Developing Intelligent Online Web Exercises for Russian

Markus Dickinson, Joshua Herring, and Chris Riley
Dept. of Linguistics, Indiana University

CALICO ’09
Tempe, AZ; March 14, 2009

These slides have been (only) slightly modified since the talk
Introduction & Motivation

Intelligent computer-aided language learning (ICALL) systems are ideal for language pedagogy & research

▶ Provide additional practice outside classroom

▶ Aid awareness of language forms & rules (see Amaral and Meurers 2006)

ICALL provides opportunities for second language acquisition (SLA) research

▶ example: track acquisition patterns via learner models Holland et al. (1993)

▶ cf. Thursday’s panel discussion, “Success and Challenges of ICALL for Learning, Teaching, & Research”

An ICALL system provides opportunities for research into more robust ICALL methods
Introduction & Motivation

However, active development in ICALL is relatively young, and the state of the art suffers from:

1. **Scarcity** - small number of systems, lack of available natural language processing (NLP) tools
2. **Expense** - in-house tool development and intense testing cycle
3. **Overspecialization** - systems are designed around particular languages & contexts
However, active development in ICALL is relatively young, and the state of the art suffers from:

1. **Scarcity** - small number of systems, lack of available natural language processing (NLP) tools
   ▶ we are only aware of ICALL systems covering
     ▶ German (Heift and Nicholson 2001)
     ▶ Japanese (Nagata 1995)
     ▶ Portuguese (Amaral and Meurers 2007a)
     ▶ soon: Russian (this project)
   ▶ few NLP tools available for many languages
Introduction & Motivation

However, active development in ICALL is relatively young, and the state of the art suffers from:

2. **Expense** - in-house tool development and intense testing cycle
   - end up reinventing the wheel
   - proprietary systems can be expensive
     - difficult to integrate once purchased
   - ability to modify components easily during testing is crucial
However, active development in ICALL is relatively young, and the state of the art suffers from:

3. **Overspecialization** - systems are designed around particular languages & contexts
   - linguistic idiosyncracies get baked in to the design, even when abstraction is possible
   - barrier to component reusability
This presentation introduces the *Boltun* project - an ICALL system for learners of Russian under development with these concerns in mind

1. Scarcity
2. Expense
3. Overspecialization
Overview

This presentation introduces the Boltun project - an ICALL system for learners of Russian under development with these concerns in mind.

1. Scarcity
   - Russian system adds a new language
     - Russian brings new development challenges (esp. regarding morphological processing)
   - develops in-house tools with an eye to reusability
     - Previous Russian systems do not seem to be widely available or in use (e.g., Felshin 1995; Loritz 1992)

2. Expense

3. Overspecialization
Overview

This presentation introduces the *Boltun* project - an ICALL system for learners of Russian under development with these concerns in mind:

1. **Scarcity**
2. **Expense**
   - few resources are available for Russian
   - forces adaptation of existing web & NLP tools
     - some tools designed with other languages in mind
   - GOAL: release source code in the future, to reduce others’ expenses
3. **Overspecialization**
Overview

This presentation introduces the *Boltun* project - an ICALL system for learners of Russian under development with these concerns in mind:

1. **Scarcity**
2. **Expense**
3. **Overspecialization**
   - Russian is a new type of language.
   - Open avenues to more general development practices, promoting reusability of components of ICALL systems
     - Modeled on an existing, modular system (TAGARELA (Amaral 2007))
     - Adapt to needs of Russian, making it easier to extend system to new language types in the future
Introduce the Boltun ("chatterbox") web-based exercise system for learners of Russian
  - starting point: the TAGARELA project for Portuguese
    - design considerations carried over from TAGARELA project promote reusability
    - introduce web framework & modular design

Discuss the challenges in adding intelligent processing to a system of Russian
  - What kind of processing do our exercises require?
    - having the system is a prerequisite for answering this question
  - How can appropriate resources be obtained quickly?
    - may help others seeking to work with other languages
Welcome to Boltun version 0.0.1!
You can select lessons from the side menu.

Boltun is currently under development, but you can help us improve! If you encounter difficulties using this application or would like to make suggestions for improvement, please select "Leave Feedback" from the sidebar to leave a comment and we will address your concerns promptly. Thank you!
Developing Intelligent Online Web Exercises for Russian

Introduction

The Boltun Project

Ideas and Motivation

System Architecture Implementation

The ‘I’ in ICALL

Challenges for Russian

Defining a lexicon

Obtaining resources

Summary & Outlook

References

Introducing Boltun

Background:

- **Public Resource** - developed under a FIPSE grant for the Slavic Department at Indiana University
- **Educational Tool** - in use by beginning students
Introducing Boltun

Background:

- **Public Resource** - developed under a FIPSE grant for the Slavic Department at Indiana University
- **Educational Tool** - in use by beginning students
  - currently in use for “survival Russian” course
  - to be expanded to include advanced learners in a healthcare exchange program
Introducing Boltun

Goals:

▶ Broad Coverage
  ▶ current students are at different levels
  ▶ future use will include all levels

▶ Domain-neutral
  ▶ general use (beginning students)
  ▶ healthcare policy (advanced students)
Introducing Boltun

Goals:

▶ Broad Coverage
  ▶ current students are at different levels
  ▶ future use will include all levels

▶ Domain-neutral
  ▶ general use (beginning students)
  ▶ healthcare policy (advanced students)
Introducing Boltun

Goals:

▶ Broad Coverage
  ▶ current students are at different levels
  ▶ future use will include all levels
▶ Domain-neutral
  ▶ general use (beginning students)
  ▶ healthcare policy (advanced students)
Introducing Boltun

Goals:

▶ Broad Coverage
  ▪ current students are at different levels
  ▪ future use will include all levels

▶ Domain-neutral
  ▪ general use (beginning students)
  ▪ healthcare policy (advanced students)
Boltun was inspired by and adopts important concepts from the TAGARELA system for Portuguese (Amaral 2007)

- **Web-Based Format** - hosted at Indiana University and is freely available to anyone with an IU login.
- **Modular Architecture** - adopts TAGARELA’s goal of strict separation between activity, error and student models
Boltun was inspired by and adopts important concepts from the TAGARELA system for Portuguese (Amaral 2007)

- **Web-Based Format** - hosted at Indiana University and is freely available to anyone with an IU login.
  - flexible development cycle
  - no distribution problem

- **Modular Architecture** - adopts TAGARELA’s goal of strict separation between activity, error and student models

**Relationship with TAGARELA**

Boltun was inspired by and adopts important concepts from the TAGARELA system for Portuguese (Amaral 2007)

- **Web-Based Format** - hosted at Indiana University and is freely available to anyone with an IU login.
  - flexible development cycle
  - no distribution problem

- **Modular Architecture** - adopts TAGARELA’s goal of strict separation between activity, error and student models
Developing Intelligent Online Web Exercises for Russian

Introduction

The Boltun Project

Ideas and Motivation

System Architecture

Implementation

The 'I' in ICALL

Challenges for Russian

Defining a lexicon

Obtaining resources

Summary & Outlook

References

Relationship with TAGARELA

Boltun was inspired by and adopts important concepts from the TAGARELA system for Portuguese (Amaral 2007)

- **Web-Based Format** - hosted at Indiana University and is freely available to anyone with an IU login.

- **Modular Architecture** - adopts TAGARELA’s goal of strict separation between activity, error and student models
  
  - flexibility: possible to work on one aspect to the exclusion of others
  
  - scalability: minimal dependencies

Web-Based Format

Modular Architecture
Developing Intelligent Online Web Exercises for Russian

Introduction

The Boltun Project

Ideas and Motivation

System Architecture

Implementation

The 'I' in ICALL

Challenges for Russian

Defining a lexicon

Obtaining resources

Summary & Outlook

References

The Architecture of TAGARELA

Amaral and Meurers (2007b)
The Three Conceptual CALL Modules

- Activity/Instructor Model
- Error Model
- Student Model
The Three Conceptual CALL Modules

- Activity/Instructor Model
  - exercise taxonomy
  - feedback requirements
- Error Model
- Student Model
The Three Conceptual CALL Modules

- Activity/Instructor Model
- Error Model
  - error taxonomy
  - feedback requirements by expected error type
- Student Model
The Three Conceptual CALL Modules

- Activity/Instructor Model
- Error Model
- **Student Model** - information about the individual student’s
  - linguistic strengths and weaknesses
  - level of progress
  - expected knowledge (based on past performance and repertoire of completed lessons)
The Activity Model

Motivations:

▶ Promotes a reusable codebase
▶ Establishes a familiar workzone for the learner
  ▶ background colors & other design features can be changed for all activities of one type at once
  ▶ familiarity with operation procedures
▶ Allows for easy specification of idiosyncracies of the task
  ▶ good for interaction with error module
Activity Model

Alphabet example
Activity Model

Ordering example
Developing Intelligent Online Web Exercises for Russian

Activity Model

Matching example

Lesson 1
Lesson 2
Lesson 3
Lesson 4

Leave Feedback

Sign Out

Drag words from the list onto the matching picture. Use the arrow keys to cycle through all the pictures. There is no need to answer in any particular order, so you can match the ones you are sure of and go back later to any that aren’t so clear.

сестра́ брат оте́ц ба́бушка дёдушка мать

dёдушка

References
Activity Model

Fill-in-the-blank example

Lesson 1
Lesson 2
Lesson 3
Vocabulary
Matching
1 2 3
Family
Scramble
1 2 3
Family
Grammar
Prepositions
Case
Lesson 4
Leave
Feedback
Sign Out

Задание: Прослушайте описание семьи Сергея. Заполните пропуски словами из рамки.

Здравствуйте! Меня зовут Сергей. Я русский студент. У меня типичная _________.

Check Answer Replay
Use Phonetic Keymap

Introduction
The Boltun Project
Ideas and Motivation
System Architecture
Implementation
The 'I' in ICALL
Challenges for Russian
Defining a lexicon
Obtaining resources
Summary & Outlook
References
The Error Model

Motivations:

- Promotes a reusable codebase
- Frees activity design from need to consider feedback implementation
- Allows for tailoring and constraint of feedback:
  - needs of the lesson
  - expected performance of the learner
  - context of lesson grouping and past feedback
Error Model

Error in matching

Developing Intelligent Online Web Exercises for Russian

Introduction

The Boltun Project

Ideas and Motivation

System Architecture

Implementation

The 'I' in ICALL

Challenges for Russian

Defining a lexicon Obtaining resources

Summary & Outlook

References
Developing Intelligent Online Web Exercises for Russian

Introduction

The Boltun Project

Ideas and Motivation

System Architecture

Implementation

The 'I' in ICALL

Challenges for Russian

Defining a lexicon

Obtaining resources

Summary & Outlook

References

Error Model

Correct matching

Lesson 1
Lesson 2
Lesson 3
Lesson 4

Leave Feedback

Sign Out

Drag words from the list onto the matching picture. Use the arrow keys to cycle through all the pictures. There is no need to answer in any particular order, so you can match the ones you are sure of and go back later to any that aren’t so clear.

направо
близко
dалеко
переулок
площадь
кинотеатр
прямо
конечно

ресторан

Правильно(Correct)!
Error Model

Error in fill-in-the-blank

Задание: Прослушайте описание семьи Сергея. Заполните пропуски словами из рамки.

Здравствуйте! Меня зовут Сергей. Я русский студент. У меня типичная семья [семья].
Error Model
Correct fill-in-the-blank

Lesson 1
Lesson 2
Lesson 3
Lesson 4
Leave Feedback
Sign Out

Задание: Прослушайте описание семьи Сергея. Заполните пропуски словами из рамки.

Здравствуйте! Меня зовут Сергей. Я русский студент. У меня типичная семья.
The Student Model

Motivations:

- Addresses concerns that computers are too impersonal to be effective in a learning environment
- Enables learner modeling
  - proficiency
  - past performance
  - native language
- Avenue for SLA research
  - database of learner responses
  - ability to test instruction strategies

⇒ Most of the potential benefit of adding intelligence to CALL is here
The Student Model

Motivations:

- Addresses concerns that computers are too impersonal to be effective in a learning environment
- Enables learner modeling
  - proficiency
  - past performance
  - native language
- Avenue for SLA research
  - database of learner responses
  - ability to test instruction strategies

⇒ Most of the potential benefit of adding intelligence to CALL is here
Web frameworks support development of dynamic websites (for example by providing libraries for session management)

We use the Pylons web framework (http://www.pylonshq.com/)

1. freely available
2. easily packageable
3. widely used
4. highly modular
5. open source
6. saves on web coding
Implementation details
Web development framework

Web frameworks support development of dynamic websites (for example by providing libraries for session management)

We use the Pylons web framework (http://www.pylonshq.com/)

1. freely available
2. easily packageable
3. widely used
4. highly modular
5. open source
6. saves on web coding
Implementation details
User Interface

Libraries provide pre-written user interface controls to allow designers to focus on more interesting aspects of dynamic websites.

We use various effects from JQuery (http://jquery.com/):

- well-maintained Javascript library
- saves on interface scripting
  - in particular on cross-browser scripting

Drawback:

- browser-dependent idiosyncracies
Implementation details

User Interface

Libraries provide pre-written user interface controls to allow designers to focus on more interesting aspects of dynamic websites.

We use various effects from JQuery (http://jquery.com/)

- well-maintained Javascript library
- saves on interface scripting
  - in particular on cross-browser scripting

Drawback:
- browser-dependent idiosyncracies
Implementation details

Intelligent backend

All of this must interface with intelligent processing in a clean way

Python and C++ Backend
  ▶ Python
  ▶ C++
Implementation details

Intelligent backend

All of this must interface with intelligent processing in a clean way

Python and C++ Backend

- **Python**
  - highly portable
  - easy interaction with Pylons (also a Python program)
  - handles student modeling, basic feedback, page rendering

- **C++**
Implementation details

Intelligent backend

All of this must interface with intelligent processing in a clean way.

Python and C++ Backend

- **Python**
- **C++**
  - fast, efficient: good for more intense NLP tasks
  - handles morphological processing, parsing and complex error feedback

**Drawbacks:**
- difficult - programmers are expensive
- not very portable
Implementation details

Intelligent backend

All of this must interface with intelligent processing in a clean way

Python and C++ Backend

- **Python** student model, activity model
- **C++** error model
Current State

- Functioning CALL project
  - used in introductory courses
  - developed by three part-time developers
- Early stages of development for ICALL
  - working morphological analyzer
  - leveraging this to build a more accurate and general tool
  - beginning to do some syntactic analysis
Current State

- Functioning CALL project
  - used in introductory courses
  - developed by three part-time developers
- Early stages of development for ICALL
  - working morphological analyzer
  - leveraging this to build a more accurate and general tool
  - beginning to do some syntactic analysis

⇒ The activity of building the system itself provides opportunities for research into effective ICALL systems
A Motivating Activity Type

- Compose a sentence from the following words:
  вы / машина / есть
  *vy* ("you") / *mashina* ("car") / *est’* ("to have, to be")
- Target answers:
  - У вас есть машина
    *U vas est’ mashina* ("You have a car")
  - У вас машина
    *U vas mashina* ("You have a car")
A Motivating Activity Type

ISSUES:

▶ word order
▶ morphological analysis
▶ reasonable feedback:
  ▶ prioritize response
  ▶ avoid marking grammatical deviations as incorrect
  ▶ provide helpful feedback (more than a spelling exercise)
    ▶ be aware of dependencies between errors (example: switching gender)
Challenges for Russian Morphologically Rich

- Russian allows for relatively free word order
  - grammatical relations are largely encoded in the morphology
- “Syntax-morphology loop”
  - need accurate morphological analysis to parse a sentence
  - → provision of morphological feedback depends on an accurate parse
    - EFFECT: There are many failure points for a computer system

**THE PROBLEM:** How do we provide useful feedback without sacrificing accuracy?
There seem to be few freely-available NLP tools for Russian (see discussion in Sharoff et al. 2008).

Available tools are not tailored to the task of analyzing learner input:
- most tools rely on grammatical input
- learner input is noisy

**THE PROBLEM:** How do we adapt existing tools to the task of analyzing learner language?
So we have two problems to solve in order to be able to provide intelligent feedback:

1. How to provide useful analysis in such a noisy environment?
2. How to get the tools we need for the task?

We start with an examination of acquiring morphological analysis tools
Requirements

- **Flexibility** - should return a set of analyses,
- **Generality** - should not confine itself to grammatical strings, but should be sensitive to learner errors
Requirements

- **Flexibility** - should return a set of analyses, which can be filtered/constrained by:
  - activity and student models
  - other components of the error model (example: syntax constrains morphology)

- **Generality** - should not confine itself to grammatical strings, but should be sensitive to learner errors
Developing Intelligent Online Web Exercises for Russian

Introduction
The Boltun Project
Ideas and Motivation
System Architecture
Implementation
The 'I' in ICALL
Challenges for Russian
Defining a lexicon
Obtaining resources
Summary & Outlook
References

Requirements

- **Flexibility** - should return a set of analyses,
- **Generality** - should not confine itself to grammatical strings, but should be sensitive to learner errors
  - learners innovate
  - learners make paradigm mistakes

For more information, see Dickinson and Herring (2008, 2009)
A Proposed Solution

General morphological analysis can be brought to bear here (Roark and Sproat 2007)

- **Finite-State Morphological Analyzer** - Represents words as a chain of connected letters
  - every wordform in lexicon
  - correct word has a “path” through the lexicon
  - paths are annotated with grammatical information
  - crucially: feedback is based on *alterations* to the path
A Proposed Solution

- **Finite-State Morphological Analyzer** - Represents words as a chain of connected letters
  - every wordform in lexicon
  - correct word has a “path” through the lexicon
  - paths are annotated with grammatical information
  - crucially: feedback is based on *alterations* to the path
A Proposed Solution

Finite-State Morphological Analyzer - Represents words as a chain of connected letters

- every wordform in lexicon
- correct word has a “path” through the lexicon
- paths are annotated with grammatical information
- crucially: feedback is based on *alterations* to the path
A Proposed Solution

Finite-State Morphological Analyzer - Represents words as a chain of connected letters

- every wordform in lexicon
- correct word has a “path” through the lexicon
- paths are annotated with grammatical information
- crucially: feedback is based on *alterations* to the path
A Proposed Solution

Finite-State Morphological Analyzer - Represents words as a chain of connected letters

- every wordform in lexicon
- correct word has a “path” through the lexicon
- paths are annotated with grammatical information
- crucially: feedback is based on *alterations* to the path
A Proposed Solution

- **Finite-State Morphological Analyzer** - Represents words as a chain of connected letters
  - every wordform in lexicon
  - correct word has a “path” through the lexicon
  - paths are annotated with grammatical information
  - crucially: feedback is based on *alterations* to the path
A Proposed Solution

Finite-State Morphological Analyzer - Represents words as a chain of connected letters

- every wordform in lexicon
- correct word has a “path” through the lexicon
- paths are annotated with grammatical information
- crucially: feedback is based on *alterations* to the path
A Proposed Solution

Finite-State Morphological Analyzer - Represents words as a chain of connected letters

- every wordform in lexicon
- correct word has a “path” through the lexicon
- paths are annotated with grammatical information
- crucially: feedback is based on *alterations* to the path
Advantages

ADVANTAGES:

1. robust against errors
2. facility with multiple analyses
3. flexibility
4. broad language coverage
5. well understood algorithms
Overcoming the Resource Constraint

But → building this kind of analyzer requires a lexicon...

- **The Problem**
  - resources are scarce for Russian
  - existing resources are not flexible enough for use with learners

- **The Solution**
  - adapt existing resources to fit our requirements
Overcoming the Resource Constraint

BUT → building this kind of analyzer requires a lexicon... 

- The Problem
  - resources are scarce for Russian
  - existing resources are not flexible enough for use with learners

- The Solution
  - adapt existing resources to fit our requirements
Overcoming the Resource Constraint

BUT → building this kind of analyzer requires a lexicon...

- **The Problem**
  - resources are scarce for Russian
  - existing resources are not flexible enough for use with learners

- **The Solution**
  - adapt existing resources to fit our requirements
Fortunately, it is possible to adapt an available part of speech tagger lexicon for Russian:

- **INPUT**: a list of words with appropriate category tags (3rd singular accusative noun, etc.)
- **OUTPUT**: a set of affixes appropriate to each tag

⇒ System is already more than 80% accurate on analysis
Fortunately, it is possible to adapt an available part of speech tagger lexicon for Russian

- **INPUT**: a list of words with appropriate category tags (3rd singular accusative noun, etc.)
- **OUTPUT**: a set of affixes appropriate to each tag
  
  ⇒ *System is already more than 80% accurate on analysis*
Advantages to our morphological analysis system include:

1. Solves Russian Problem
2. Generalizes well
3. Configurable
4. Inexpensive
Advantages

Advantages to our morphological analysis system include:

1. **Solves Russian Problem**
   - captures shared structure over multiple wordforms
   - allows for guiding “repair” across morpheme boundaries

2. **Generalizes well**

3. **Configurable**

4. **Inexpensive**
Advantages

Advantages to our morphological analysis system include:

1. **Solves Russian Problem**
2. **Generalizes well**
   - compatible with any alphabetic writing system
3. **Configurable**
4. **Inexpensive**
Advantages

Advantages to our morphological analysis system include:

1. **Solves Russian Problem**
2. **Generalizes well**
3. **Configurable**
   - behavior easily adapted to fit activity and student models
4. **Inexpensive**
Advantages

Advantages to our morphological analysis system include:

1. Solves Russian Problem
2. Generalizes well
3. Configurable
4. Inexpensive
   ▶ uses well understood algorithms
   ▶ there are freely-available finite state tools
Summary

This presentation has

1. Introduced the *Boltun* project and explained some of how it operates
   - added a language to the pool of available ICALL systems

2. Conveyed the challenges specific to working with Russian
   - suggested some solutions to the resource acquisition problem for underrepresented languages

3. Suggested a common architecture & analysis framework to promote resource sharing between future projects
Summary

This presentation has

1. Introduced the *Boltun* project and explained some of how it operates
   - added a language to the pool of available ICALL systems

2. Conveyed the challenges specific to working with Russian
   - suggested some solutions to the resource acquisition problem for underrepresented languages

3. Suggested a common architecture & analysis framework to promote resource sharing between future projects
This presentation has

1. Introduced the *Boltun* project and explained some of how it operates
   - added a language to the pool of available ICALL systems
2. Conveyed the challenges specific to working with Russian
   - suggested some solutions to the resource acquisition problem for underrepresented languages
3. Suggested a common architecture & analysis framework to promote resource sharing between future projects
Outlook

1. Focus on making the exercises and system design more communicative & interactive/exploratory
2. Finish testing the morphological analyzer and put it into place with real learner language
3. Begin using a student model for feedback
   - learner information is currently tracked, but is not yet used for provision of feedback
4. Add an interface for instructors to tailor feedback
Acknowledgments

We would like to thank

- Detmar Meurers and Luiz Amaral for providing us with the TAGARELA sourcecode & insights into ICALL systems
- Olena Chernishenko for designing exercises
- Anna Feldman & Jirka Hana for advice on Russian resources
- The Indiana University Computational Linguistics discussion group for comments & feedback

This research was supported by grant P116S070001 through the U.S. Department of Education’s Fund for the Improvement of Postsecondary Education.


Dickinson, Markus and Joshua Herring (2009). Morphological Analysis for Russian Learner Language. Talk given Workshop on Automatic Analysis of Learner Language (AALL’09); March 10, 2009; Tempe, AZ.


Heift, Trude and Devlan Nicholson (2001). Web delivery of adaptive and
Developing Intelligent Online Web Exercises for Russian

Introduction

The Boltun Project

Ideas and Motivation

System Architecture

Implementation

The 'I' in ICALL

Challenges for Russian

Defining a lexicon

Obtaining resources

Summary & Outlook

References


