

A modular system for porting advanced interactive programs to new, morphology-rich languages

Lene Antonsen, Ciprian Gerstenberger,
Ryan Johnson, Trond Trosterud, Heli Uibo
Centre for Sami Language Technology
<http://giellatekno.uit.no/>



Contents

Introduction

Programs

Group 1: Lexicon excercises

Group 2: Morphology

Group 3: Morphology + syntax

Usage statistics for Oahpa

Programming

A general point

Conclusion

Introduction

- ▶ Idea:
 - ▶ Port our interactive programs OAHPA! to other languages

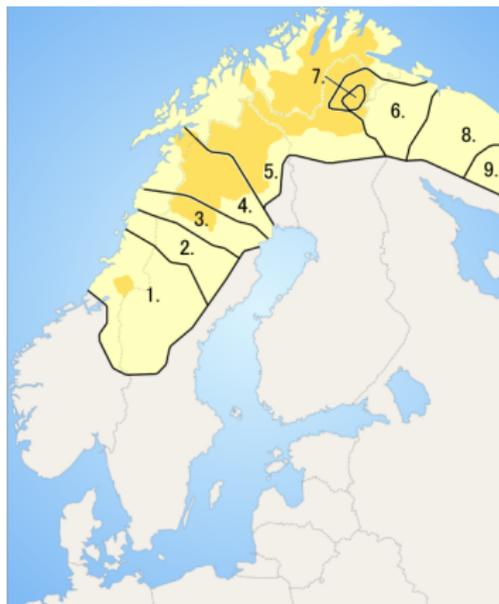
Introduction

- ▶ Idea:
 - ▶ Port our interactive programs OAHPA! to other languages
- ▶ What we have:
 - ▶ An open-source infrastructure for advanced interactive teaching of morphologically complex languages

Introduction

- ▶ Idea:
 - ▶ Port our interactive programs OAHPA! to other languages
- ▶ What we have:
 - ▶ An open-source infrastructure for advanced interactive teaching of morphologically complex languages
- ▶ What you need in order to join in
 - ▶ Basic vocabulary, a morphological analyser, (a syntactic analyser)
 - ▶ A language teacher, a programmer, and a computational linguist

The Sami Languages



- 1. South Sami
- 2. Ume Sami
- 3. Pite Sami
- 4. Lule Sami
- 5. North Sami
- 6. Skolt Sami
- 7. Inari Sami
- 8. Kildin Sami
- 9. Ter Sami

Darkened area
represents
municipalities that
recognize Sami as
an official language.

North Sami

- ▶ Morphologically complex – a suffixing language with many stem-changing processes
 - ▶ «a combination of Turkish and Icelandic»
- ▶ Inflects nouns in 7 cases, and verbs in 3 persons and 3 numbers
- ▶ Does not use «yes» or «no» in turntaking
 - ▶ ... but answers «yes» with repeating the verb but changing the inflection, and «no» with inflecting the negation verb in person and number

North Sami

- ▶ Morphologically complex – a suffixing language with many stem-changing processes
 - ▶ «a combination of Turkish and Icelandic»
- ▶ Inflects nouns in 7 cases, and verbs in 3 persons and 3 numbers
- ▶ Does not use «yes» or «no» in turntaking
 - ▶ ... but answers «yes» with repeating the verb but changing the inflection, and «no» with inflecting the negation verb in person and number
- ▶ This calls for a learning methodology with focus on word inflection

The pedagogical motivation behind OAHPA!

To develop a language tutoring system which

- ▶ has free-form dialogues and sophisticated error analysis
- ▶ gives immediate error feedback and advice to the user
- ▶ is flexible
- ▶ is easily integrated to the instruction in school and university
- ▶ enables the choice of main dialect and metalanguage
- ▶ is freely accessible via Internet

ICALL programs – <http://oahpa.no/davvi/>

HELP

OAHPA!

Bures boahтин!

Veahkkegiella
English

Suopman
Guovdageaidnu

<p>MORFA-S</p>  <p>Hárjehala sojahit sániid</p>	<p>VASTA</p>  <p>Vástit gažaldagaide. Sániit ja jorgalusat Answer to questions</p>	<p>LEKSA</p>  <p></p>
<p>MORFA-C</p>  <p>Hárjehala sojahit sániid cealkagis</p>	<p>SAHKA</p>  <p>Ságastallamat</p>	<p>NUMRA</p>  <p>Hárjehala loguid</p>

OAHPA lea interneahhtaprográmma nuoraide ja rávesolbmuide geat leat oahpahallame davvisámegiela. Prográmma sáhtát heivehit fáttáid ja dási mielde, ja odđa bargobihtát ráhkaduvvojit automáhtalaččat.

ICALL programs – <http://oahpa.no/>

HELP

Veahkkegiella
English

Suopman
Guovdageaidnu

OAHPA!

Bures boahthin!

MORFA-S  Hárjehala sojahit sániid	VASTA  Vástit gažaldanaide. Sániid ja jorgalusat Answer to questions	LEKSA  Sániid ja jorgalusat Answer to questions
MORFA-C  Hárjehala sojahit sániid cealkagis	SAHKA  Ságastallamat	NUMRA  Hárjehala logiid

OAHPA lea internethtaprográmma nuoraide ja rávesolbmuide geat leat oahpahallame davvisámegiela. Prográmma sáhtát heivehit fáttáid ja dási mielde, ja odđa bargobihtát ráhkaduvvojit automáhtalaččat.

ICALL programs – <http://oahpa.no/>

HELP

OAHPA!

Bures boahhtin!

Veahkkegiella
English

Suopman
Guovdageaidnu

MORFA-S  Hárjehala sojahit sániid	VASTA  Vástit gažaldagaide. Sániit ja jorgalusat Answer to questions	LEKSA 
MORFA-C  Hárjehala sojahit sániid cealkagis	SAHKA  Ságastallamat	NUMRA  Hárjehala loguid

OAHPA lea internettaográmma nuoraide ja rávesolbmuide geat leat oahpahallame davvisámegiela. Prográmma sáhtát heivehit fáttáid ja dási mielde, ja ođđa bargobihtát ráhkaduvvojit automáhtalaččat.

ICALL programs – <http://oahpa.no/>

The screenshot shows the OAHPA! web interface. At the top left is a 'HELP' button. In the center, the title 'OAHPA!' is written in a large, hand-drawn font, with the subtitle 'Bures beahtin!' below it. On the top right, there are two dropdown menus: 'Veahkkegiella' (Language) set to 'English' and 'Suopman' (Country) set to 'Guovdageaidnu'. The main content area features six activity cards arranged in a 2x3 grid. The central card, 'VASTA', is circled in blue. Each card includes a title, an illustration, and a description. At the bottom of the page, there is a paragraph of text in Sami.

Activity	Icon	Description
MORFA-S	Red crescent moon	Hárjehala sojahit sániid
VASTA	Green vegetable character	Vástit gažaldanaide. Sániit ja jorgalusat Answer to questions
LEKSA	Yellow star	Sániit ja jorgalusat
MORFA-C	Red crescent moon with arrow	Hárjehala sojahit sániid cealkagis
SAHKA	Green vegetable character	Ságastallamat
NUMRA	Blue number 1	Hárjehala loguid

OAHPA lea interneahttaprogámma nuoraide ja rávesolbmuide geat leat oahpahallame davvisámegiela. Prográmma sáhtát heivehit fáttáid ja dási mielde, ja odđa bargobihtát ráhkaduvvojit automáhtalaččat.

A modular system for porting advanced interactive programs to new, morphology-rich languages

└ The pedagogical programs

└ Group 1: Lexicon exercises

Group 1: Lexicon exercises

Group 1: Lexicon exercises

- ▶ Numra — number expressions
 - ▶ ordinal and cardinal numbers
 - ▶ clock
 - ▶ dates

Group 1: Lexicon exercises

- ▶ Numra — number expressions
 - ▶ ordinal and cardinal numbers
 - ▶ clock
 - ▶ dates
- ▶ Leksa — training basic vocabulary
 - ▶ words grouped by semantic domain or textbook
 - ▶ placenames grouped by area

Numra

oahpa.uit.no/aarjel/numra/klokka/

oahpa.uit.no/aarjel/numra/klokka/

OAHPA!

MORFA-R MORFA-B LEKS

Grammar

Select how many points of time to include.

easy

medium

hard

Select the direction

Strings to numerals

Numerals to strings

New set

njealjehs avtelen golme

bielie akte

njealjehs avtelen gaektsie

uktsie

Enter the 10:21)

NUMRA

Cardinals

Ordinals

Clock

Dato

A modular system for porting advanced interactive programs to new, morphology-rich languages

└ The pedagogical programs

└ Group 1: Lexicon exercises

Leksa

oahpa.uit.no/sjdoahpa/lekxa/

LEKSA NUMF

Выберите языки Книга

С кильдин-саамского на русский Все

Места в доме
Путешествие
Погода
Природа
Растения
Рукоделие
Терминология языка
Школа и образование
Выражения из нескольких слов
Церковь
Абстрактные слова
Глаголы – уровень 1
Глаголы – уровень 2
✓ Глаголы – уровень 3
Местоимения
Имена
Ласкательная форма
Все слова

йҥкуувэ

рөдтлаххтэ

ләххтэ

ләлле

әйтнэ

Переведите слова. Вы можете выбрать набор или уровень, но не оба.

Ответы на упражнения

The group 1 programs come as a side effect of the other programs

- ▶ Numra

- ▶ was made as an automaton, we needed it for text-to-speech

The group 1 programs come as a side effect of the other programs

- ▶ Numra
 - ▶ was made as an automaton, we needed it for text-to-speech
- ▶ Leksa
 - ▶ contains the words used for the inflection exercises

Building lexical content

- ▶ Lexicon: approx. 3000 basic words
 - ▶ These may be available from existing teaching material
 - ▶ By marking the vocabulary with textbook, the program may be tailored to specific courses

Building lexical content

- ▶ Lexicon: approx. 3000 basic words
 - ▶ These may be available from existing teaching material
 - ▶ By marking the vocabulary with textbook, the program may be tailored to specific courses
- ▶ Numbers:
 - ▶ We can port the number automaton to other languages
 - ▶ A number-clock-date automaton can be made in less than a day

A modular system for porting advanced interactive programs to new, morphology-rich languages

└ The pedagogical programs

└ Group 2: Morphological exercises

Group 2: Morphology: Morfa

Group 2: Morphology: Morfa

- ▶ Practice inflection
 - ▶ without context
 - ▶ embedded in context

Morfa S

The screenshot shows the Morfa S web application interface for the language Oahpa. At the top, the word "OAHPA!" is written in a stylized, hand-drawn font. To the right of the title are four icons representing different modules: MORFA-R (red crescent), MORFA-B (red crescent), LEKSA (yellow star), and NUMI (blue arrow). Below the title, there are three dropdown menus: "Case" (set to "illative"), "Stem" (with checkboxes for "bisyllabic" and "trisyllabic"), and "Book" (set to "All"). A "Grammar explanations" button is located to the right of these menus. A "New set" button is positioned below the dropdowns. The main content area displays a list of words: "barkoefaaleidahke", "barkoefaaleidahkese", "jávlebiejjieh", "jávlebiejjiide" (with a red border and a "help" icon), "baakoe", "baakose", "tjarme", "tjarmese", "baahkoe", and "baahkose". On the right side, there are two text boxes: "Practise illative" with the instruction "Add nouns in correct forms. You get translation if you click the word." and a note: "'jávlebiejjieh' has an even-syllabled stem. -ide-ending." At the bottom, there are buttons for "Test answers" and "Show the correct answers", and a score display: "Your score: 4/5". A sidebar on the left contains a red crescent icon and a list of menu items: "MORFA-B", "ns", "os", "jectives", "ouns", "erence", "terials", "dning", "ionary", and "nmar".

Morfa C

oahpa.no/davvi/morfac/a

OAHPA!

MORFA-C MORFA-S VASTA SAHKA LEKSA NUN

Veahkkegiel
English

Bargobihát

attributive positive

Ođđa bargobihát

Grammar explanations

Njálggeshildu lea ruoná. Makkár njálggeshildu dát lea? (ruoná)

Diet lea njálggeshildu.

Mu bargobivttas lea vuogas. Makkár bargobivttas mus lea? (vuogas)

Dus lea bargobivttas.

Mu násteboahkánat leat seavdnjat. Makkár násteboahkánat mus leat? (seavdnjat)

Dus leat násteboahkánat.

Bálggis lea oanehaš. Makkár bálggis dát lea? (oanehaš)

Diet lea bálggis.

Mu árgabivttas lea ođđaigásaš. Makkár árgabivttas mus lea? (ođđaigásaš)

Dus lea árgabivttas.

Hárjehala positivva attribuhtahámiid.
Sojat adjektiivvaid. Jus coahkkalat sáni, de oáččut dárogiel jorgalusa.

HELP

MORFA-C

Substantiivvat
Vearbbat
Adjektiivvat
Pronomenat
Lohkosánit
Suorgádusat

Resursat
Bagadus

How it works

- ▶ All the wordforms are stored in a MySQL database
 - ▶ In principle, the paradigms may be typed in manually
 - ▶ We prefer to let a morphological generator make the paradigms automatically

How it works

- ▶ All the wordforms are stored in a MySQL database
 - ▶ In principle, the paradigms may be typed in manually
 - ▶ We prefer to let a morphological generator make the paradigms automatically
- ▶ Wordforms are used both by Leksa and by Morfa
 - ▶ advantage: the students know the words they shall inflect

Morphological analysis / generation

Analyser:

```
$ echo walks | analyse-eng  
walks walk+N+Pl  
walks walk+V+Prs+Sg3
```

Generator:

```
$ echo walk+V+Prs+Sg3 | generate-eng  
walk+V+Prs+Sg3 walks
```

Group 3: Morphology and syntax

- ▶ Answer to open questions (Vasta)
- ▶ Participate in QA drills (Sahka)

Group 3: Morphology and syntax

- ▶ Answer to open questions (Vasta)
- ▶ Participate in QA drills (Sahka)
 - ▶ You may type whatever answer you like
 - ▶ ... and the program will comment upon your agreement and case errors, etc.

Group 3 programs require more resources

- ▶ Prerequisites:
 - ▶ Full-scale lexicon, covering at least 90-95% of running text
 - ▶ Full-coverage morphological analysers
 - ▶ Full-scale Constraint Grammars (CGs) for syntactic analysis

Group 3 programs require more resources

- ▶ Prerequisites:
 - ▶ Full-scale lexicon, covering at least 90-95% of running text
 - ▶ Full-coverage morphological analysers
 - ▶ Full-scale Constraint Grammars (CGs) for syntactic analysis
- ▶ How it works:
 - ▶ **Rules** for error detection
 - ▶ **Error messages** to the user for each error type
 - ▶ **Question frames** for generating open questions (Vasta)
 - ▶ Dialogues with **navigation instructions** (Sahka)

Vasta

OAHPA!



MORFA-C MORFA-S VASTA SAHKA LEKSA NUMRA

Dássi
First level

English

Odda bargobihtát

Grammar explanations

Maid mii oaidnit?
Dii oaidnibehtet nieida. ✘

Iskka vástádusaid
Nominative doesn't go with a transitive verb.

Du čuoggát: 0/1

Vástit olles cealkagiin.
Fuomá ahte jus
jearaldagas lea moai/mii,
de don vástidat doai/dii.

VASTA
Vasta-S
Vasta-F

Resurssat
Bagadus
Neahttasátnegirji
Grammatihkka

Copyright 2012 Romssa universitehta
Contact oahpa@hum.uit.no

[Linjka dán hárhjussii](#)

HELP

Sahka

OAHPA!



MORFA-C MORFA-S VASTA SAHKA LEKSA NUMRA

Vástit olles cealkagiin. Muite ahte báikenamat áiget stuora bustávain. Veahkkegillia English

Buorre beaivi! Bures boahhtin mu geahčái!

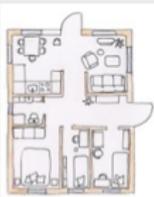
Mun lean aiddo fárren sisa iežan odda orrunsdajái. Mus leat lossa viessogáivvut dáppe feaskáris. Gillešit go veahkehit mu?
Jua, mun gillen veahkehit du.

Mus lea TV dás. Gude lanjas TV lea du orrunsašis?
 ✕

Your answer should contain a [locative](#).

[HELP](#)

SAHKA

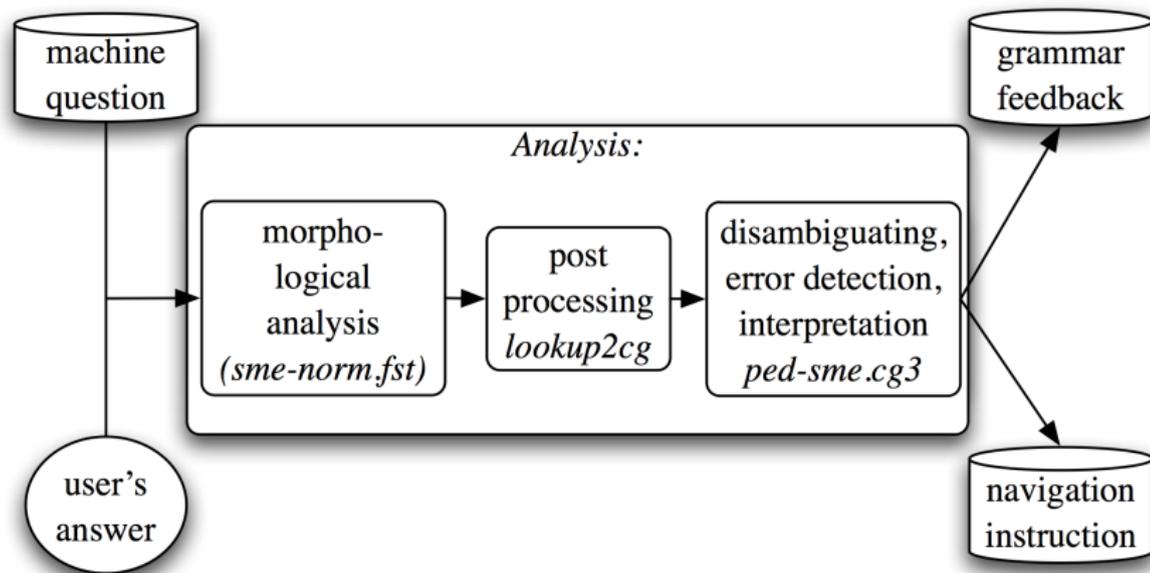


basadanlatnja,
lokta,
oaddenlatnja,
stohpu, feaskkir,
gievkkán, hivsset

Resurssat
Bagadus

Copyright 2012 Romssa universitehta
Contact oahpa@hum.uit.no

Schematic overview of the treatment of the free input



Analysis: Searching for the missing illative

```
"<Gude>"  
  "guhte" Pron Interr Sg Gen &grm-missing-Ill  
"<latnjii>"  
  "latnja" N Sg Ill  
"<moai>"  
  "mun" Pron Pers Du1 Nom  
"<bidje>"  
  "bidjat" V TV Ind Prs Du1  
"<mu>"  
  "mun" Pron Pers Sg1 Gen  
"<TV>"  
  "TV" N ACR Sg Acc  
"<^sahka>"  
  "sahka" QDL where_place_TV  
"<Moai>"  
  "mun" Pron Pers Du1 Nom  
"<bidje>"  
  "bidjat" V TV Ind Prs Du1  
"<TV>"  
  "TV" N ACR Sg Gen  
"<gievkkanis>"  
  "gievkkkan" N Sg Loc  
"<.>"  
  "." CLB
```

Usage statistics

- ▶ The programs are popular:
 - ▶ The North Sami language community has some 17000 speakers
 - ▶ Our programs get on average 400 queries / workday
- ▶ Our primary target group was adult L2 students
 - ▶ ... the log shows that they are used in primary and secondary schools as well

Oahpa languages

- ▶ All programs:
 - ▶ North Sami
- ▶ Lexicon and morphology
 - ▶ South Sami
- ▶ Experimental versions (lexicon)
 - ▶ Kildin, Skolt and Inari Sami; Russian

Programming

Programming

- ▶ The programs are developed using **Django**
 - ▶ open-source framework for creating web applications supporting the model-view-controller (MVC) design
 - ▶ database-driven applications (Model)
 - ▶ web templates by means of HTML, CSS, jQuery and javascript (View)
 - ▶ implemented in Python (Controller)

Programming

- ▶ Porting the programs to a new language requires relatively few changes:
 - ▶ change settings (paths to linguistic tools, database name and password etc.)
 - ▶ correct the lists of linguistic categories (case lists etc.)
 - ▶ localise the user interface to more languages if needed (this is an automated process where a linguist just has to translate a number of strings in a file)

Example cases

- ▶ Porting the infrastructure of Leksa and Numra from North Sami to Kildin Sami took a couple of days.
- ▶ Porting the infrastructure of Leksa, Numra, Morfa-S and Morfa-C from South Sami to North Sami took a couple of weeks.

Work ahead: Modularising the infrastructure

- ▶ Language-independent and language-specific code should be separated

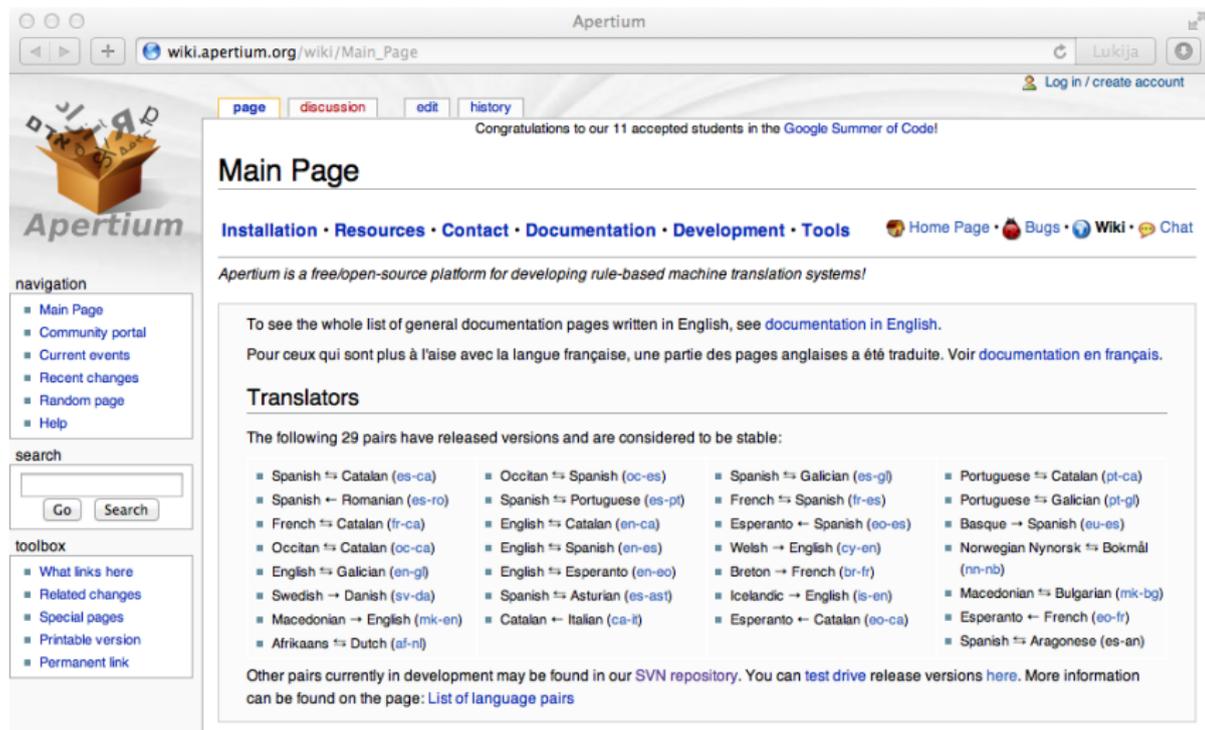
Work ahead: Modularising the infrastructure

- ▶ Language-independent and language-specific code should be separated
- ▶ Today this is only partly the case
 - ▶ Porting to a new language today thus means changing language-specific content of general files

Work ahead: Modularising the infrastructure

- ▶ Language-independent and language-specific code should be separated
- ▶ Today this is only partly the case
 - ▶ Porting to a new language today thus means changing language-specific content of general files
- ▶ Goal: Install new languages by:
 - ▶ exchanging language-specific files
 - ▶ keeping the language-independent infrastructure constant, in separate files

Apertium as a repository of morphological analysers



The screenshot shows the Apertium website's main page in a browser window. The browser's address bar shows `wiki.apertium.org/wiki/Main_Page`. The page has a navigation menu with options like 'page', 'discussion', 'edit', and 'history'. A congratulatory message for Google Summer of Code students is visible. The main heading is 'Main Page', followed by a list of links: 'Installation', 'Resources', 'Contact', 'Documentation', 'Development', and 'Tools'. There are also icons for 'Home Page', 'Bugs', 'Wiki', and 'Chat'. A navigation sidebar on the left includes links to 'Main Page', 'Community portal', 'Current events', 'Recent changes', 'Random page', and 'Help'. A search box is located below the sidebar. A 'toolbox' section at the bottom left provides links for 'What links here', 'Related changes', 'Special pages', 'Printable version', and 'Permanent link'. The main content area features a paragraph in English and French, a 'Translators' section listing 29 stable language pairs, and a note about development versions in the SVN repository.

Apertium

wiki.apertium.org/wiki/Main_Page

page discussion edit history

Congratulations to our 11 accepted students in the [Google Summer of Code!](#)

Main Page

[Installation](#) • [Resources](#) • [Contact](#) • [Documentation](#) • [Development](#) • [Tools](#) [Home Page](#) • [Bugs](#) • [Wiki](#) • [Chat](#)

Apertium is a free/open-source platform for developing rule-based machine translation systems!

To see the whole list of general documentation pages written in English, see [documentation in English](#).

Pour ceux qui sont plus à l'aise avec la langue française, une partie des pages anglaises a été traduite. Voir [documentation en français](#).

Translators

The following 29 pairs have released versions and are considered to be stable:

■ Spanish ↔ Catalan (es-ca)	■ Occitan ↔ Spanish (oc-es)	■ Spanish ↔ Galician (es-gl)	■ Portuguese ↔ Catalan (pt-ca)
■ Spanish ↔ Romanian (es-ro)	■ Spanish ↔ Portuguese (es-pt)	■ French ↔ Spanish (fr-es)	■ Portuguese ↔ Galician (pt-gl)
■ French ↔ Catalan (fr-ca)	■ English ↔ Catalan (en-ca)	■ Esperanto ↔ Spanish (eo-ca)	■ Basque → Spanish (eu-es)
■ Occitan ↔ Catalan (oc-ca)	■ English ↔ Spanish (en-es)	■ Welsh → English (cy-en)	■ Norwegian Nynorsk ↔ Bokmål (nn-nb)
■ English ↔ Galician (en-gl)	■ English ↔ Esperanto (en-eo)	■ Breton → French (br-fr)	■ Macedonian ↔ Bulgarian (mk-bg)
■ Swedish → Danish (sv-da)	■ Spanish ↔ Asturian (es-ast)	■ Icelandic → English (is-en)	■ Esperanto ← French (eo-fr)
■ Macedonian → English (mk-en)	■ Catalan ↔ Italian (ca-it)	■ Esperanto ↔ Catalan (eo-ca)	■ Spanish ↔ Aragonese (es-an)
■ Afrikaans ↔ Dutch (af-nl)			

Other pairs currently in development may be found in our [SVN repository](#). You can [test drive](#) release versions [here](#). More information can be found on the page: [List of language pairs](#)

Languages with morphological resources 1

- ▶ Commonly taught foreign languages:
 - ▶ English, French, German*, Russian, Spanish
- ▶ Nordic states' languages:
 - ▶ Danish, Norwegian, Swedish*, Finnish*, Icelandic, Faroese*
- ▶ Nordic indigenous minority languages:
 - ▶ North Sami, Lule Sami*, South Sami*

* = resources available, but not via Apertium

Languages with morphological resources 2

- ▶ Celtic:
 - ▶ Welsh, Breton, Irish*
- ▶ Romance:
 - ▶ Aragonese, Asturian, Catalan, Galician, Italian, Occitan, Portuguese, Romanian, (Sardinian)
- ▶ Germanic:
 - ▶ Afrikaans, Dutch
- ▶ Uralic:
 - ▶ (Estonian), Hungarian*

(...) = resources available, but not under open licenses

Languages with morphological resources 3

- ▶ Slavic:
 - ▶ Serbo-Croatian, Slovenian, Macedonian, Bulgarian, (Czech), (Polish)
- ▶ Semitic:
 - ▶ Maltese, Arabic
- ▶ Turkic:
 - ▶ Kyrgyz, Kazakh, Tatar, Chuvash, (Bashkir)
- ▶ Other:
 - ▶ Basque, Albanian, (Latvian)

Notable languages missing

- ▶ Slavic:
 - ▶ Belarusian, Rusyn, Slovak, Sorbian, Ukrainian
- ▶ Other:
 - ▶ Greek, Scottish Gaelic, Lithuanian

Conclusion

Conclusion

- ▶ Morphology-rich languages need morphology-aware ICALL programs

Conclusion

- ▶ Morphology-rich languages need morphology-aware ICALL programs
- ▶ Our Oahpa programs may be ported to new languages, by
 - ▶ utilizing a common infrastructure
 - ▶ and reuse linguistic resources from other contexts

Conclusion

- ▶ Morphology-rich languages need morphology-aware ICALL programs
- ▶ Our Oahpa programs may be ported to new languages, by
 - ▶ utilizing a common infrastructure
 - ▶ and reuse linguistic resources from other contexts
- ▶ Your result will be as good as the amount of time and resources you put in...
 - ▶ ... but at least we did the initial developmental work.

Conclusion

- ▶ Morphology-rich languages need morphology-aware ICALL programs
- ▶ Our Oahpa programs may be ported to new languages, by
 - ▶ utilizing a common infrastructure
 - ▶ and reuse linguistic resources from other contexts
- ▶ Your result will be as good as the amount of time and resources you put in...
 - ▶ ... but at least we did the initial developmental work.

Thank you for listening — Any questions?