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

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 (Hefft and Nicholson 2001)
   - Japanese (Nagata 1995)
   - Portuguese (Amaral and Meurers 2007a)
   - soon: Russian (this project)
   - few NLP tools available for many languages

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

3. **Overspecialization** - systems are designed around particular languages & contexts
   - linguistic idiosyncrasies get baked in to the design, even when abstraction is possible

---

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
- 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

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**
   - 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**
   - 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)

---

Outline of talk

- 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

The Boltun Project

Welcome to Boltun version 0.0.1!
You can select lessons from the side menu.

Lesson 1
Lesson 2
Lesson 3
Lesson 4

Leave Feedback
Sign Out
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

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)

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
  - flexibility: possible to work on one aspect to the exclusion of others
  - scalability: minimal dependencies

The Architecture of TAGARELA
Amaral and Meurers (2007b)

The Three Conceptual CALL Modules

Activity/Instructor Model
- exercise taxonomy
- feedback requirements

Error Model
- error taxonomy
- feedback requirements by expected error type

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

You will hear the name of a Russian letter. Using the keyboard alone as a guide, type the letter you hear.

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

Activity Model

Ordering example

Activity Model

Matching example

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

Activity Model

Fill-in-the-blank example

Error Model

Error in matching

Error Model

Correct matching
### Error Model

**Error in fill-in-the-blank**

![Image of fill-in-the-blank exercise]

**Correct fill-in-the-blank**

![Image of correct fill-in-the-blank exercise]

### 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

### 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

### 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
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:
  - vy ("you") / mashina ("car") / est’ ("to have, to be")
- Target answers:
  - У вас есть машина
  
  - У вас машина

**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)

**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
  - learners innovate
  - learners make paradigm mistakes

For more information, see Dickinson and Herring (2008, 2009)
### 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**
   - compatible with any alphabetic writing system
3. **Configurable**
   - behavior easily adapted to fit activity and student models
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

---

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

---

A Proposed Solution

---

POS Tagger Lexicon

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
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.

References

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.