

Apertium: A rule-based machine translation platform

Francis M. Tyers

HSL-fakulteha
UiT Norgga árkalaš universitehta
9018 Romsa (Norway)

28th September 2014



Outline

Introduction

- Design
- Development
- Status

Teaching

- Courses
- Google Summer of Code
- Google Code-in

Research

- New language pairs
- Applying unsupervised methods
- Hybrid systems

Future work and challenges

- Challenges
- Plans
- Collaboration



History

- ▶ 2004 — Spain gets a new government which launches a call for proposals to build machine translation systems for the languages of Spain
- ▶ The Universitat d'Alacant (UA), in consortium with EHU, UPC, UVigo, Eleka, Elhuyar (Basque Country) and Imaxin (Galicia) get funded to develop two MT systems:
 - ▶ **Apertium**: Spanish–Catalan, Spanish–Galician
 - ▶ Matxin: Spanish→Basque
- ▶ Apertium was not built from scratch, but was rather a rewrite of two existing closed-source systems which had been built by UA



Focus

- ▶ **Marginalised:** Languages which are on the periphery either societally or technologically (from Breton to Bengali).
- ▶ **Lesser-resourced:** Languages for which few *free/open-source* language resources exist.
- ▶ **Closely-related:** Languages which are suited to shallow-transfer machine translation.



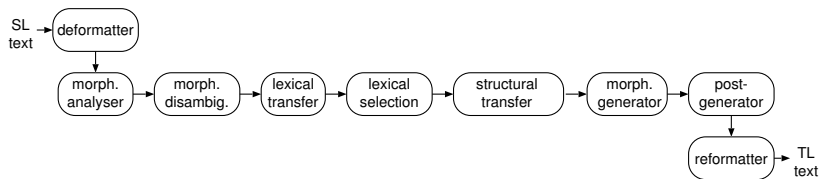
Translation philosophy

- ▶ Build on word-for-word translation
- ▶ Avoid complex linguistic formalisms
 - ▶ We like saying “only secondary-school linguistics required”¹
 - ▶ Intermediate representation based on morphological information
- ▶ Transformer-style systems
 - ▶ Analyse the source language (SL), then
 - ▶ Apply rules to ‘transform’ the SL to the TL

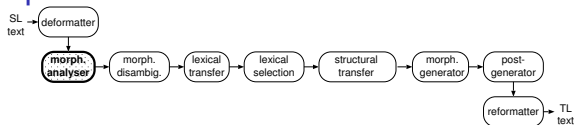
¹Otte, P. and Tyers, F. (2011) “Rapid rule-based machine translation between Dutch and Afrikaans”. Proceedings of the 16th Annual Conference of the European Association of Machine Translation



Pipeline



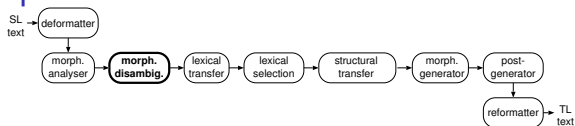
Step-by-step



Morphological analysis

```
$ echo "Tiilitalojen keskellä hän kirjoitti minulle nimensä kiinalaisin kirjaimin." | apertium -d  
. fin-eng-morph  
^Tiilitalojen/tiilitalo<n><pl><gen>$ ^keskellä/keskellä<adv>/keskellä<post>$  
^hän/hän<prn><pers><sg><nom>$ ^kirjoitti/kirjoittaa<vblex><actv><past><p3><sg>$  
^minulle/minä<prn><pers><sg><all>$  
^nimensä/nimi<n><pl><nom><pxsp3>/nimi<n><sg><gen><pxsp3>/nimi<n><sg><nom><pxsp3>$  
^kiinalaisin/kiinalainen<adj><pos><pl><ins>/kiinalainen<adj><v+n><sup><sg><nom>/kiinalainen<n><pl><ins>$  
^kirjaimin/kirjain<n><pl><ins>$^./.<sent>$
```

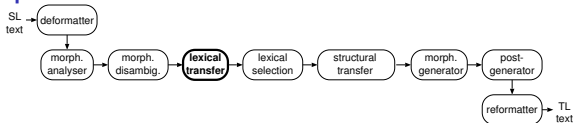
Step-by-step



Morphological disambiguation

```
$ echo "Tiilitalojen keskellä hän kirjoitti minulle nimensä kiinalaisin kirjaimin." | apertium -d  
. fin-eng-tagger  
^Tiilitalo<n><pl><gen>$ ^keskellä<post>$ ^hän<prn><pers><sg><nom>$  
^kirjoittaa<vblex><actv><past><p3><sg>$ ^minä<prn><pers><sg><all>$ ^nimi<n><sg><gen><pxsp3>$  
^kiinalainen<adj><pos><pl><ins>$ ^kirjain<n><pl><ins>$^.<sent>$
```

Step-by-step

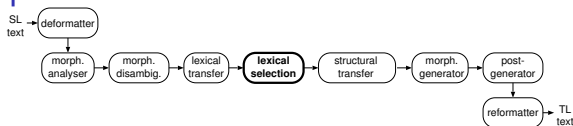


Lexical transfer

```
$ echo "Tiilitalojen keskellä hän kirjoitti minulle nimensä kiinalaisin kirjaimin." | apertium -d  
. fin-eng-biltrans  
^Tiilitalo<n><pl><gen>/Brick house<n><pl><gen>$ ^keskellä<post>/in between<pr>$  
^hän<prn><pers><sg><nom>/she<prn><pers><f><sg><nom>/they<prn><pers><mf><pl><nom>/  
  he<prn><pers><m><sg><nom>$ ^kirjoittaa<vblex><actv><past><p3><sg>/author<vblex><actv><past><p3><sg>/  
  cut<vblex><actv><past><p3><sg>/write<vblex><actv><past><p3><sg>/type<vblex><actv><past><p3><sg>/  
  draft<vblex><actv><past><p3><sg>/spell<vblex><actv><past><p3><sg>/...$  
^minä<prn><pers><sg><all>/I<prn><pers><mf><sg><all>$  
^nimi<n><sg><gen><pxsp3>/name<n><sg><gen><pxsp3>/title<n><sg><gen><pxsp3>/header<n><sg><gen><pxsp3>$  
^kiinalainen<adj><pos><pl><ins>/Chinese<adj><pos><pl><ins>$  
^kirjain<n><pl><ins>/letter<n><pl><ins>/character<n><pl><ins>$^<sent>/<sent>$
```



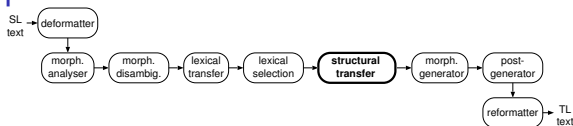
Step-by-step



Lexical selection

```
$ echo "Tiilitalojen keskellä hän kirjoitti minulle nimensä kiinalaisin kirjaimin." | apertium -d  
. fin-eng-lextor  
^Tiilitalo<n><pl><gen>/Brick house<n><pl><gen>$ ^keskellä<post>/in between<pr>$  
^hän<prn><pers><sg><nom>/they<prn><pers><p3><mf><sg><nom>$  
^kirjoittaa<vble><actv><past><p3><sg>/write<vble><actv><past><p3><sg>$  
^minä<prn><pers><sg><all>/I<prn><pers><mf><sg><all>$  
^nimi<n><sg><gen><pxsp3>/name<n><sg><gen><pxsp3>$  
^kiinalainen<adj><pos><pl><ins>/Chinese<adj><pos><pl><ins>$  
^kirjain<n><pl><ins>/character<n><pl><ins>$^.<sent>/.<sent>$
```

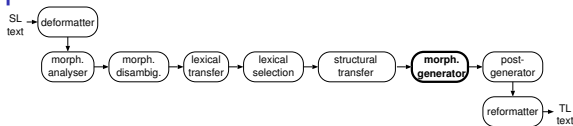
Step-by-step



Structural transfer

```
$ echo "Tiilitalojen keskellä hän kirjoitti minulle nimensä kiinalaisin kirjaimin." | apertium -d  
. fin-eng-postchunk  
^In between<pr>$ ^brick house<n><pl>$ ^they<prn><pers><p3><mf><sg>$ ^write<vble><past>$  
^to<pr>$ ^I<prn><pers><p1><mf><sg><acc>$ ^name<n><sg>$ ^with<pr>$ ^Chinese<adj>$  
^character<n><pl>$^.<sent>$
```

Step-by-step



Morphological generation

```
$ echo "Tiilitalojen keskellä hän kirjoitti minulle nimensä kiinalaisin kirjaimin." | apertium -d . fin-eng
```

In between brick houses they wrote to me name with Chinese characters.

Modularity

The pipeline architecture makes it straightforward to insert or substitute modules:

- ▶ **HFST:**

- ▶ Replacement for `lttoolbox`
- ▶ Used for most “morphologically complex” languages

- ▶ **VISL CG:**

- ▶ Accompanies `apertium-tagger`
- ▶ Rule-based morphological disambiguation and shallow-syntactic analysis



What kind of systems are developed?

Dissemination	Assimilation
High precision	Low precision
Min→Min	Gisting
Maj→Min	Min→Maj
Closely-related languages	Unrelated or distant languages
Real <i>translation</i> aids	Reading aids



WIKIPEDIA
The Free Encyclopedia

Publicar traducción

Trädgårdsföreningen

español [ver página](#)

La Sociedad de Horticultura de Gotemburgo, en sueco **Trädgårdsföreningen** es un **jardín botánico** situado en una zona céntrica de la ciudad **sueca** de **Gotemburgo**, no lejos de la Estación Central y del parque *Brunnsparken*.

El jardín propiamente dicho posee una extensión de pocas hectáreas, y está contiguo y adosado al *Jardín Botánico de Gotemburgo* y sin solución de continuidad a la gran zona verde del parque *Slottsskogen*, que entonces sumaría una extensión de 175 **ha** (unos 430 acres), estando incluido el *Arboretum* y la reserva natural.



Puerta de entrada al Trädgårdsföreningen de Gotemburgo.

Trädgårdsföreningen

català

La Societat d'Horticultura de Gotemburgo, en suec **Trädgårdsföreningen** és un jardí botànic situat en una zona cèntrica de la ciutat sueca de Gotemburg, no lluny de l'Estació Central i del parc *Brunnsparken*.

El jardí pròpiament dit posseeix una extensió de poques hectàrees, i està contigu i adossat al *Jardí Botànic de Gotemburgo* i sense solució de continuïtat a la gran zona verda del parc *Slottsskogen*, que llavors sumaria una extensió de 175 ha, estant inclòs el Arboretum i la reserva natural.

+ Añadir traducción

Buscar una palabra

Traducción automática
De Apertium, del \$2 al \$3 -

Restaurar

☐ Usar texto de origen

☒ Borrar traducción



Máilmmi vuosttas eamiálbmogiid Pride-festivála ordnejuvvo Gironis | Yle Uutiset | yle.fi - iceweasel

Eikey Editat Yksilöitä Historial Adreces d'interès Eiges Ajuda

Máilmmi vuosttas... Verdensens de fa...

yle.fi/uutiset/mailmmi_vuosttas_eamiálbmogiid_pride-festivála_ordnejuvvo_gironis/7450625

Google

Sápmi 8.9.2014 klo 14:29 | beaivduvvon 8.9.2014 klo 14:45

Máilmmi vuosttas eamiálbmogiid Pride-festivála ordnejuvvo Gironis

Sámi Nuorra ja Queering Sápmi -prošeakta ordnejit máilmmi vuosttas eamiálbmogiid Pride-festivála golggoimánus Gironis. Festivála sisttisdoallá earret eará ovtastallama, bargobájiid, logaldallamiid ja kulturdoaluid.



Sápmi Pride -leavga lea sáme- ja arvedávgeleavvga seaguhus. Govva: Sápmi Pride

Máilmmi vuosttas eamiálbmogiid Pride-festivála ordnejuvvo Gironis golggoimánus 16.-19. beivviid. Festivála ordne Ruota sámenuoraid organisašuvdna Sámi Nuorra ja ruhtada Queering Sápmi -prošeakta.

Seksuála- ja sohka-beallevehádagaid Pride-festivála sisttisdoallá Gironis earret eará

Váldoodđasat

Sápmi klo 8:50 0

Sovkina ii bala – giitá sámiid veahkis



Sii geat ná láhttejit min vuostá ja hehtejit min searvamis dákkár čoahkkimii, sii dat ballet, Guoládaga sámi parlameantta ságadoalli Valentina Sovkina dajai marjebárga New Yorkkas.

Sápmi klo 14:56 0

Áile Javo ON:s: Árktaš guovlu vuordá, ahte loahppadokumeantta boadusin ruvkeláгат rievdađuvvojit

Sámiráđi ságadoalli Áile Javo doalai árktaš guovlu virggálaš sáhkavuoru máilmmikonferanssa eananvuolgtvuochtaságastallamis New Yorkkas marjebárgga.

Sápmi klo 12:04 0

Dutkamuš: Eanandoallit nagodedje ealihit eambo mánáid go sápmelaččat

Šibit- ja eanandoallu ealihedje 1700-1900-jođuin buorebut go árbevrolaš bivdu ja boazodoallu. Dan dihte suomelaččat lassáneđje jođáneabbot go sápmelaččat, čájeha sámevuovlu girkojirjiid vuodul

Verdenens de første urfolkenes Pride-festivalen ordnes i Kiruna | Yle Uutiset | yle.fi - Icestream

Eibar Editar Visualiza Historial Añades d'interès Eines Ajuda

Máilmmi vuostt... x Verdenens de f... x


file:///tmp/test.nob.html

Google

Samen 8.9.2014 klo 14:29 | beaivdávven 8.9.2014 klo de 14:45

Verdenens første urfolks Pride-festivalen ordnes i Kiruna

Samens Ung og Queering Sameland -prosjektet ordner verdenens de første urfolkenes Pride-festivalen i oktober i Kiruna. Festivalen inneholder blant annet samvær, bargoåbjid, forelesningene og kulturarrangementet.




Sameland Pride -flagget er samens- og arvedávgeleavvga en blanding. Et bilde: En same Pride

Verdenens de første urfolkenes Pride-festivalen ordnes i Kiruna oktobers de 16.-19. dagene. Festivalen ordner Sveriges sameungdommer organisasjonen Samelands Ungdom og finansierer Queering Sameland -prosjektet.

Váldoodđasat

Samen klo den 8:50 0.

Sovkina frykter ikke – han takker på samenes hjelp



De som så oppfører de seg mot oss og de hindrer oss å slutte seg sammen til slikt møte, de frykter det, Kolas sames parlamentets talsmann Valentina Sovkina sa tirsdagen New Yorkkas.

Sameland klo 14:56 0.

Åile Javo i ON: Den arktiske retningen venter, at sluttokumentet som resultat forandres ruvkeláat

Samerådets talsmann Åile Javo holdt det arktiske retnings offisielle innleggs verdenskonferansen eanamuigatvuochtaságastallamis New Yorkkas tirsdagens.

Sameland klo 12:04 0.

Dutkamuš: Gårdbrukerne klarte livberge mye barn enn samer

Husdyr- og jordbruket livberget 1700-1900- med et tall bedre enn den tradisjonelle fisket og reindriften. for Det økte flirderne Jodnesabot når samene, hun viser sameområdet på grunn av kirkebøkene gjort dutkamuš.



Effort

How much effort does it take to develop a system with Apertium ?

- ▶ 2 weeks minimum
 - ▶ Fastest system made (Macedonian→English)
- ▶ After 3 months, success rate is around 50%
 - ▶ Around half of the students that participate in the Google Summer of Code are able to finish their systems in 12 weeks
- ▶ After 6 months, perhaps 75% success rate
 - ▶ Students who don't quite make it in the Google Summer of Code can have their projects picked up and finished in about three months more.
- ▶ After 1 year, ...
 - ▶ When asking for funding, this is probably the minimum amount of time you'd put.



Success factors

Technical:

- ▶ How much data is freely available
 - ▶ In terms of dictionaries, rule descriptions and corpora
- ▶ Stability and compatibility of tagsets (intermediate representation)

Linguistic:

- ▶ Morphological complexity of the languages
- ▶ Genetic and structural similarity of the languages

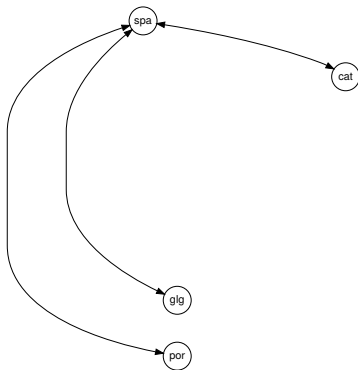
Personal:

- ▶ Experience of the developer with the tools
- ▶ Staying power of the developer
 - ▶ Do they have the staying power to check 100s of lexicon entries/day for weeks on end ?



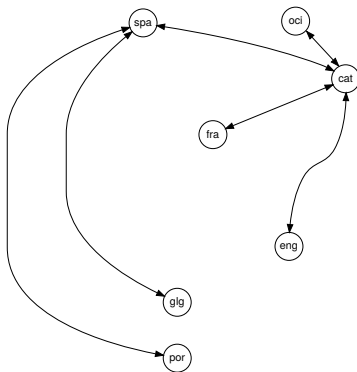
Language pairs

2005



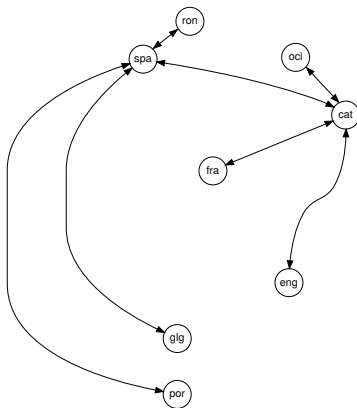
Language pairs

2006



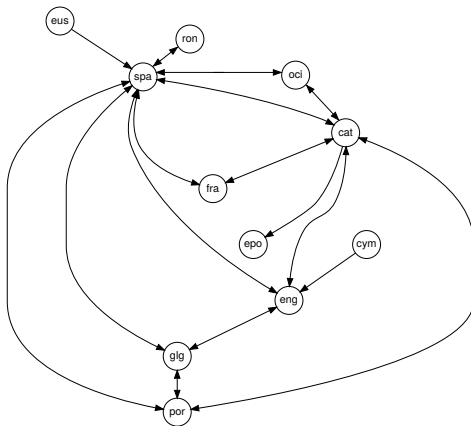
Language pairs

2007



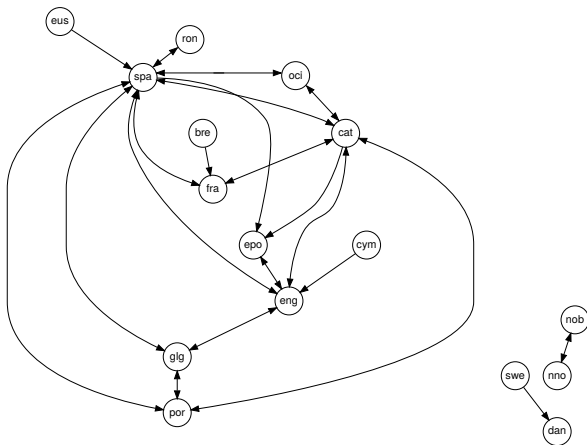
Language pairs

2008



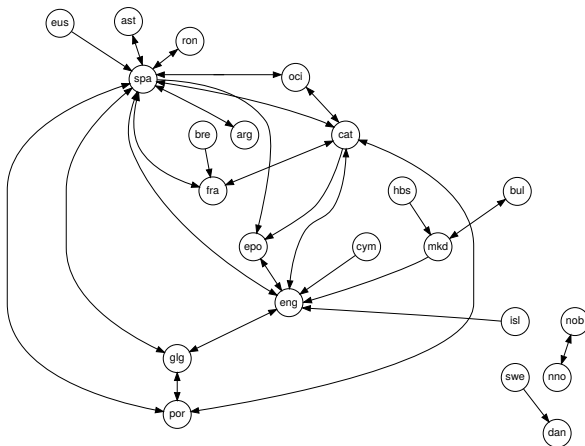
Language pairs

2009



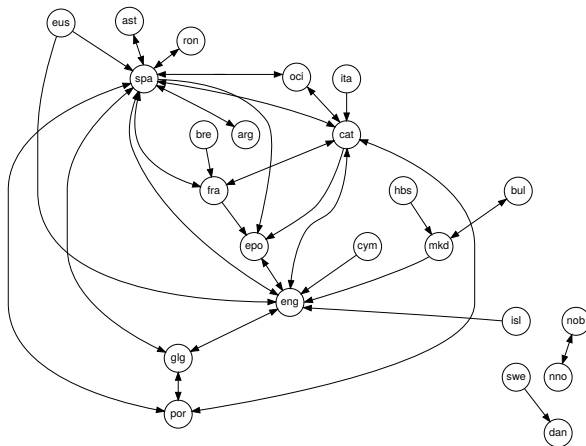
Language pairs

2010



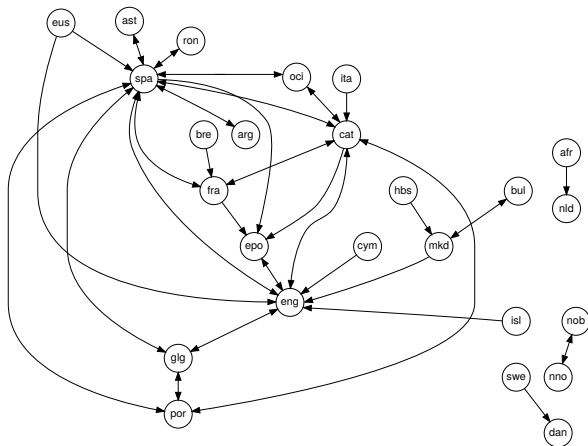
Language pairs

2011



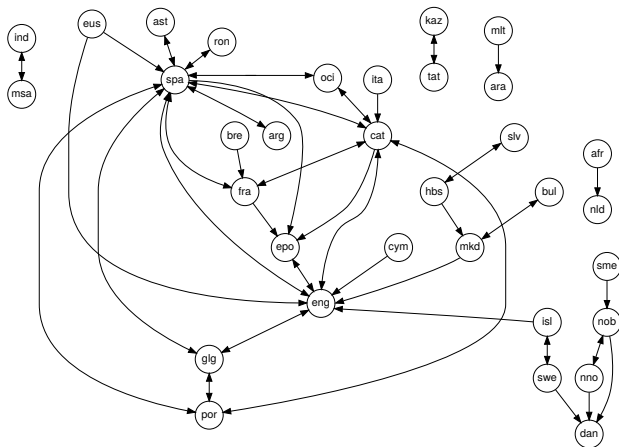
Language pairs

2012



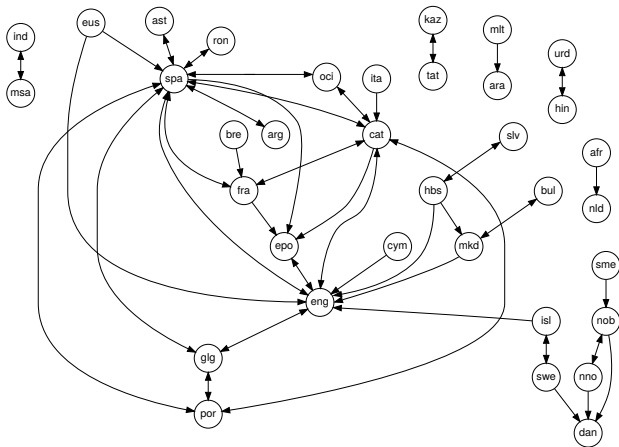
Language pairs

2013



Language pairs

2014

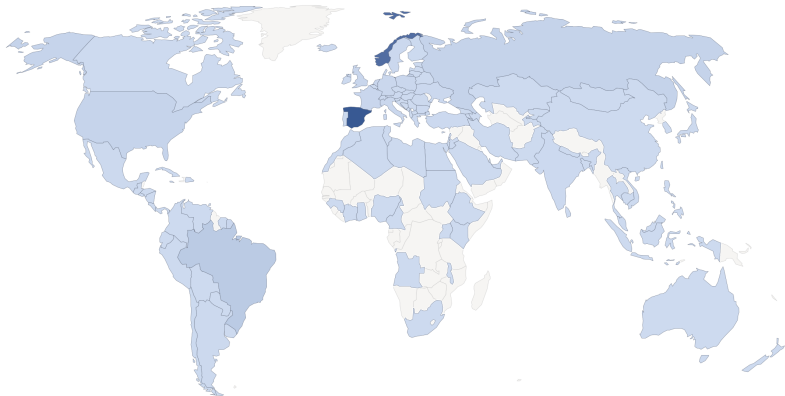


Major users

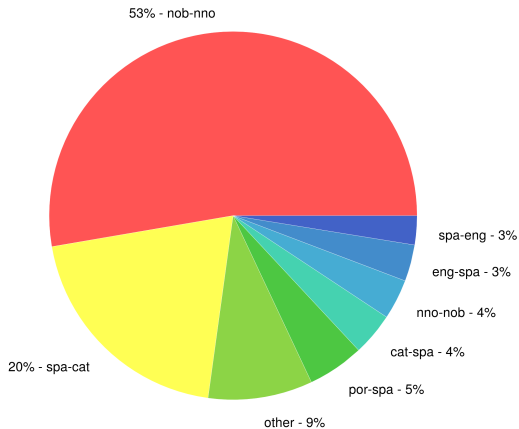
- ▶ *La Voz de Galicia*
- ▶ *La Generalitat de Catalunya*
- ▶ *Ofis Publik ar Brezhoneg*
- ▶ *WikiMedia*
- ▶ *Oslo School District* ☺

Online users

Visitor Map



Online translation statistics



Outline

Introduction

- Design
- Development
- Status

Teaching

- Courses
- Google Summer of Code
- Google Code-in

Research

- New language pairs
- Applying unsupervised methods
- Hybrid systems

Future work and challenges

- Challenges
- Plans
- Collaboration



Courses organised

- ▶ Luxembourg, May 2011
- ▶ Shupashkar, January 2012
- ▶ Helsinki, May 2013

- ▶ 2-day course, funded by the European Commission DGT
- ▶ Held at the European Commission
- ▶ Organised for translators
- ▶ Produced an 80-page course book used in later courses

- ▶ 2-day course, funded by the European Commission DGT
- ▶ Held at the European Commission
- ▶ Organised for translators
- ▶ Produced an 80-page course book used in later courses
- ▶ Translators not ever so interested in making MT systems ☺

Shupashkar, January 2012

- ▶ 5-day course, funded by Apertium
- ▶ Held in Shupashkar (Cheboksary), Chuvashia
- ▶ 8 modules, covering all aspects of the system
- ▶ 3 teachers
- ▶ 30 participants from Russia
- ▶ At least two participants have gone on to work on MT

- ▶ Course organised at the Dept. Linguistics
- ▶ 3 ECTS credits!
- ▶ Around 20 students
- ▶ Uralicists and linguistics students

Google Summer of Code

What is the Google Summer of Code ?

- ▶ Offers stipends (\$5,500, 12 weeks) for students
- ▶ Open-source projects propose 3-month project ideas
- ▶ Students apply for projects

Apertium in the Google Summer of Code:

- ▶ Participated: 2009, 2010, 2011, 2012, 2013, 2014
- ▶ Around 60 completed projects
- ▶ Projects can be:
 - ▶ Language pairs
 - ▶ Engine development
 - ▶ Interfaces
 - ▶ Research work
- ▶ Massive organisation effort



Important projects

2009:	Norwegian Bokmål–Nynorsk
2010:	HFST integration
2011:	–
2012:	Kazakh–Tatar Apertium for Android
2013:	Finite-state constraint grammar
2014:	Assimilation evaluation toolkit

Outcomes

- ▶ Student retention rate: around 10%
- ▶ Over 10 released language pairs
- ▶ Approx. €5,000 / year for project funds!
 - ▶ Sending students to conferences
 - ▶ Organising workshops
 - ▶ Funding limited non-research work

Google Code-in

What is the Google Code-in ?

- ▶ Competition organised for 13–17 year olds
- ▶ Organised as 'tasks'
 - ▶ One task should take experienced developer 2 hours
 - ▶ Each task is worth one point
- ▶ Two kids from each organisation get fully-paid trip to SF

Apertium in the Google Code-in:

- ▶ Participated: 2011, 2012, 2013, 2014
- ▶ Thousands of completed tasks
- ▶ Tasks can be:
 - ▶ Language pairs
 - ▶ Engine development
 - ▶ Interfaces
 - ▶ Research work
- ▶ Massive organisational effort



Outcomes

- ▶ Afrikaans–Dutch language pair
- ▶ Support for compound words in `1tttoolbox`
- ▶ Free tagged corpus of English WP articles (30,000 words)
- ▶ Python interface to Apertium
- ▶ New web site



Outcomes

- ▶ Afrikaans–Dutch language pair
- ▶ Support for compound words in `1tttoolbox`
- ▶ Free tagged corpus of English WP articles (30,000 words)
- ▶ Python interface to Apertium
- ▶ [New web site](#)



Outcomes

Apertium: an open-source machine translation engine and toolbox - Iceweasel

Fitxer Editar Visualitza Historial Adreces d'interès Eines Ajuda

Apertium | Una pl... x Apertium: an ope... x

xixona d'isi ua.es/~fran/webpace/

24-11-2011: ibeque English 6.3.0 Released, [read more...](#)

Apertium

Plataforma lliure de codi font obert per a la traducció automàtica

[proveu-lo](#) | [què es Apertium?](#) | [descàrregues](#) | [documentació](#) | [contacte](#)


catallà Catal

Traducció de textos

Apertium ofereix tant traducció de textos, com a traducció de documents i navegar i traduir. També es pot buscar als diccionaris en línia i provar les versions en desenvolupament dels nostres traductors.

Directori:

Mark unknown words in



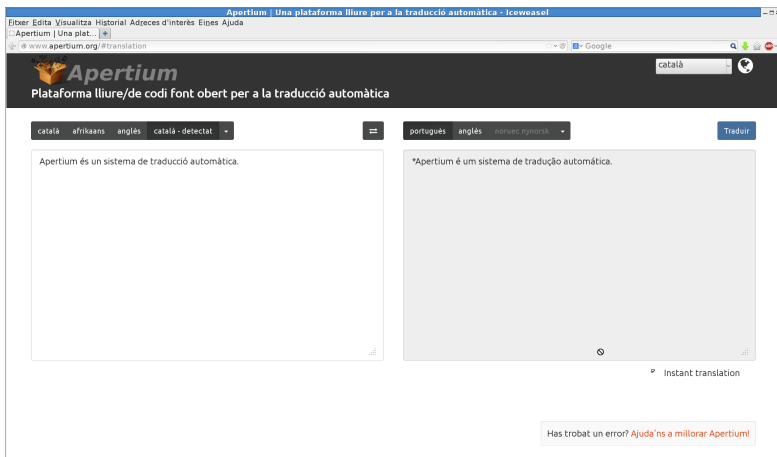
Copyright © 2011 - 2010 Universitat de València

Copyright © 2011 - 2010 Universitat de València

BYTEMARK



Outcomes



The screenshot shows the Apertium web interface in a browser window. The browser's address bar displays `www.apertium.org/#translation`. The page header includes the Apertium logo and the text "Plataforma lliure/de codi font obert per a la traducció automàtica". Below the header, there are two language selection dropdown menus. The left menu is set to "català" and "afrikaans", and the right menu is set to "português" and "anglès". A blue "Traduir" button is located to the right of the second dropdown. The main content area is split into two columns. The left column contains the text "Apertium és un sistema de traducció automàtica." and the right column contains the translated text "*Apertium é um sistema de tradução automática." Below the right column, there is a link for "Instant translation". At the bottom of the page, a message box says "Has trobat un error? Ajuda'ns a millorar Apertium!".

Apertium | Una plataforma lliure per a la traducció automàtica - Icceweasel

Editar Editar Visualitza Historial Adreces d'interès Eipex Ajuda

Apertium | Una plat...

www.apertium.org/#translation

Google

Apertium

Plataforma lliure/de codi font obert per a la traducció automàtica

català

català afrikaans anglès català - detectat

português anglès noruec nynorsk

Traduir

Apertium és un sistema de traducció automàtica.

*Apertium é um sistema de tradução automática.

Instant translation

Has trobat un error? Ajuda'ns a millorar Apertium!



Outline

Introduction

- Design
- Development
- Status

Teaching

- Courses
- Google Summer of Code
- Google Code-in

Research

- New language pairs
- Applying unsupervised methods
- Hybrid systems

Future work and challenges

- Challenges
- Plans
- Collaboration



New language pairs

Creating a new language pair involves:

- ▶ Describing the morphology as a finite-state transducer
- ▶ Writing a disambiguation grammar, or annotating a corpus
- ▶ Constructing a bilingual dictionary
- ▶ Developing a contrastive grammar of two languages

Many languages that Apertium works with have

- ▶ Limited grammatical descriptions
- ▶ Few or inexistent lexical resources



Unsupervised methods

« Learning statistical models from unlabelled data »

In Apertium:

- ▶ How to improve an existing MT system
- ▶ Remove a module from the pipeline and try and relearn it
- ▶ Applied to part-of-speech tagging and lexical selection

General method:

- ▶ Generate all possible outcomes
- ▶ Score the outcomes on a target-language model
- ▶ Use fractional counts as input into supervised algorithm



Unsupervised: Lexical selection

Monolingual corpus

Koivun syyt ovat yleensä suorat.

Se on sinun syytäsi.

Uniapnean syyt ja siihen vaikuttavat tekijät.

Ilmastonmuutoksen syyt ja seuraukset.

Onko tarkoitus tukkia puun syyt?

Petsi myös jättää puun syyt enemmän näkyviin.



Unsupervised: Lexical selection

Expand possible translations

Koivun syyt ovat yleensä suorat.	Birch grains are usually straight. Birch reasons are usually straight. Birch faults are usually straight. It is your grain.
Se on sinun syytäsi.	It is your reason. It is your fault.
Uniapnean syyt ja siihen vaikuttavat tekijät.	Sleep apnea reasons and factors contributing to it. Sleep apnea grains and factors contributing to it. Sleep apnea faults and factors contributing to it.
Ilmastomuutoksen syyt ja seuraukset.	Global warming reasons and consequences. Global warming grains and consequences. Global warming faults and consequences.
Siinä pintaan maalataan puun syyt näkyviin.	There on surface is painted tree reasons become visible. There on surface is painted tree grains become visible. There on surface is painted tree faults become visible. There on surface is painted wood reasons become visible. There on surface is painted wood grains become visible. There on surface is painted wood faults become visible.
Petsi myös jättää puun syyt enemmän näkyviin.	Stain also leaves tree reasons more visible. Stain also leaves tree grains more visible. Stain also leaves tree faults more visible. Stain also leaves wood reasons more visible. Stain also leaves wood grains more visible. Stain also leaves wood faults more visible.



Unsupervised: Lexical selection

Score on language model and normalise

Koivun syyt ovat yleensä suorat.	Birch grains are usually straight.	0.20
	Birch reasons are usually straight.	0.36
	Birch faults are usually straight.	0.44
	It is your grain.	0.04
Se on sinun syytäsi.	It is your reason.	0.12
	It is your fault.	0.84
	Sleep apnea reasons and factors contributing to it.	0.55
Uniapnean syyt ja siihen vaikuttavat tekijät.	Sleep apnea grains and factors contributing to it.	0.26
	Sleep apnea faults and factors contributing to it.	0.19
	Global warming reasons and consequences.	0.98
	Global warming grains and consequences.	0.01
Ilmastonmuutoksen syyt ja seuraukset.	Global warming faults and consequences.	0.01
	There on surface is painted tree reasons become visible.	0.01
	There on surface is painted tree grains become visible.	0.01
	There on surface is painted tree faults become visible.	0.01
Siinä pintaan maalataan puun syyt näkyviin.	There on surface is painted wood reasons become visible.	0.03
	There on surface is painted wood grains become visible.	0.92
	There on surface is painted wood faults become visible.	0.02
	Stain also leaves tree reasons more visible.	0.01
	Stain also leaves tree grains more visible.	0.01
	Stain also leaves tree faults more visible.	0.32
Petsi myös jättää puun syyt enemmän näkyviin.	Stain also leaves wood reasons more visible.	0.01
	Stain also leaves wood grains more visible.	0.06
	Stain also leaves wood faults more visible.	0.59



Unsupervised: Lexical selection

Extract n -grams and count

n-gram	'reason'	'grain'	'fault'
syyt	2.07	1.51	2.42
koivun syyt	0.36	0.20	0.44
syyt ovat	0.36	0.20	0.44
sinun syytäsi	0.12	0.04	0.84
uniapnean syyt	0.55	0.26	0.19
syyt ja	1.53	0.27	0.20
ilmastonmuutoksen syyt	0.98	0.01	0.01
puun syyt	0.06	1.00	0.94
syyt näkyviin	0.04	0.93	0.03
syyt enemmän	0.02	0.07	0.91



Unsupervised: Lexical selection

What then ?

- ▶ Take counts and feed into ML-algorithm of choice
 - ▶ For example MaxEnt, as in my thesis
- ▶ Get the same quality as TL-model, using only SL-information
- ▶ No need to make multiple translations at runtime!
- ▶ Learn translation probabilities without a parallel corpus!



Hybrid systems

Using rule-based systems inside SMT:

- ▶ Many approaches:
 - ▶ Synthetic data (either wordforms or phrases)
 - ▶ Incorporating linguistic information (morphology or syntax)
- ▶ A system with data from Apertium came second in WMT

Using statistics in rule-based systems:

- ▶ Most rule-based systems already have statistics to some extent



Outline

Introduction

- Design
- Development
- Status

Teaching

- Courses
- Google Summer of Code
- Google Code-in

Research

- New language pairs
- Applying unsupervised methods
- Hybrid systems

Future work and challenges

- Challenges
- Plans
- Collaboration



Major challenges

- ▶ Coverage
- ▶ Finding and motivating users

Coverage

Given adequate ($> 95\%$) coverage, we are competitive with Google translate:

- ▶ Evaluations of Slovenian→Serbo-Croatian, Afrikaans→Dutch, Swedish→Danish, Danish→Norwegian, Maltese→Arabic have shown this

However, most pairs are prototypes:

- ▶ Coverage around 80%–85%
- ▶ Increasing lexical coverage beyond this is pretty boring
- ▶ Difficult to motivate people



Finding users

Challenge:

- ▶ Apertium often finds itself translating in ‘non-canonical’ translation directions
- ▶ People are fond of translating from English (or French)

Successes:

- ▶ Spanish–Portuguese (Prompsit + AutoDesk)
- ▶ Romance languages in general

“to do”:

- ▶ Turkic languages
- ▶ Slavic languages
- ▶ Uralic languages



Factoring out language independent resources

- ▶ Apertium began with three language pairs, now has thirty
- ▶ Most language pairs developed by *copying* previous data
- ▶ Result → 13 copies of Spanish morphological dictionary
- ▶ This year we started to separate out monolingual data
- ▶ Now many languages have their own language directory
- ▶ Pairs then depend on a single monolingual source



Factoring out language independent resources

- ▶ Minor tagset differences that were introduced
- ▶ Classification differences
- ▶ “Multiwords”

Pair	Example
spa-cat	el<det><def><f><sg> mismo<adj><f><sg> casa<n><f><sg>
spa-eng	el mismo<det><ind><f><sg> casa<n><f><sg>

- ▶ Problem #1: Generation errors
- ▶ Problem #2: Solving these errors will not solve any problems

Tree-based transfer

Languages with radically different word order could benefit from:

- ▶ Long distance reordering and agreement
- ▶ “recursive” transfer rules

Current prototype inspired by METAL and MorphoLogic:

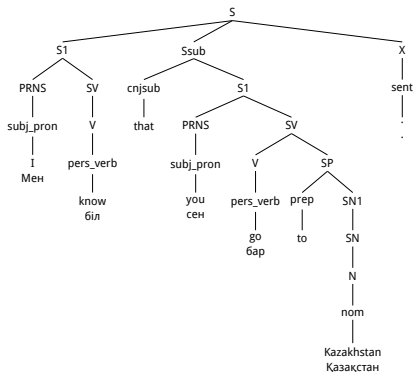
- ▶ Bison (CFG) based transformer
- ▶ Build target-language tree as you parse
- ▶ Pattern-Action, like current Apertium

`Ssub → that SV { $2 $1 }`

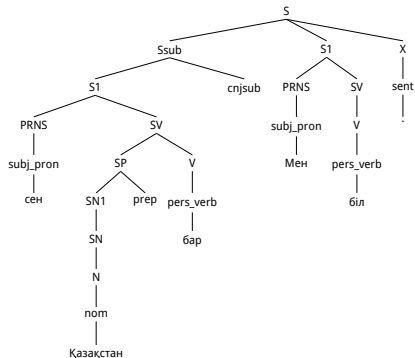


Tree-based transfer

Subordinate clauses in English→Kazakh



I know that you went to Kazakhstan.



Мен сіздің Қазақстанға барғаныңызды білемін.
Minä sinun Kazakstaniin menneen tiedän.



Weighted pipeline

Apertium includes probabilistic components in the pipeline:

- ▶ Part-of-speech tagger calculates tag probabilities
- ▶ Lexical-selection module allows for translation probabilities

However:

- ▶ No probabilistic information encoded in the lexicon
- ▶ Rules are not weighted

In general:

- ▶ We should be able to take into account weights at all stages
- ▶ We should never output something less probable
- ▶ But rules should always allow us to control the output
 - ▶ And output should be predictable!



Lexicon graph

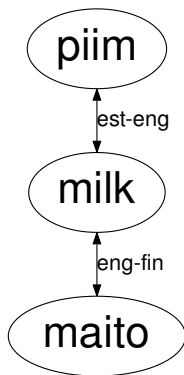
- ▶ There are 40 'released' bilingual dictionaries in Apertium
- ▶ We have a program that 'crosses' two dictionaries, e.g.
 - ▶ Spanish-French + Spanish-Occitan → French-Occitan

Questions:

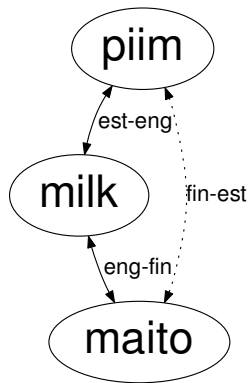
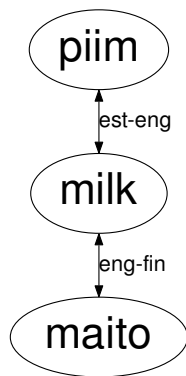
- ▶ What happens with ambiguities ?
- ▶ Can all other dictionaries help ?



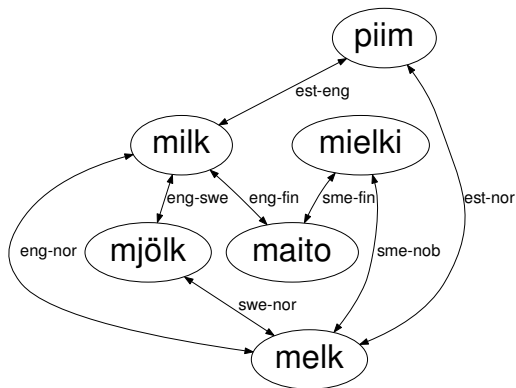
Lexicon graph



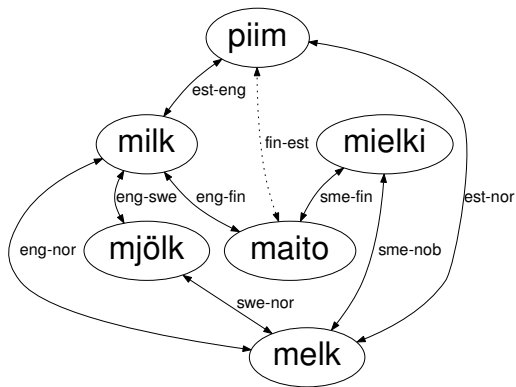
Lexicon graph



Lexicon graph



Lexicon graph



Constructions

- ▶ Some things don't translate well if you translate word-by-word!
- ▶ The meaning is not so much in the "content" words, but in the expression.



Constructions

- ▶ Some things don't translate well if you translate word-by-word!

Ei oo kaikki muumit laaksossa
Not all Moomin's are in the valley

- ▶ The meaning is not so much in the “content” words, but in the expression.



Constructions

- ▶ Some things don't translate well if you translate word-by-word!

Ei oo kaikki muumit laaksossa
Not all Moomins are in the valley

A few sandwiches short of a picnic

Ei oo kaikki X:t Y:ssa \rightarrow A few X short of a Y.
« A few Moomins short of a valley »

- ▶ The meaning is not so much in the “content” words, but in the expression.



Software-engineering practices

With large software projects, following good engineering principles is often a challenge

- ▶ When Apertium was only pair-based this was easier. Each pair was an island.
- ▶ Now we have separated out monolingual data, changes in one place can have knock-on effects in other places.
- ▶ We have no full test suite that developers can rely on.



Collaboration

Developers working on rule-based systems should collaborate!

- ▶ **Morphological descriptions**
 - ▶ Lemma lists categorised by morphological paradigm
- ▶ **Dictionaries**
 - ▶ Bi-/ multi-lingual correspondences between lemma + POS
- ▶ **Ideas**
 - ▶ Anything you can think of!



Giitu · Takk · Kiitos · Tack!

