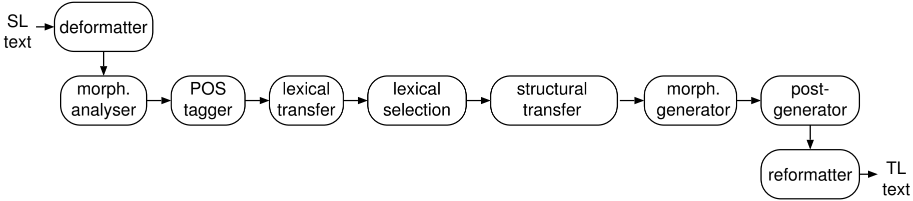
Apertium Tutorial

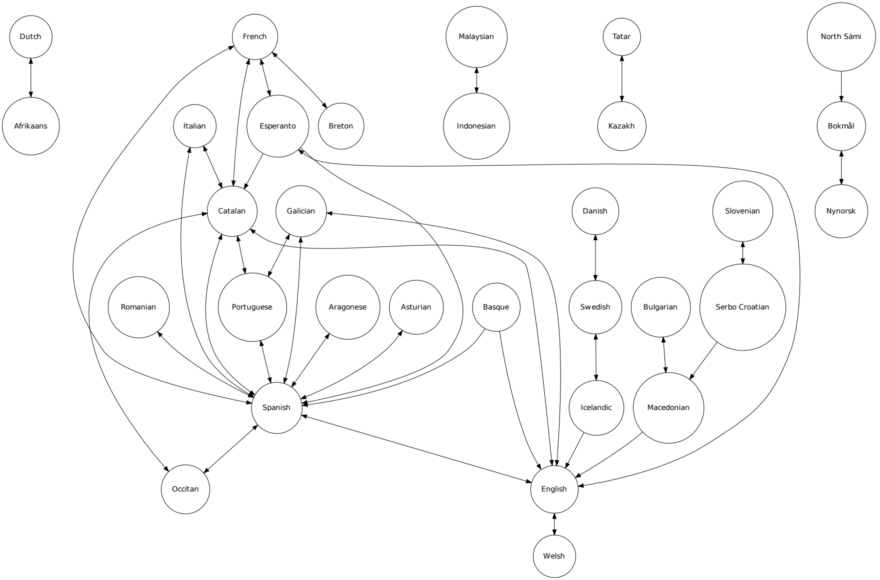
Francis Morton Tyers

17th July 2013

Introduction



- ‘Platform’ for making rule-based machine translation systems
- Free/open-source (GPL): Pretty militant about this
- 37 released language pairs
    - Perhaps around 200 in total
- Many developers (mostly on the linguistic side)


Dutch
French
Malaysian
Tatar
North Sámi
Afrikaans
Italian
Esperanto
Breton
Indonesian
Kazakh
Bokmål
Catalan
Galician
Danish
Slovenian
Nynorsk
Romanian
Portuguese
Aragonese
Asturian
Basque
Swedish
Bulgarian
Serbo Croatian
Spanish
Icelandic
Macedonian
Occitan
English
Welsh

State of the art?

For the vast majority of languages that Apertium supports, it is not state-of-the-art. However,

Monolingual resources:

- Occitan, Afrikaans, Maltese, Welsh, Breton, Kazakh, Tatar, Macedonian, Norwegian, Aragonese, Asturian, …

MT systems:

- Icelandic–Swedish, Norwegian Nynorsk–Bokmål, Serbo-Croatian–Slovenian, Spanish–Catalan, Occitan–Spanish, Occitan–Catalan, Breton–French, Kazakh–Tatar, Afrikaans–Dutch, …

And for some other languages (e.g. Albanian, Armenian, Kyrgyz) it may be ‘state-of-the-art’ solely for lack of competing resources.
Do we really need state-of-the-art tools to improve MT ?

Some languages

Afrikaans · Albanian · Arabic · Aragonese · Armenian · Asturian · Basque · Bengali · Breton · Bulgarian · Catalan · Chuvash · Czech · Danish · Dutch · English · Esperanto · Faroese · French · Galician · Hebrew · Hindi · Icelandic · Indonesian · Iranian Persian · Irish · Italian · Kazakh · Kyrgyz · Latvian · Macedonian · Malaysian · Maltese · Norwegian Bokmål · Norwegian Nynorsk · Occitan · Polish · Portuguese · Quechua · Romanian · Russian · Sardinian · Scottish Gaelic · Serbo-Croatian · Slovenian · Spanish · Swedish · Tajik · Tatar · Tetum · Ukrainian · Urdu · Welsh

```
Následný|následný|AAIS1----1A---- postup|postup|NNIS1-----A-
na|na-1|RR--6---------- základě|základ|NNIS6-----A----
usnesení|usnesení_^(*5ést)|NNNS2-----A----
Parlamentu|parlament|NNIS2-----A---- :|:|Z:-------------
viz|viz_:W_^(odkaz_na_jiné_místo)|Vi-S---2--A---1
zápis|zápis|NNIS4-----A----
```

```
Action|action|NN taken|take|VBN on|on|IN
Parliament|Parliament|NNP '|'s|POS s|'s|POS
resolutions|resolution|NNS :|:|: see|see|VB
Minutes|minute|NNS
```

```
Человек
  ЧЕЛОВЕК   С мр,рд,им,ед,мн,
,
  ,         С ср,жр,мр,пр,тв,вн,дт,рд,им,ед,мн,
его
 ЕГО     МС-П 0,но,од,ср,жр,мр,зв,пр,тв,вн,дт,рд,им,ед,мн,
  ОН        МС 3л,мр,вн,рд,ед,
  ОНО       МС 3л,ср,вн,рд,ед,
^Následný/následný<adj><sint><mi><sg><nom>$
^postup/postup<n><mi><sg><nom>$ ^na/na<pr>$
^základě/základ<n><mi><sg><loc>$
^usnesení/usnesení<n><nt><sg><gen>$
^Parlamentu/parlament<n><mi><sg><gen>$^:/:<sent>$
^viz/viz<vblex><imp><p2><sg>$ ^zápis/zápis<n><mi><sg><acc>$

^Action/action<n><sg>$
^taken/take<vblex><pp>$ ^on/on<pr>$
^Parliament/parliament<n><sg>$ ^’s/’s<gen>$
^resolution/resolution<n><sg>$^:/:<sent>$
^see/see<vblex><imp>$ ^minutes/minute<n><pl>$

^Человек/человек<n><m><aa><sg><nom>$^,/,<cm>$
^его/его<det><pos>$ ^права/право<n><nt><nn><pl><nom>$
^и/и<cnjcoo>$ ^свободы/свобода<n><f><nn><pl><nom>$
^являются/являться<vblex><impf><pres><p3><pl>$
^высшей/*высшей$ ^ценностью/ценность<n><f><nn><sg><ins>$
```

Repository layout

`https://svn.code.sf.net/p/apertium/svn`

- `incubator`: Scratchpad, snippets
- `nursery`: Pairs which have had quite a bit of development effort
- `staging`: Pairs which are almost release ready
- `trunk`: Released language pairs,

- `languages`: New top-level module, for monolingual language resources, morphological analysis and disambiguation.

Browse:
`https://sourceforge.net/p/apertium/svn/HEAD/tree/`

Documentation and contacting us

Documentation:

- A load of documentation on our wiki:
- `http://wiki.apertium.org/wiki/Main_Page`

Courses:

- `http://wiki.apertium.eu/index.php/Main_Page`
- `http://wiki.apertium.org/wiki/Курсы_машинного_перевода_для_языков_России`

Contact:

- IRC: #apertium, irc.freenode.net
- Mailing lists: apertium-stuff@lists.sourceforge.net, apertium-turkic@lists.sourceforge.net, apertium-celtic@lists.sourceforge.net

Install

```
$ export APERTIUMSVN=\
 https://svn.code.sf.net/p/apertium/svn/
$ svn co $APERTIUMSVN/trunk/lttoolbox
$ svn co $APERTIUMSVN/trunk/apertium
$ svn co $APERTIUMSVN/trunk/apertium-lex-tools
$ svn co $APERTIUMSVN/trunk/apertium-tools/apertium-moses
```

```
$ export PREFIX=$HOME/apertium
$ export PATH=$PREFIX/bin:$PATH
$ export PKG_CONFIG_PATH=$PREFIX/lib/pkgconfig
$ export LD_LIBRARY_PATH=$PREFIX/lib
```

```
$ ./autogen.sh --prefix=$PREFIX
$ make ; make install
```

You may also need VISL constraint grammar:

http://wiki.apertium.org/wiki/Apertium_and_Constraint_Grammar

Testing

Let’s try a language pair:
https://svn.code.sf.net/p/apertium/svn/trunk/apertium-mk-en

```
$ svn co $APERTIUMSVN/trunk/apertium-mk-en

$ cd apertium-mk-en
$ ./autogen.sh --prefix=$PREFIX
$ ./configure --prefix=$PREFIX
$ make

$ echo ”Водачите на Косово и Србија на 9-ти септември направија
  значајни подготовки за локалните избори во Косово, што
  претставува нивни последен состанок пред изборите
  на 3-ти ноември.” | apertium -d . mk-en
```

Source text: http://tinyurl.com/setimesart1

Analysing and tagging

```
$ cat f | apertium-destxt | lt-proc -w mk-en.automorf.bin
```

```
$ cat f | apertium-destxt | lt-proc -w mk-en.automorf.bin |\
  cg-proc mk-en.rlx.bin
```

```
$ cat f | apertium-destxt | lt-proc -w mk-en.automorf.bin |\
 cg-proc mk-en.rlx.bin  | apertium-tagger -p -g mk-en.prob
```

```
$ cat f | apertium-destxt | lt-proc -w mk-en.automorf.bin |\
 cg-proc mk-en.rlx.bin  | apertium-tagger -p -g mk-en.prob  |\
  python3 apertium-moses/tagger-to-factored.py
```