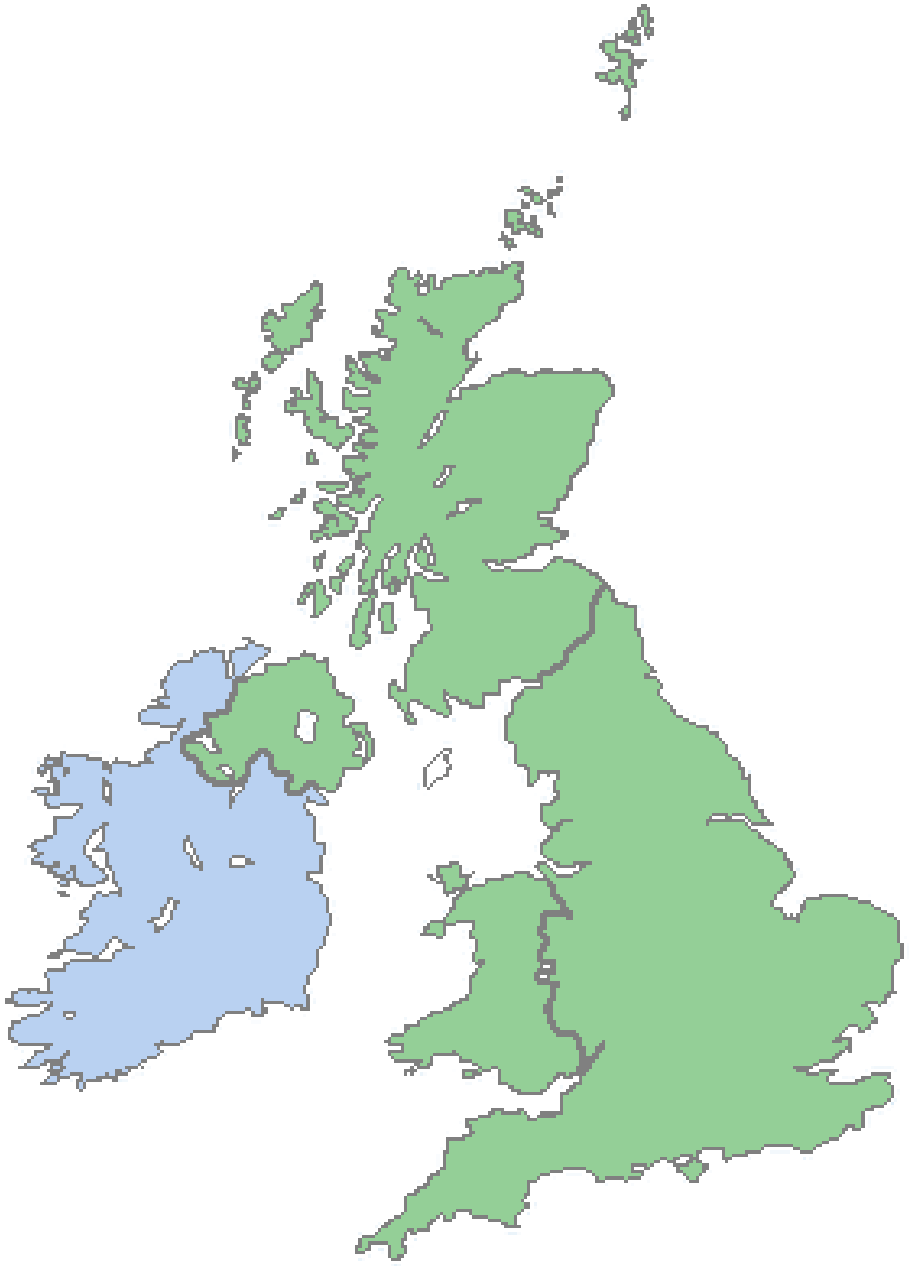
So what is Apertium?

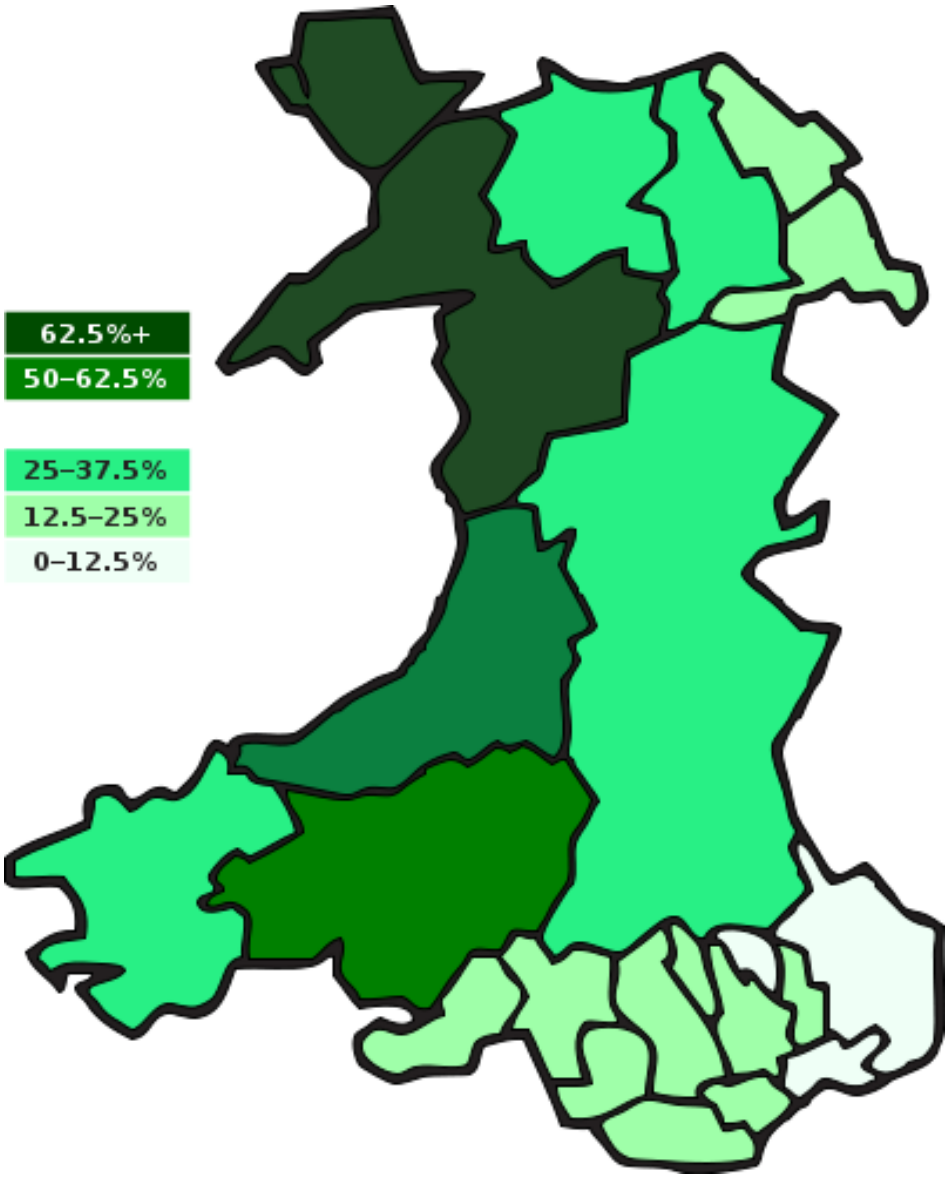
- GPL-licensed platform for machine translation
- Modular – made up of stand-alone programs which communicate using Unix pipes
- Developed by universities, companies and independent developers
- 17 available “stable” language pairs
- More in development
  - And language data in development for many other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

Language pairs







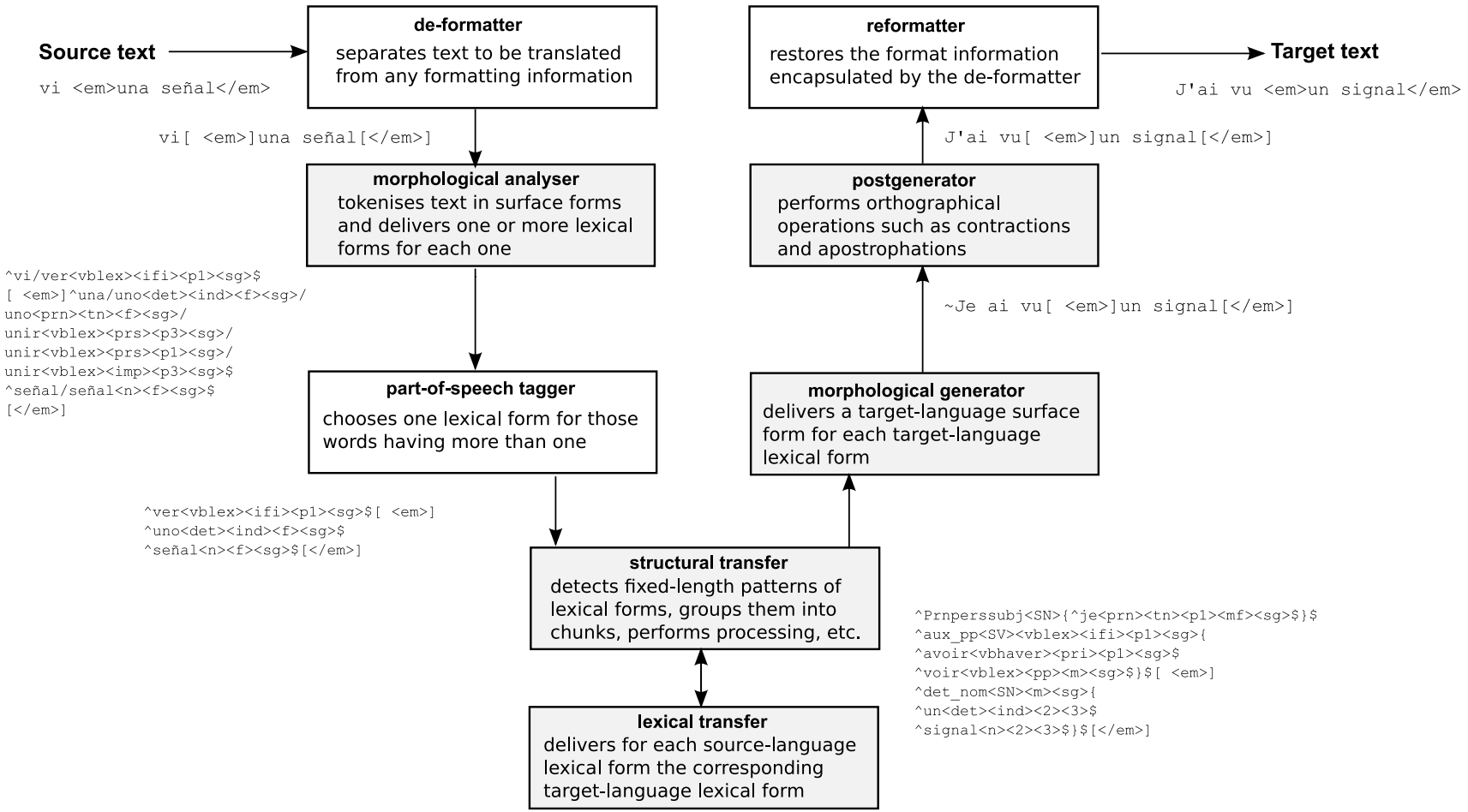

62.5%+
50–62.5%
25–37.5%
12.5–25%
0–12.5%

Development

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 (`nom adj` → `adj nom`)
  - Chunking (`adj adj nom` → `SN[adj adj nom]`)
  - Insertions, deletions and substitutions of lexical units and chunks

Translation model



Existing data (Welsh→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

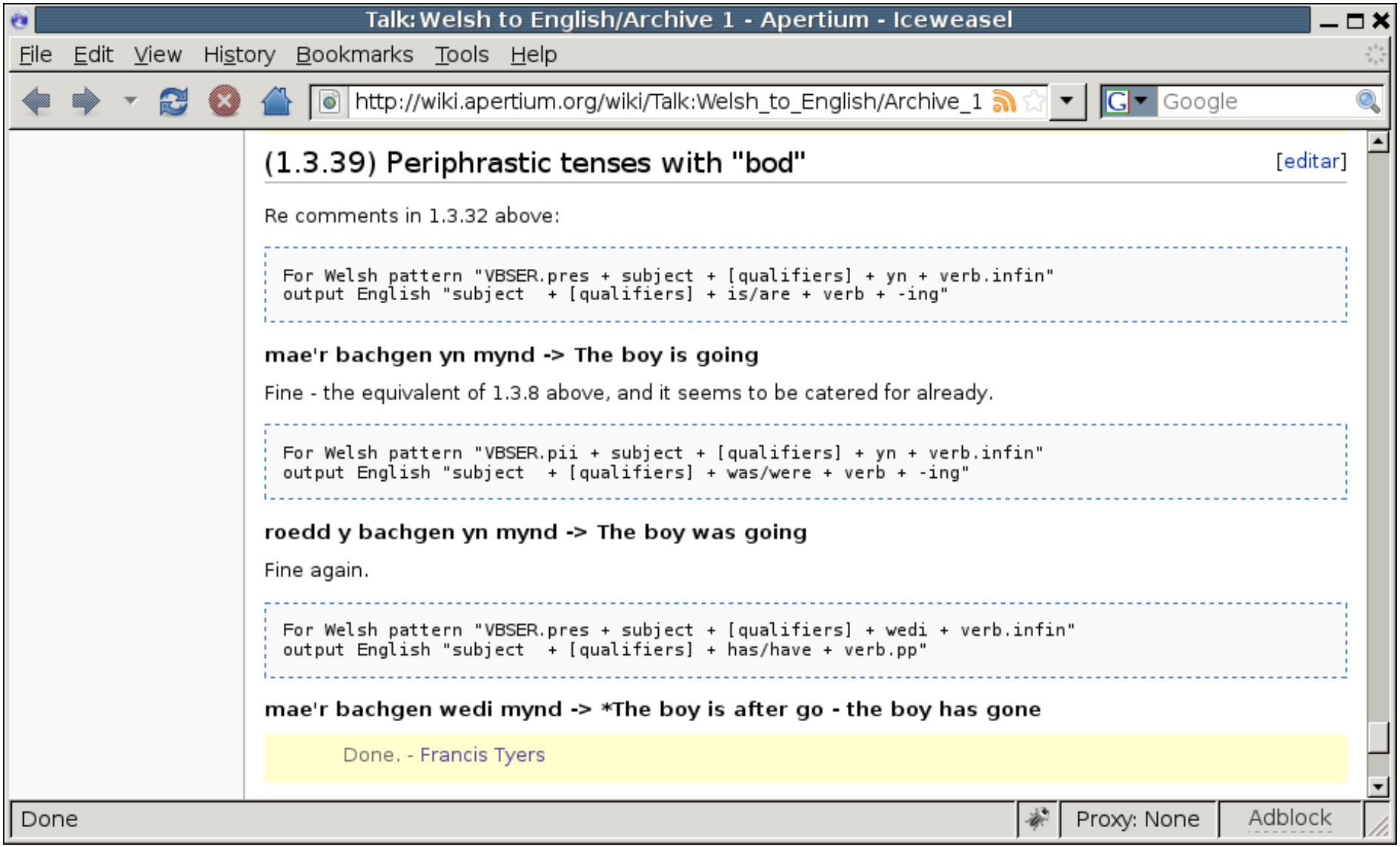
[1] http://www.eurfa.org.uk

[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


Talk:Welsh to English/Archive 1 - Apertium - Iceweasel
File Edit View History Bookmarks Tools Help
http://wiki.apertium.org/wiki/Talk:Welsh_to_English/Archive_1
Google
(1.3.39) Periphrastic tenses with "bod"
[editar]
Re comments in 1.3.32 above:
For Welsh pattern "VBSER.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 "VBSER.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 "VBSER.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
Done
Proxy: None
Adblock

Transfer rules

```
<when>
  <test>
    <and>
      <equal>
        <clip pos="1" part="tipus">
        <lit-tag v="noun">
      </equal>
      <equal>
        <clip pos="2" part="tipus"/>
        <lit v="det.def"/>
      </equal>
      <equal>
        <clip pos="3" part="tipus"/>
        <lit-tag v="noun"/>
      </equal>
    </and>
  </test>
  <out>
     <clip pos="2" part="whole"/>
    <b/>
     <clip pos="1" part="whole"/>
    <b/>
     <lit v="of"/>
    <b/>
     <clip pos="2" part="whole"/>
    <b/>
     <clip pos="3" part="whole"/>
  </out>
</when>
```

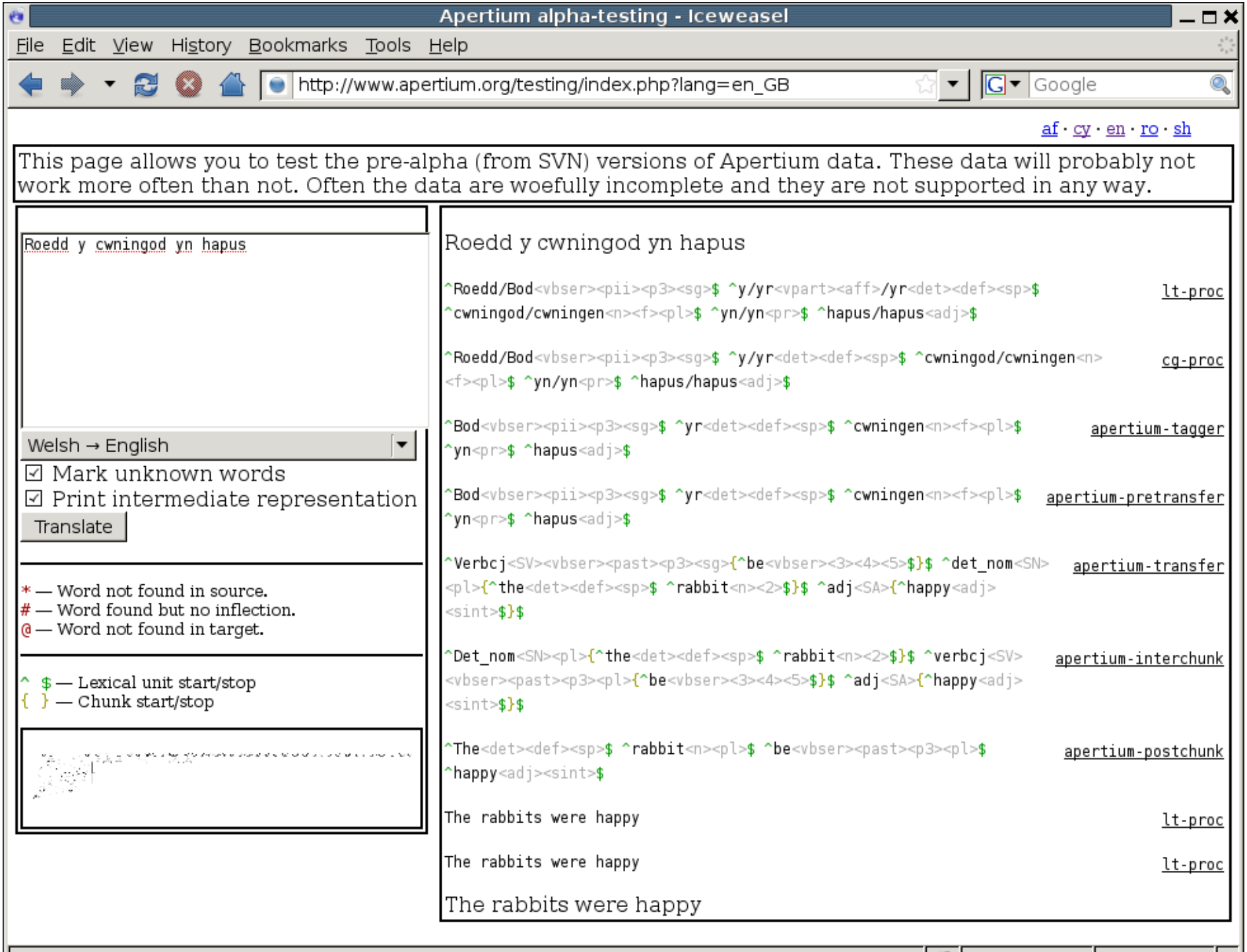
Transfer rules

“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

	Number	Coverage
Lexicon	10,994	90.1%
Disambiguation	56	-
Chunk	84	-
Inter-chunk	36	-

Table: **Statistics from current SVN revision #8140**


Apertium alpha-testing - Iceweasel
File Edit View History Bookmarks Tools Help
http://www.apertium.org/testing/index.php?lang=en_GB
Google
af · cy · en · ro · sh
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
Welsh → English
Mark unknown words
Print intermediate representation
Translate
* — Word not found in source.
# — Word found but no inflection.
@ — Word not found in target.
^ $ — Lexical unit start/stop
{ } — Chunk start/stop
Roedd y cwningod yn hapus
^Roedd/Bod<vbser><pii><p3><sg>$ ^y/yr<vpart><aff>/yr<det><def><sp>$ ^cwningod/cwningen<n><f><pl>$ ^yn/yn<pr>$ ^hapus/hapus<adj>$
lt-proc
^Roedd/Bod<vbser><pii><p3><sg>$ ^y/yr<det><def><sp>$ ^cwningod/cwningen<n><f><pl>$ ^yn/yn<pr>$ ^hapus/hapus<adj>$
cg-proc
^Bod<vbser><pii><p3><sg>$ ^yr<det><def><sp>$ ^cwningen<n><f><pl>$ ^yn<pr>$ ^hapus<adj>$
apertium-tagger
^Bod<vbser><pii><p3><sg>$ ^yr<det><def><sp>$ ^cwningen<n><f><pl>$ ^yn<pr>$ ^hapus<adj>$
apertium-pretransfer
^Verbcj<SV><vbser><past><p3><sg>{^be<vbser><3><4><5>$}$ ^det_nom<SN><pl>{^the<det><def><sp>$ ^rabbit<n><2>$}$ ^adj<SA>{^happy<adj><sint>$}$
apertium-transfer
^Det_nom<SN><pl>{^the<det><def><sp>$ ^rabbit<n><2>$}$ ^verbcj<SV><vbser><past><p3><pl>{^be<vbser><3><4><5>$}$ ^adj<SA>{^happy<adj><sint>$}$
apertium-interchunk
^The<det><def><sp>$ ^rabbit<n><pl>$ ^be<vbser><past><p3><pl>$ ^happy<adj><sint>$
apertium-postchunk
The rabbits were happy
lt-proc
The rabbits were happy
lt-proc
The rabbits were happy

Evaluation

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.

Quantitative (I)

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**[3] – 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

Quantitative (II)

	WER	**PER**	**BLEU**
Wikipedia true-case	55.78	30.59	30.70
Wikipedia lowercased	**53.40**	**27.22**	**32.21**
PNAW true-case	65.99	35.44	15.12
PNAW lowercased	**64.94**	**34.35**	**15.68**

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

Qualitative

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

Comparative

We also did some brief comparative tests with InterTran[4] and default Moses configuration trained on the PNAW corpus.[5]

`or` *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` **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.[6]

[4] `http://www.tranexp.com/`
[5] `http://xixona.dlsi.ua.es/corpora/`
[6] `http://elx.dlsi.ua.es/~fran/welsh/`

Geriaoueg - Iceweasel

File Edit View History Bookmarks Tools Help

http://elx.dlsi.ua.es/geriaoueg/navegador.php

Google

cy-en **Page:** http://news.bbc.co.uk/welsh/hi/newsid_7730000/newsid_7734100/7734129.stm Browse Home

Off On Translate

BBC Home Search Explore the BBC

Fersiwn graffeg isel | Adborth | Cymorth

BBC Cymru Newyddion

Chwiliwch yn Gymraeg EWCH

Newyddion
Gwleidyddiaeth
Chwaraeon
Y Tywydd
Lleol
Pobl ifanc
Dysgwyr

BBC Cymru'r Byd
Chwaraeon
Y Tywydd

Dysgu Cymraeg
Learn Welsh

Lleol i Mi

BBC VOCAB
OFF / I FFWRDD
»Tum ON
Troi YMLAEN
»What is VOCAB?
Beth yw GEIRFA?

Diweddarwyd: Dydd Mawrth, 18 Tachwedd 2008, 12:39 GMT

Fersiwn brintiadwy

Cynllun i greu cymuned Aeleg

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

Bwriad y cynllun yw cryfhau'r defnydd o'r iaith Aeleg

Dywedodd ymgyrchwyr y byddai cymuned o fwy na 1000 o siaradwyr Gaeleg yn "hwb enfawr" i'r iaith.

Mae nifer y siaradwyr wedi lleihau'n fawr yn y blynyddoedd diwethaf, cyfanswm o 58,652 yn ôl Cyfrifiad 2001, a'r mwyafrif yn 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.

HEFYD
- Sianel Aeleg newydd yn dechrau
  19 Medi 08 | Newyddion
- Gaeleg: Newyddion ar wefan
  28 Chwef 08 | Newyddion

CYSYLLTIADAU RHYNGRWYD:
- Alba (Gaeleg / Saesneg)
- Llywodraeth yr Alban (Saesneg)
- Ymddiriedolaeth y BBC yn cymeradwyo'r Gwasanaeth Digidol Gaeleg yn amodol

Dyw'r BBC ddim yn gyfrifol am gynnwys safleoedd rhyngrwyd allanol

PRIF STRAEON NEWYDDION
- Dwr: Cyflenwadau yn ôl
- 'Gweledigaeth' darpar arweinydd?
- Cnydau'n 'hwb i gefn gwlad'

RSS | Beth yw RSS?

Done Proxy: None Adblock

Geriaoueg - Iceweasel

File Edit View History Bookmarks Tools Help

http://elx.dlsi.ua.es/geriaoueg/navegador.php

Google

cy-en **Page**: http://news.bbc.co.uk/welsh/hi/newsid_7730000/newsid_7734100/7734129.stm Browse Home

Off On Translate

BBC *Home Search *Explore BBC tea

Feedback low graphics Version | Help

BBC Cymru Newyddion

Search Welsh EWCH

News
Politics
Games
THE Weather
Local
young Peoples
Learners

the BBC Cymru of the World
Games
THE Weather

Teach Welsh
*Learn *Welsh

Local to me

BBC VOCAB
OFF / To WAY
»*Turn ON Tum FORWARD
»*What a lower VOCAB? What

were Updated : Tuesday, 18 Novembers 2008, 12:39 GMT

Version *brintiadwy

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 Intention of the plan is strengthen the material of the language Scots Gaelic

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'

ALSO
- new Scots Gaelic Channel beginning
  19 Septembers 08 News
- a Scots Gaelic: News on a website
  28 *Chwef 08 News

INTERNET ASSOCIATIONS:
- Alba (English Scots Gaelic)
- Scotland Government (English)
- the Trust of the BBC approving '*r Scots Gaelic Digital Service conditional

the BBC Are not responsible for positions content an external internet

PRINCIPAL STORIES NEWS
- Heap: Supplies according to
- 'leader' designate Vision?
- Crops in 'push to countryside'

RSS | Beth yw RSS?

Done Proxy: None Adblock

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.

Geriaoueg - Iceweasel

File Edit View History Bookmarks Tools Help

http://elx.dlsi.ua.es/geriaoueg/navegador.php

Google

cy-en **Page:** http://news.bbc.co.uk/welsh/hi/newsid_7730000/newsid_7734100/7734129.stm Browse Home

Off On Translate

BBC Home Search Explore BBC

Fersiwn graffeg isel Adborth Cymorth

BBC Cymru Newyddion Chwiliwch yn Gymraeg EWCH

Newyddion
Gwleidyddiaeth
Chwaraeon
Y Tywydd
Lleol
Pobl ifanc
Dysgwyr

BBC Cymru 'r Byd
Chwaraeon
Y Tywydd

Dysgu Cymraeg
Learn Welsh

Lleol i Mi

BBC VOCAB
OFF I FFWRDD
Tum ON
Troi YMLAEN
What is VOCAB
?Beth yw GEIRFA ?

Diweddarwyd : Dydd Mawrth ,18 Tachwedd 2008 ,12 :39 GMT

Fersiwn brintiadwy

Cynllun i greu cymuned Aeleg

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

Bwriad y cynllun yw cryfhau 'r defnydd o'r iaith Aeleg

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

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

(Ucheldir) highland (n)

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 .

HEFYD
- Sianel Aeleg newydd yn dechrau 19 Medi 08 Newyddion
- Gaeleg :Newyddion ar wefan 28 Chwef 08 Newyddion

CYSYLLTIADAU RHYNGRWYD :
- Alba (Gaeleg Saesneg)
- Llywodraeth yr Alban (Saesneg)
- Ymddiriedolaeth y BBC yn cymeradwyo r Gwasanaeth Digidol Gaeleg yn amodol

Dyw 'r BBC ddim yn gyfrifol am gynnwys safleoedd rhyngrwyd allanol

PRIF STRAEON NEWYDDION
- Dwr :Cyflenwadau yn ôl
- Gweledigaeth darpar arweinydd ?
- Cnydau 'n hwb i gefn gwlad

RSS | Beth yw RSS?

Done Proxy: None Adblock

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!