# 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

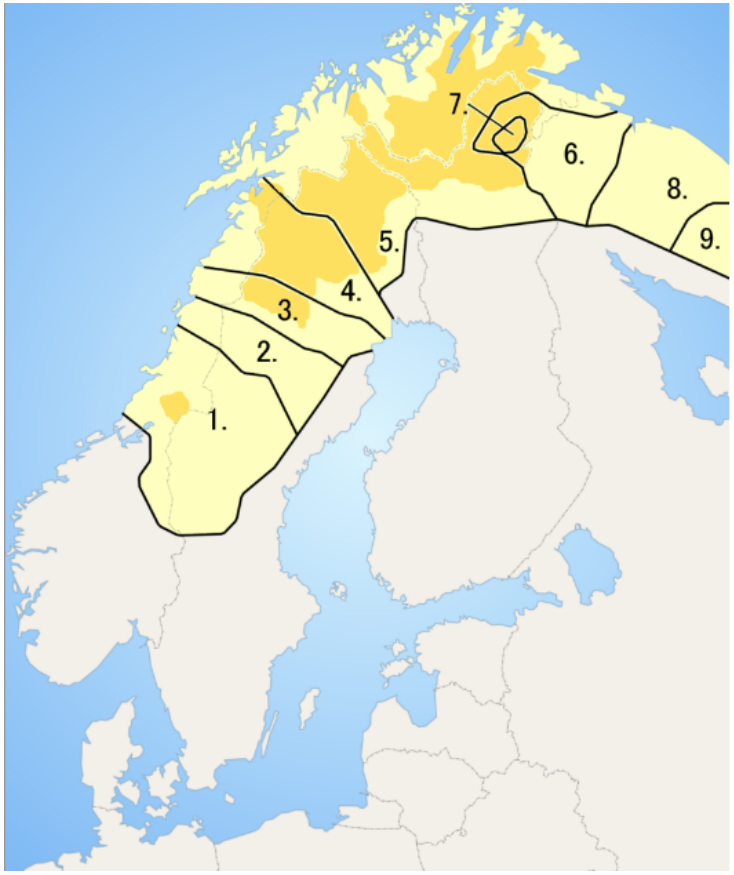
# Introduction

- 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!

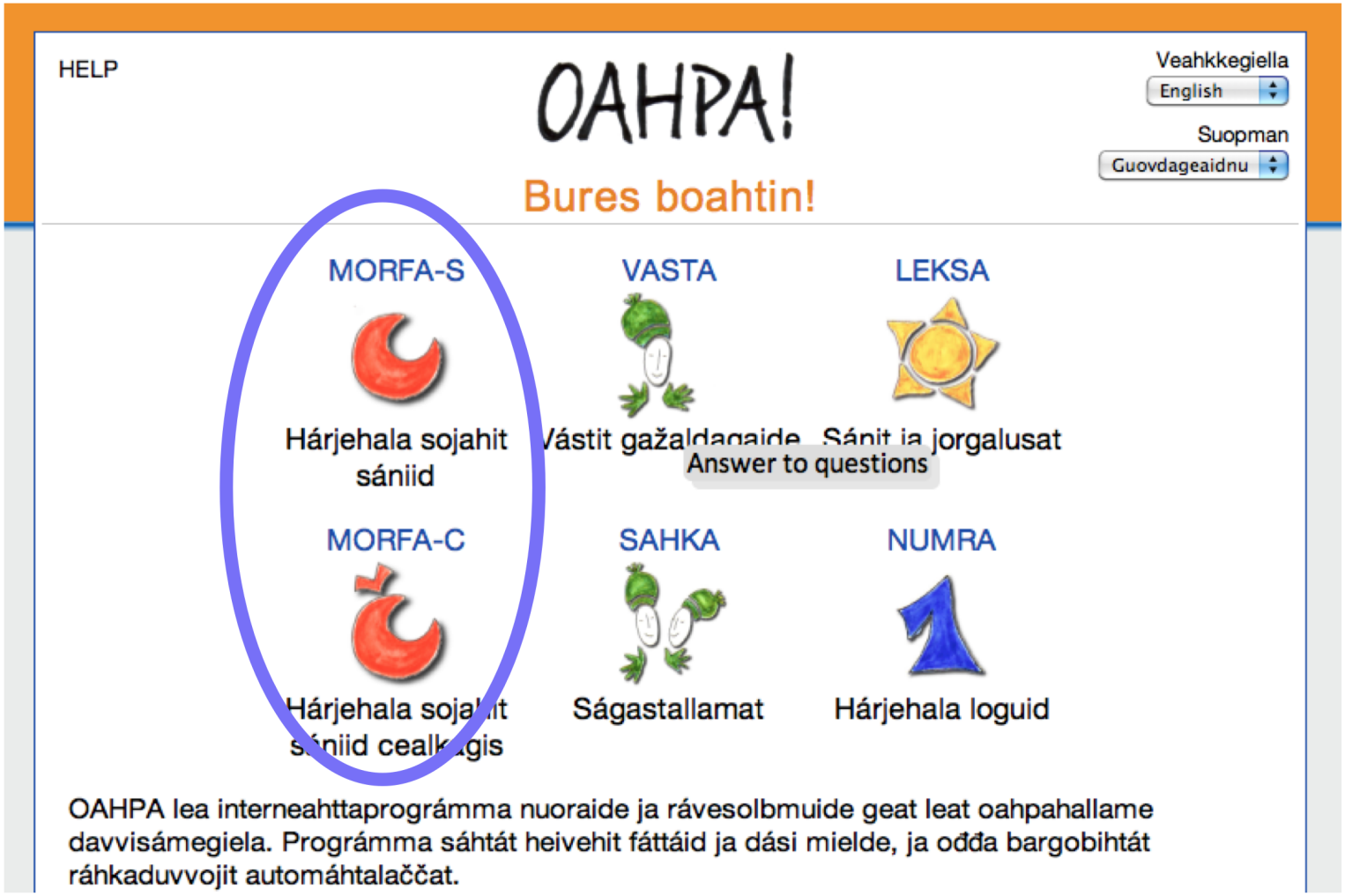
To develop a language tutoring system which

- 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

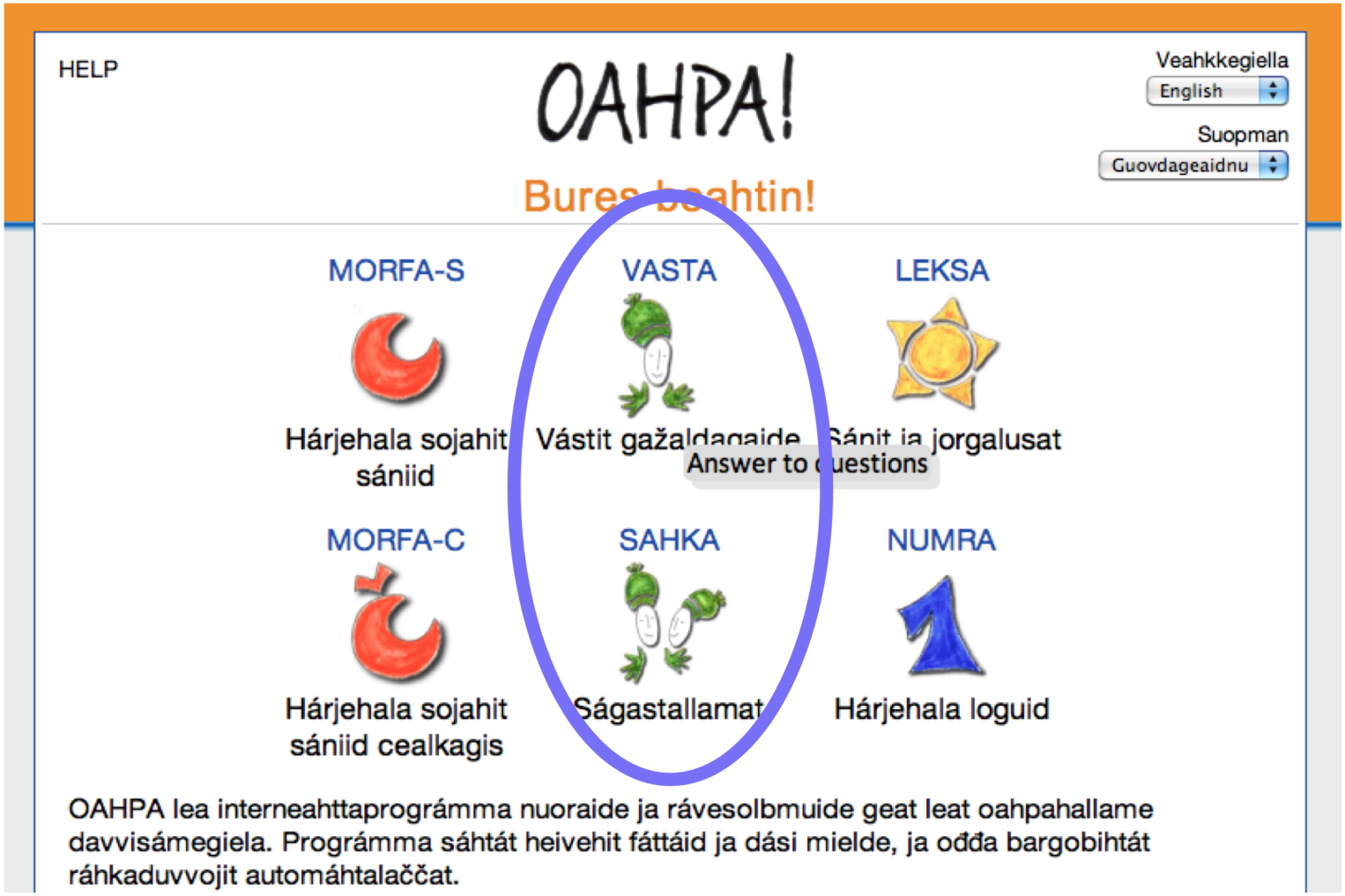
# ICALL programs – http://oahpa.no/davvi/

# ICALL programs – http://oahpa.no/

# ICALL programs – http://oahpa.no/



# ICALL programs – http://oahpa.no/



# Group 1: Lexicon exercises

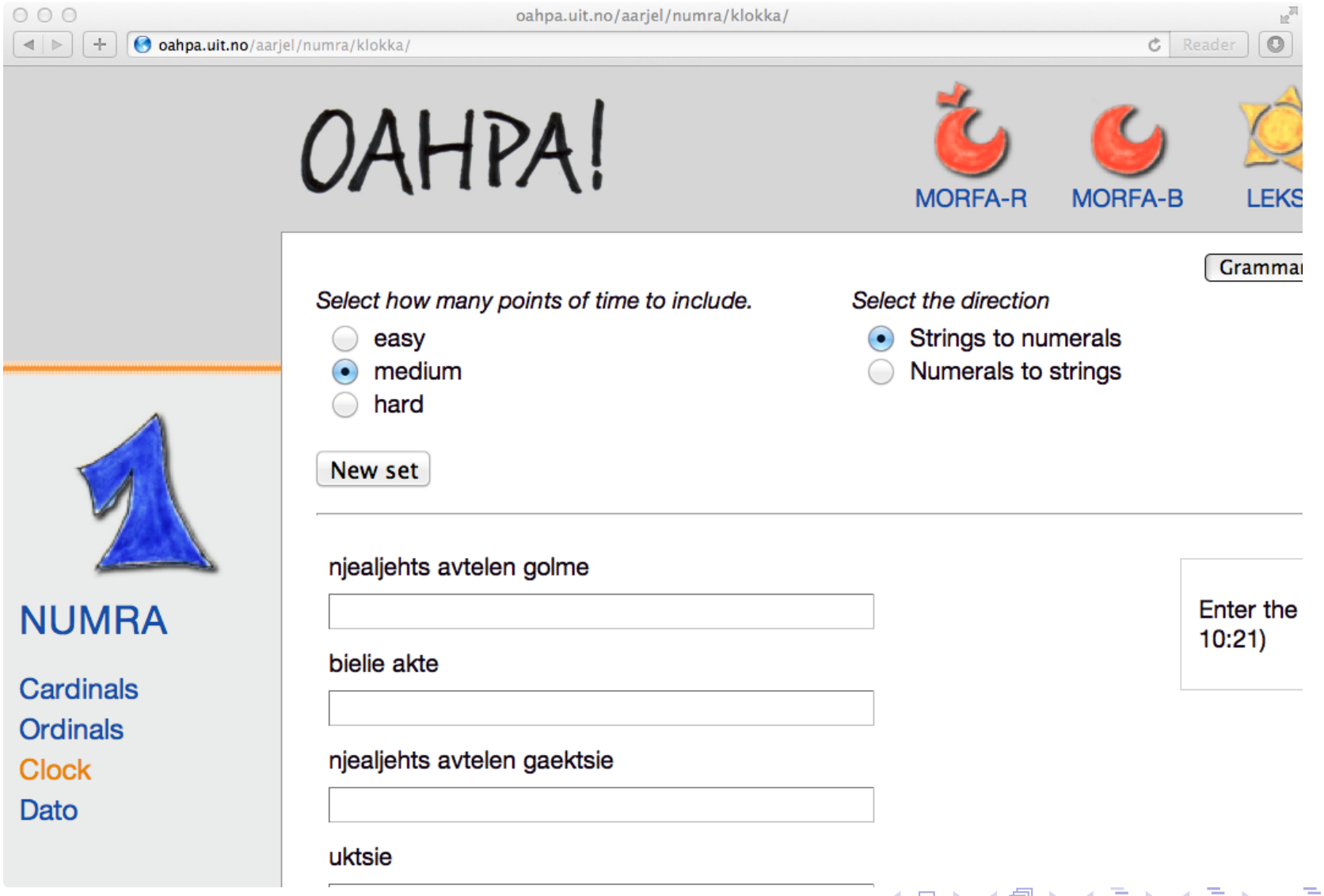
# Group 1: Lexicon exercises

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

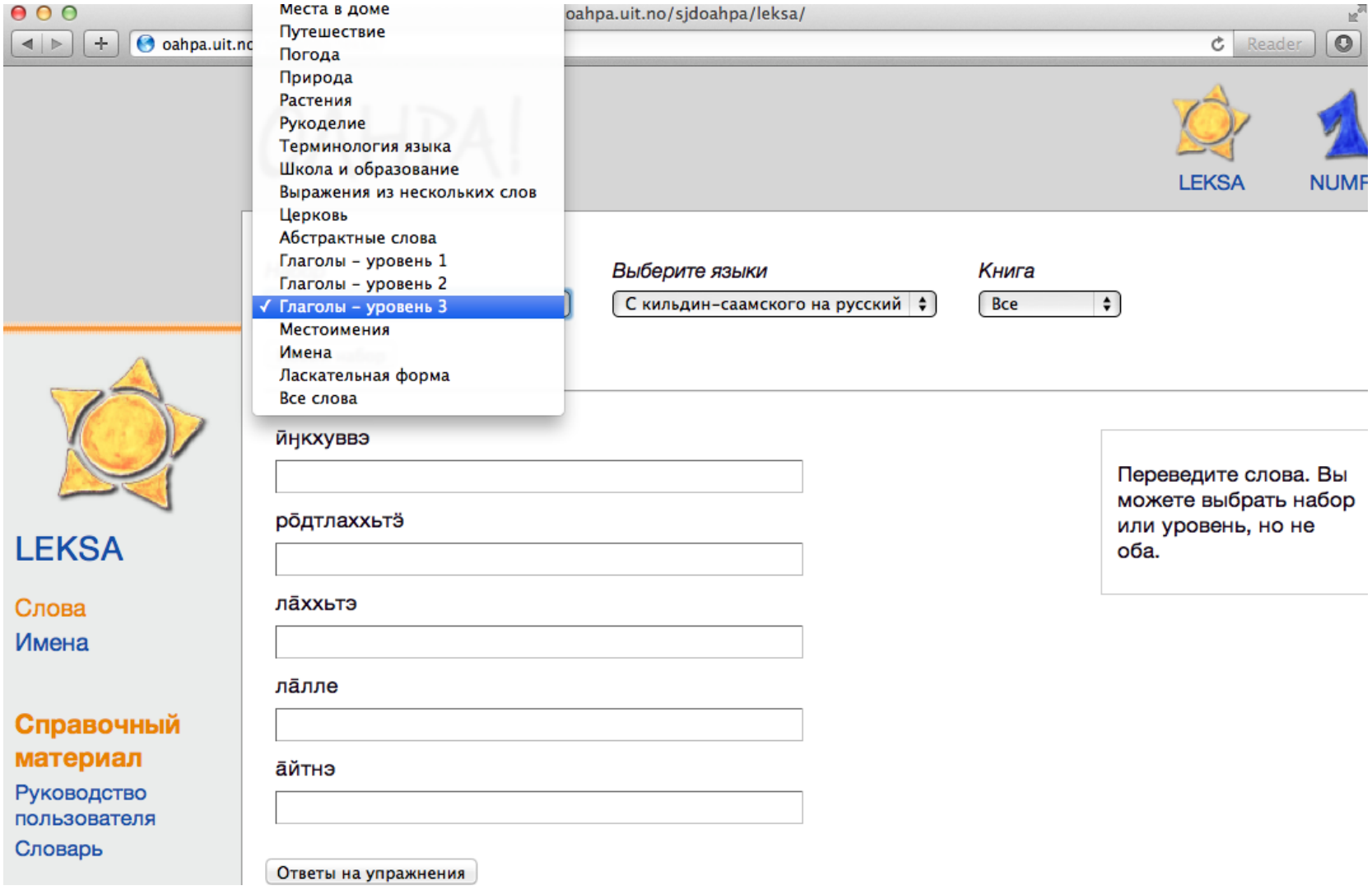
# Group 1: Lexicon exercises

- 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/
Reader
OAHPA!
MORFA-R
MORFA-B
LEKS
Gramma
Select how many points of time to include.
easy
medium
hard
Select the direction
Strings to numerals
Numerals to strings
New set
NUMRA
Cardinals
Ordinals
Clock
Dato
njealjehts avtelen golme
bielie akte
njealjehts avtelen gaektsie
uktsie
Enter the
10:21)

# Leksa


oahpa.uit.no/sjdoahpa/leksa/
Reader
LEKSA
NUMP
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
Выберите языки
С кильдин-саамского на русский
Книга
Все
LEKSA
Слова
Имена
Справочный материал
Руководство пользователя
Словарь
ӣŋкхуввэ
рӧдтлаххьтӭ
лӓ̄ххьтэ
лӓ̄лле
а̄йтнэ
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

# Group 2: Morphology: Morfa

# Group 2: Morphology: Morfa

- Practice inflection
  - without context
  - embedded in context

# Morfa S


OAHPA!
MORFA-R
MORFA-B
LEKSA
Grammar explanations
Case
illative
Stem
bisyllabic
trisyllabic
Book
All
New set
barkoefaaleldahke
barkoefaaleldahkese
jåvlebiejjieh
jåvlebiejjiide
help
baakoe
baakose
tjarme
tjarmese
baahkoe
baahkose
Practise illative
Add nouns in correct forms. You get translation if you click the word.
"jåvlebiejjieh" has an even-syllabled stem. -ide-ending.
Test answers
Show the correct answers
Your score: 4/5

# Morfa C


oahpa.no/davvi/morfac/a/
OAHPA!
MORFA-C
MORFA-S
VASTA
SAHKA
LEKSA
Bargobihtát
attributive positive
English
Ođđa bargobihtát
Grammar explanations
MORFA-C
Substantiivvat
Vearbbat
Adjektiivvat
Pronomenat
Lohkosánit
Suorggádusat
Resurssat
Bagadus
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 seavdnjadat. Makkár násteboahkánat mus leat? (seavdnjat)
Dus leat násteboahkánat.
Bálggis lea oanehaš. Makkár bálggis dát lea? (oanehaš)
Diet lea bálggis.
Mu árgabivttas lea ođđaáigásaš. Makkár árgabivttas mus lea? (ođđaáigásaš)
Dus lea árgabivttas.
Hárjehala positiivva attribuhttahámiid.
Sojat adjektiivvaid. Jus coahkkalat sáni, de oačču dárogiel jorgalusa.
HELP

# 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

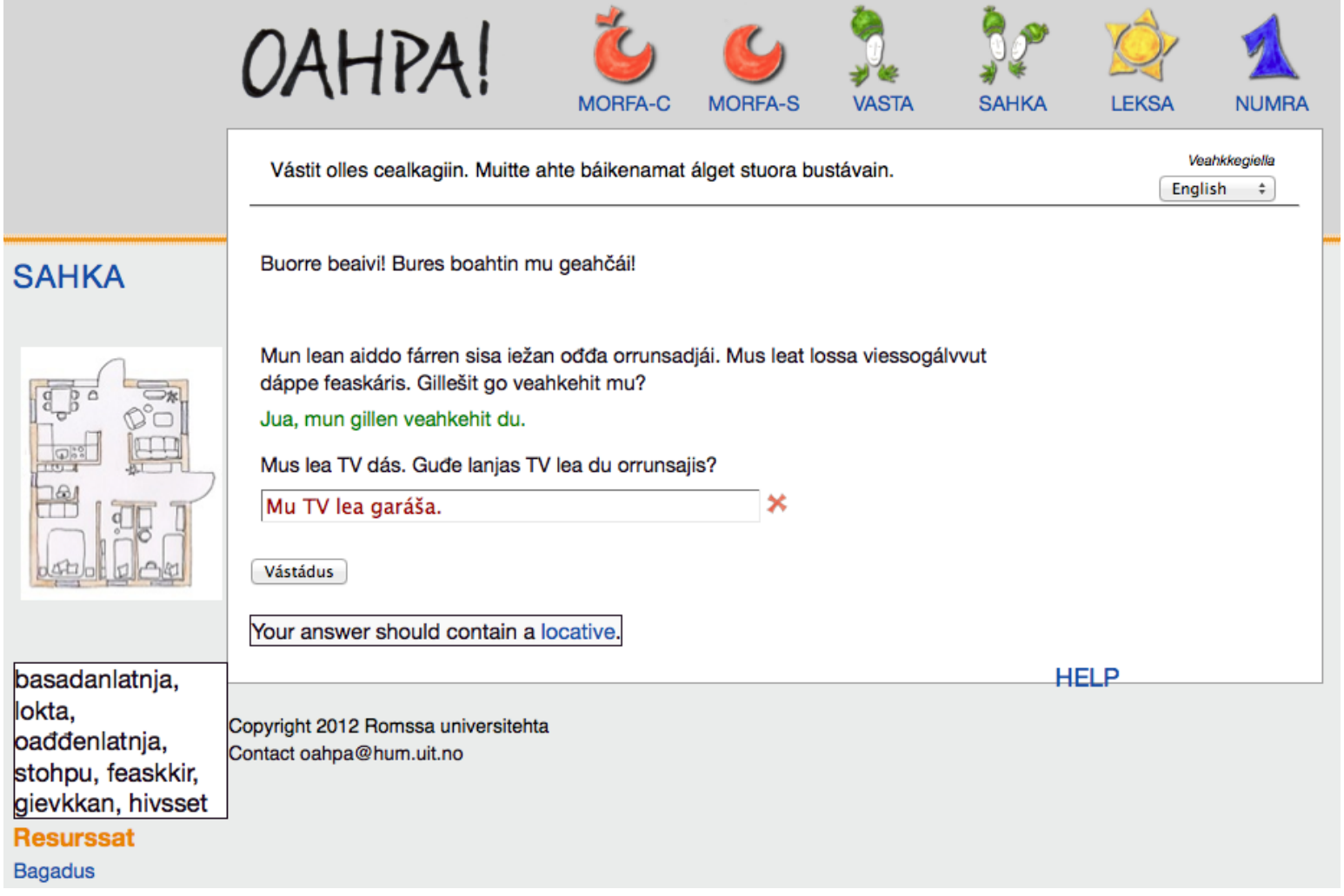
# Group 3 programs require more resources

- 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)

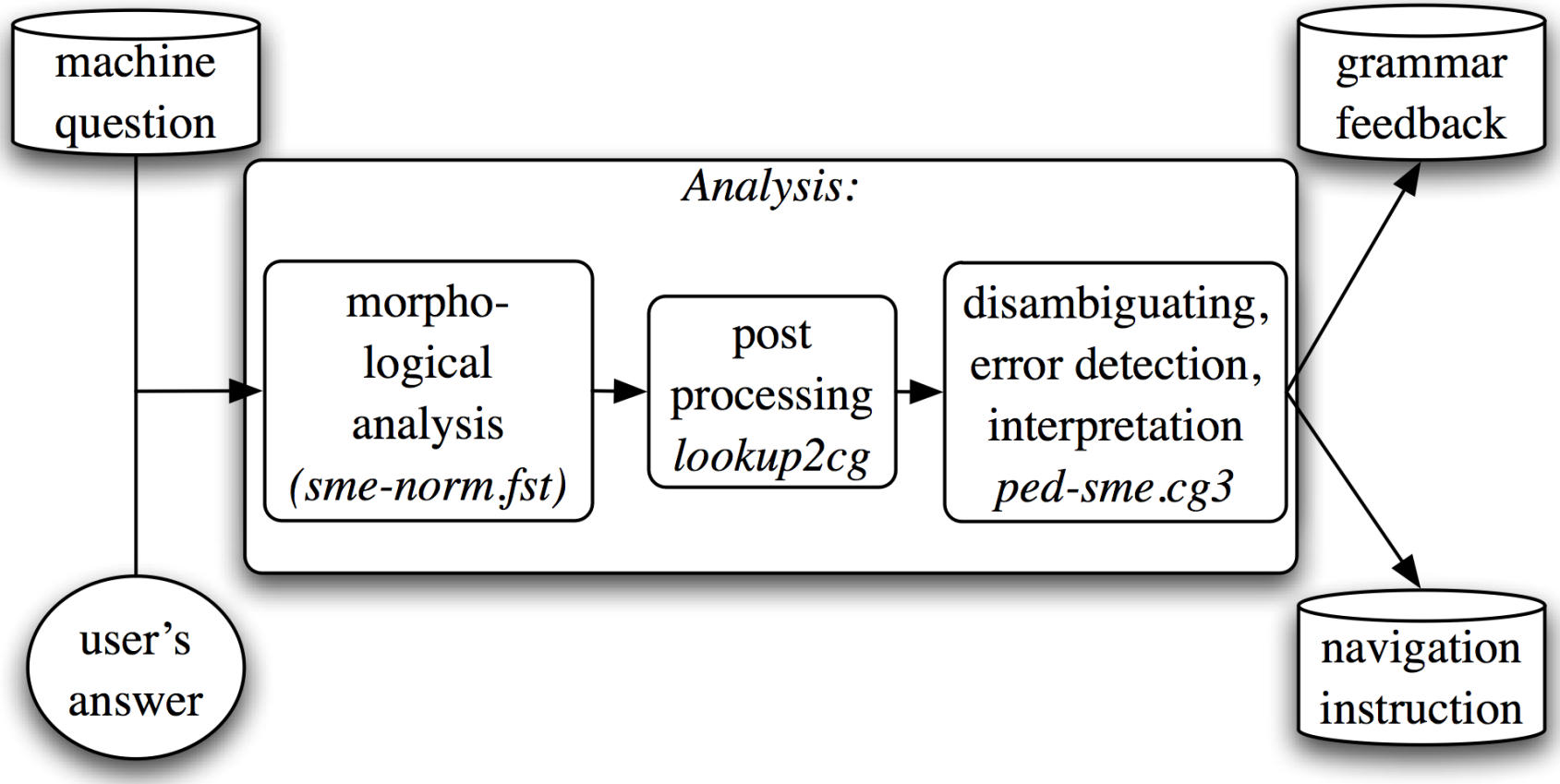
# Vasta


OAHPA!
MORFA-C
MORFA-S
VASTA
SAHKA
LEKSA
NUMRA
Dássi
First level
Veahkkegiella
English
Ođđa bargobihtát
Grammar explanations
Maid mii oaidnit?
Dii oaidnibehtet nieida.
Vástit olles cealkagiin. Fuobmá ahte jus jearaldagas lea moai/mii, de don vástidat doai/dii.
Iskka vástádusaid
Nominative doesn't go with a transitive verb.
Du čuoggát: 0/1
VASTA
Vasta-S
Vasta-F
Copyright 2012 Romssa universitehta
Contact oahpa@hum.uit.no
Liŋka dán hárjehussii
Resurssat
Bagadus
Neahttasátnegirji
Grammatihkka
HELP

# Sahka


OAHPA!
MORFA-C
MORFA-S
VASTA
SAHKA
LEKSA
NUMRA
Vástit olles cealkagiin. Muitte ahte báikenamat álget stuora bustávain.
Veahkkegiella
English
SAHKA
Buorre beaivi! Bures boahtin mu geahčái!
Mun lean aiddo fárren sisa iežan ođđa orrunsadjái. Mus leat lossa viessogálvvut dáppe feaskáris. Gillešit go veahkehit mu?
Jua, mun gillen veahkehit du.
Mus lea TV dás. Guđe lanjas TV lea du orrunsajis?
Mu TV lea garáša.
Vástádus
Your answer should contain a locative.
HELP
basadanlatnja,
lokta,
oađđenlatnja,
stohpu, feaskkir,
gievkkan, hivsset
Copyright 2012 Romssa universitehta
Contact oahpa@hum.uit.no
Resurssat
Bagadus

# Schematic overview of the treatment of the free input



# Analysis: Searching for the missing illative

```
"<Guđe>"
   "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>"
   "gievkkan" 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

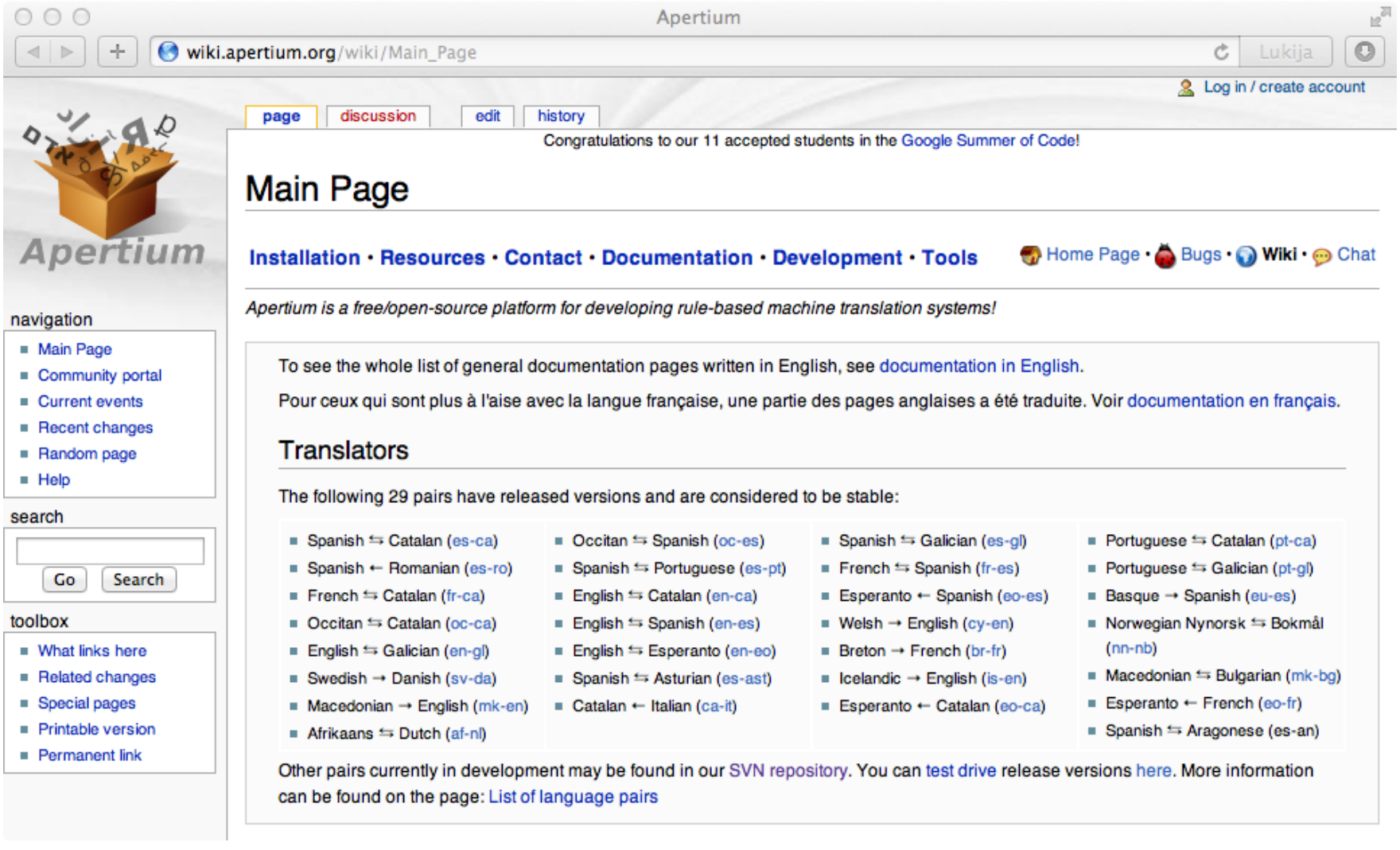
# Work ahead: Modularising the infrastructure

- 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



# 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?