Language and Computers

Language Tutoring Systems

L245
(Based on Dickinson, Brew, & Meurers (2013))

Spring 2017

Some common computer uses

- Computers are widely used in support of **foreign language teaching (FLT)**. For example, they
  - provide access to foreign language newspapers, radio, and TV programs through the internet
  - connect language learners with native speakers through email/chat
  - support multimedia presentations providing an audio-visual foreign language context
  - enable the learner to search for real-life examples in electronic corpora
- Essentially, such computer usage helps language learners experience a foreign language and culture in a more direct, real-life fashion.

Language Tutoring Systems

Overarching question: How computers can help provide foreign language learners with experiences that are:

- richer,
- more personalized, and
- more effective?

First Language Acquisition

Second language learning differs in many ways from **first language acquisition**:

- Researchers disagree on how much of language learning ability
  - is **innate**, i.e., a biological endowment
  - **emerges** from experience, i.e., a rich social and physical environment.
- But, crucially, children become **native speakers** without explicit instruction
  - They typically follow the same **stages of acquisition** (babbling, word learning, simple utterances, etc.)

Second Language Acquisition

Awareness of language forms

Adults do not automatically acquire a second language

- Even after living in a foreign country for a long time, listening to & talking in a foreign language there
- Research since the 90s has shown that **awareness** of language forms and rules is important for an adult learner to successfully acquire a foreign language.
  - e.g., the use of the **articles** *the* and *a* in English is difficult to learn
    - especially for those whose native language does not make use of articles (Chinese, Russian, etc.)
    - requires awareness of: **mass nouns** (e.g., *rice*) & **generics** (e.g., *milk* in *I like to drink milk*)

Language Tutoring Systems (LTSs) can provide an opportunity to enhance awareness of a language's rules

Needs of second language learners

- The time a student can spend with an instructor/tutor typically is very limited
  - Work on form and grammar is often de-emphasized and confined to homework
  - The time with the instructor is used for purely communicative activities
- Learners have relatively few opportunities to gain awareness of forms & rules and receive individual feedback

An opportunity for CALL

- The situation seems like an excellent opportunity for developing Computer-Aided Language Learning (CALL) tools to
  - provide individual feedback on learner errors and
  - foster learner awareness of relevant language forms and categories.
- But for existing CALL systems which offer exercises:
  - they typically are limited to uncontextualized multiple choice, point-and-click, or simple form filling
  - feedback usually is limited to yes/no or letter-by-letter matching of the string with a pre-stored answer
    - An example for letter-by-letter feedback on the "Spanish Grammar Exercises" site (B. K. Nelson)

Basic uses of computers for CALL

Lots of general possiblities for using a computer to learn:

- multimedia presentations
- online dictionaries with fast access
- extensive databases of information
- digital audio files
- digital videos of people speaking in L2

And then some more specific cases where natural language processing could help:

- interactive games & puzzles
- exercises for students to complete

CALL systems
Multiple choice

Computers can explicitly store knowledge about words or grammar necessary to complete a specific exercise

1. Fred lives _____ Mill Street, doesn't he?
- ○ in
- ○ on
- ○ at

2. My father was born _____ Christmas Eve.
- ○ at
- ○ on
- ○ in

3. Come here _____ once! I need your help right now!
- ○ at
- ○ on
- ○ in

(Source: http://www.eslcafe.com/quiz/prep3.html)

Multiple choice exercises work well for practicing or testing specific choices of forms or meanings

- Include so-called **distractors** as incorrect choices

CALL systems
Fill-in-the-blank

Other possible exercises include:

- Pull-down menus listing the choices
- **Fill-in-the-blank (FIB)** texts: a word in a sentence is erased & the learner must type in the missing word
  - Also referred to as **cloze** exercises
  - Often include a **fallback case** to respond to any unexpected input
    - i.e., **canned text responses**

Putting questions on the web or another computer-based platform makes it possible to provide immediate feedback

- How to provide feedback for more open-ended exercise types?
  - Simple answer: write out all possibilities

Early CALL systems

Frame-based systems "match student answers with a set of correct and incorrect answers stored in a frame"

- These systems differ in their strategies for selecting questions, but they rely on preset questions & answers
- In principle, could be used with NLP techniques

Many also feature a dynamic **sequencing of instruction**

Problems with frame-based systems

Frame-based systems are fairly simple and generally do not involve much linguistic knowledge

- There is no deep understanding of question domain
- They generally only match answers with questions, but language use is more varied
- There is not much tailoring to particular student needs

Language awareness
Making generalizations

What happens when teachers must specify all options for answering an exercise?

(1) Today is November 5. What date is tomorrow?
Tomorrow is ________________________.

Possible correct answers (among others):

- 06. 11.
- Nov., the 6th
- the sixth
- November, the sixth
- 11/06
- 6. Nov.

- Many different ways to misspell any of these options
- Many different possible incorrect answers

⇒ We need linguistic generalizations, in this case:

- **Named entity recognition** to identify special expressions, e.g., dates, addresses, names

Language awareness
Semantic generalizations

More broadly: refer to classes instead of individual strings

- Consider fill-in-the-blank exercise modeled on a German exercise in Trude Heift's E-Tutor system:

(2) John works in New York City, but his family lives in Boston. On the weekend, he drives home. Fortunately, John has a new ________________________.

Different options for correctly filling in this blank:

- **Synonyms**: words which mean the same thing, at least in certain contexts: e.g., *car* & *automobile*
- Other **lexical semantic relations** between words:
  - **Hyponymy**; using a more specific term (**hyponym**), e.g., *pick-up*, *SUV*, or *hybrid car*
  - The more general term *car* is the **hypernym**

Language awareness
Morphological generalizations

Additionally, a single word in a language can show up in different forms.

- e.g., **citation form** or **lemma** of *bring* is*to bring*
  - Also realized as *bringing*, *brought*, *bring*, or *brings*
  - The different word forms and their function are investigated in **morphology**
- Other languages feature richer inventories of forms
  - e.g., 6 forms for one of the verbs meaning *to be* in Spanish: *soy*, *eres*, *es*, *somos*, *sois*, *son*
  - Plus over a dozen other tenses and moods

We would need to spell out the many different forms for each exercise in a CALL system

Language awareness
Syntactic generalizations

Consider exercises where learner can enter multiple words

- The various word order possibilities result in additional, systematic variation
- **Syntax** identifies different word order possibilities & the forms words have to appear in

(3) John, the radio is much too loud. Please ________________________!

(4) a. turn down the radio.
b. turn the radio down.

Many non-English languages allow freer word order

- Capturing all possible word orders is infeasible

Linguistic generalizations can compactly specify the expected correct or incorrect answers

Intelligent CALL (ICALL)

Intelligent CALL (ICALL) focuses on using linguistics and natural language processing to make CALL better.

- ICALL can also involve integrating authentic text into exercises, usually for more advanced learners
- ICALL involves providing linguistic analysis to handle real learner input

So, what types of linguistic analysis do we need to do?

Adding linguistic analysis
Tokenization

Starting point: find the words (or **tokens**)

- A text is simply a very long list of letters
- **Tokenization** (or **word segmentation**): task of finding tokens in a text

Why is this challenging?

1. **Covering ambiguity**: two or more characters may be combined to form one word or not
   - Writing systems of many languages do not use spaces between words, e.g., 要害 in Chinese:
     - Option #1: segment as two words of one character each, meaning *will hurt*
     - Option #2: segment it as a single word of two characters, meaning *vitals*
   - Context determines the segmentation

Adding linguistic analysis
Tokenization (2)

2. **Overlapping ambiguity**: a given character may either combine with the previous or with the next word
   - 布什在谈话中指出(ex. from Xiaofei Lu)
   - Meaning changes depending on which word the second to last character 指is part of

* 布什	在	谈话	中指	出
Bush	at	talk	middle-finger	out

布什	在	谈话	中	指出
Bush	at	talk	middle	point-out

Bush pointed out in his talk’

- NB: in Chinese, only the second segmentation option is grammatical

Adding linguistic analysis
Tokenization (3)

Language and Computers | Language Tutoring Systems | What is ICALL? | Second Language Acquisition | An opportunity for CALL | CALL systems | Basic uses of computers | Early CALL systems | Language awareness | ICALL | Linguistic analysis | Parser-Based ICALL | Learner modeling | Authentic Text ICALL

Even for English, spaces are not exact:
- e.g., *inasmuch as*, *insofar as*, *in spite of*

1. **Compound nouns** such as *flu shot*:

(5) a. I got my flu shot yesterday.
    b. I got my salary yesterday.

2. **Contractions**: e.g., *I’m*, *cannot*, or *gonna*
   - They should likely be treated on a par with *I am*, *can not*, and *going to*

Automatic tokenizers typically have long lists of known words & abbreviations, plus (finite-state) rules for subregularities

Adding linguistic analysis
POS tagging

Language and Computers | Language Tutoring Systems | What is ICALL? | Second Language Acquisition | An opportunity for CALL | CALL systems | Basic uses of computers | Early CALL systems | Language awareness | ICALL | Linguistic analysis | Parser-Based ICALL | Learner modeling | Authentic Text ICALL

With tokens identified, we can obtain the general classes of words we want, such as part-of-speech (POS) classes
- e.g., to support **meta-linguistic feedback** messages such as “The sentence you entered is missing a verb.”

Parts of speech are labels for classes of words which behave alike ... in three different ways:

1. **Distribution**: linear order with respect to other tokens, i.e., the slot a word appears in.
   - e.g., for *John gave him ____ ball.*:
     - Slot between *him* & *ball* is a distributional slot of a determiner such as *the* or *a*
   - For automatic POS taggers, distributional information encoded as statistics about POS (*n*-gram) sequences

Adding linguistic analysis
POS tagging (2)

Language and Computers | Language Tutoring Systems | What is ICALL? | Second Language Acquisition | An opportunity for CALL | CALL systems | Basic uses of computers | Early CALL systems | Language awareness | ICALL | Linguistic analysis | Parser-Based ICALL | Learner modeling | Authentic Text ICALL

2. **Lexical stem lookup**
   - Unambiguous part-of-speech (POS): e.g., *claustrophobic* is only an adjective
   - Ambiguous POS: e.g., *can*
     - auxiliary: *The baby can walk.*,
     - full verb: *I can tuna for a living.*
     - a noun: *Hand me that paint can, please.*
   - Words not in the lexicon: a big problem for computers

Adding linguistic analysis
POS tagging (3)

Language and Computers | Language Tutoring Systems | What is ICALL? | Second Language Acquisition | An opportunity for CALL | CALL systems | Basic uses of computers | Early CALL systems | Language awareness | ICALL | Linguistic analysis | Parser-Based ICALL | Learner modeling | Authentic Text ICALL

3. **Morphology**: the form of words
   - Markings (e.g., **suffixes** added to stem endings) encode information only appropriate for particular POS
     - e.g., the *-ed* indicates past tense
   - **Inflectional suffixes**: information such as tense or agreement (e.g., *-s* on third person singular verbs)
   - **Derivational affixes** (e.g., *-er* turns verbs into nouns: *walk – walker*).
     - Automatic POS-taggers use *suffix analysis* as a **fallback step**
     - If a word has not been seen before, **suffix analysis** determines the most likely POS

Adding linguistic analysis
POS tagging (4)

Language and Computers | Language Tutoring Systems | What is ICALL? | Second Language Acquisition | An opportunity for CALL | CALL systems | Basic uses of computers | Early CALL systems | Language awareness | ICALL | Linguistic analysis | Parser-Based ICALL | Learner modeling | Authentic Text ICALL

Complication: dealing with **interlanguage**

Consider these sentences written by Spanish learners of English (from the NOCE corpus):

(6) a. ... to be **choiced** for a job ...
    b. RED helped him **during** he was in the prison.

- *choiced*:
  - distributionally appears in a verbal slot
  - morphologically carries verbal inflection (’*-ed*’)
  - lexically the stem *choice* is a noun (or adjective)
- *during*:
  - morphologically is a preposition
  - distributionally a conjunction

POS tagging for learner language need to be extended to take into account such potentially mismatching evidence

Parser-Based ICALL

Parser-Based ICALL systems generally fall along the following lines:

- System presents the learner with an exercise
- Learner inputs an answer, possibly with errors, i.e., a potentially **ill-formed** sentence
- The parser processes this sentence
  - Identifying where, if at all, it was incorrect
  - Providing information on the nature of the error
- Feedback is then presented to the student

We'll look at two example systems:

- e-Tutor (German Tutor): Heift & Nicholson
- TAGARELA: Amaral & Meurers

Parser-Based ICALL

A note on detecting errors

Parsers, morphological analyzers, etc. are designed to handle well-formed input

How do we adapt technology to find errors?

- Use so-called **mal-rules** = rules which are added to the grammar to handle error cases.
  - e.g., A singular noun and a plural verb are allowed to combine, but it is marked as an error.
  - $S_{error} \rightarrow NP_{sg}\ VP_{pl}$
- Modify the technology: a parser can be reworked to handle ill-formed input.
  - e.g., It will parse *John are big*, but will say that the parse failed and how it failed

e-Tutor (German Tutor)

e-Tutor (Heift & Nicholson 2001) is used at Simon Fraser University to teach German to students; it is:

- general, i.e., allows for any native language (L1)
- able to capture different kinds of errors
  - because in large part the exercises are very constrained

Student input is put through the following modules and stops with feedback when the first error is encountered

1. String match: if the input matches a pre-defined correct answer, we know it's good.
   - Prevents time-consuming analysis for perfect answers
2. Punctuation check: is any punctuation missing?

More on system architecture

3. Spell check: run an off-the-shelf spell checker on the input and get the **lemmas**
   - Idea: eliminate the really basic errors.
   - Problem: sometimes a "misspelled" word is a sign of lack of grammatical competence, e.g. *runned*
4. Example check: are the right words being used?
5. Missing word check: are any words missing?
6. Extra word check: are any words added?
   - These 3 steps (example, missing word, and extra word checks) all are based on the notion that the exercise has *pre-defined* all the acceptable words

More on system architecture (cont.)

7. Word order check: match the user word order with the correct word order
8. Grammar check
   - This is the most complicated part of the process, the one which requires linguistic knowledge (syntax)
   - About 60% of errors make it to this stage.
9. Catch-all: just in case everything else fails

Note:

- Heift's system works so well because the exercises themselves are constrained, as we will see
- The approach is very **modular** = each check is an independent program

e-Tutor

Build a Sentence

Use all the given words (lemmas) and create a grammatical German sentence.

Advanced learner output here: "There is an error in gender with the subject."

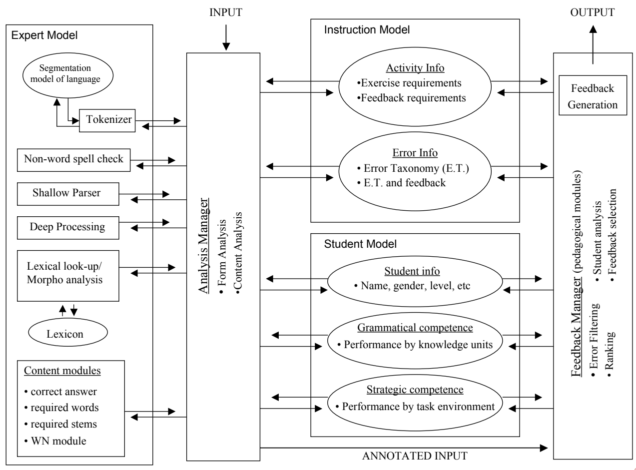
TAGARELA

TAGARELA is a system developed for individualized instruction of Portuguese at Ohio State

- It features standard exercises, as found in foreign language workbooks
- NLP processing is used to detect spelling, morphological, syntactic, and semantic errors
- A student model is kept to track performance and to choose appropriate feedback
  - An instruction model allows the instructor to state what is important

TAGARELA system overview

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL



Demand-driven architecture

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

Different from the e-Tutor, TAGARELA works in a **demand-driven** fashion; the analysis manager:

- receives input from the student
- gathers the necessary information from:
  - instruction model
  - student model
- decides on the best processing strategy
  - which NLP modules to call
  - in which order (as opposed to linearly)
- calls NLP modules to process input, producing an input annotated with linguistic properties
- hands the annotated input to the feedback manager

Sources of information for CALL systems

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

Generally, we have three sources of information by which to analyze a learner production:

1. Language/linguistic properties
   - General information we already discussed about linguistic generalizations
2. Exercise information
   - e.g., what is known about errors for “build a sentence” exercises
3. Information about the learner ...

Modeling the learner

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

Learner modeling includes two types of information:

1. Learner properties which are more or less permanent
   - e.g., gender, native language, learning style
2. Dynamic record of learner performance so far: whether a learner successfully used particular words/structures

Both types of information are relevant for feedback

- e.g., native language (L1) of a learner influences words & constructions used & mistakes made
  - Positive and negative **L1-transfer**
  - Negative transfer: many native speakers of languages such as Chinese or Czech, find *the* & *a* difficult
    - L1s do not include articles of the kind found in English
    - Tutoring system should provide feedback on article misuse for learners with such native languages

Modeling the learner
Obtaining learner information

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

How do we obtain dynamic record of learner performance?

- The system needs to draw **inferences** from the learner’s interaction with the system.
  - Need to abstract to general linguistic properties & classes which a learner answer provides evidence for
    - e.g., whether a learner answer contained a finite verb, provided evidence for subject-verb agreement, etc.
  - After seeing answers with instances of a particular property, we can infer that the learner has mastered it
    - e.g., deprioritize feedback on it in the future
- Models may help **sequence teaching material**
  - e.g., by guiding the learner to additional material on concepts not yet mastered

Authentic Text ICALL

Authentic Text ICALL attempts to connect learners to appropriate naturally-occurring texts

- Allows students to find examples in target language related to their interests
- Allows for more exploration and something akin to "immersion"

Basic uses of computers for CALL

Concordancers

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

One of the simplest ways to show authentic text is via a **concordance**:

- Keyword in context (KWIC)
- Concordancers help learners understand how a given word is used.
  - For example, is the word *data* in English singular or plural?

contract to supply voice and	**data**	communications within the Tunnel in
giving control over how much	**data**	is sent over the network
humanists to fit their special	**data**	to the software , rather
27 mm . But these	**data**	are for fourth-year crabs .

The WERTi System

Visual Enhancement of the Web

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

VIEW is "an ICALL system designed to provide supplementary language learning activities using authentic texts selected by the learner"

- Multi-lingual extension of: WERTi - Working with English Real-Texts: An Intelligent Workbook for English
- Learners select a topic which fits their interests
- Webpages are returned, which learners interact to learn about, e.g., prepositions
  - Learners can choose to see prepositions in color; click on them; or fill in blanks

Crucially, the exercises are **generated** on the fly

- Pre-existing NLP technology (e.g., a POS tagger) is used to spot the relevant categories

The REAP Project

Reader-Specific Lexical Practice for Improved Reading Comprehension

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

In the REAP system:

- Teachers have target vocabulary items
- REAP finds appropriate texts for learners, based on their individual profile
  - Learners get individualized vocabulary practice from authentic web texts

There are several challenges in extracting text for reading

- Each extracted text is analyzed for its "syntactic features, readability, length, and the occurrence of target vocabulary"
- Information retrieval and statistical NLP techniques are used to find appropriate texts

GLOSSER

Language and Computers
Language Tutoring Systems
What is ICALL?
Second Language Acquisition
An opportunity for CALL
CALL systems
Basic uses of computers
Early CALL systems
Language awareness
ICALL
Linguistic analysis
Parser-Based ICALL
Learner modeling
Authentic Text ICALL

GLOSSER facilitates dictionary look-up

- System uses lemmatization and morphological analysis
- Look-up is 100 times faster (Nerbonne 2003)
  - Otherwise very challenging for highly-inflected languages