Language and Computers
Language Tutoring Systems

Based on Dickinson, Brew, & Meurers (2013)
Indiana University
Fall 2013

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 world-wide-web
  - 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.

First Language Acquisition

Stages of First Language Acquisition

Typical stages of first language acquisition:
- Babies play with making sounds and around six months, typical babies begin to babble
  - use sequences of consonants and vowels, e.g., bababa
- Quickly start learning words by first birthday
- Form simple 2-word utterances by the time they turn 2
- Voice & understand complex sentences by 3
- Continue acquiring words & complex language structures over next 9 or 10 years
  - Some structures, e.g., passives, added relatively late

Essentially the same pattern across all languages and cultures (with some individual variation)

A child can be a native speaker of multiple first languages, acquiring each of them without explicit instruction

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)

CALL can provide an opportunity to enhance awareness of a language’s rules
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

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 existing CALL systems which offer exercises typically are limited to uncontextualized multiple choice, point-and-click, or simple form filling, and feedback usually is limited to yes/no or letter-by-letter matching of the string with a pre-stored answer.

Lots of general possibilities 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
  - Digital advantages: easy playback, easy isolation of problematic spots, etc.
- interactive games & puzzles
- exercises for students to complete

The last two examples potentially require sophisticated natural language processing

Take a text and create a concordance = display of words in context.
- Concordancers help learners understand how a given word is used.
  - For example, is the word data in English singular or plural?
  - The contract to supply voice data to the software, rather than giving control over how much humanists to fit their special 27 mm. But these data are for fourth-year crabs.

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

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

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

(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 distracters as incorrect choices
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

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):

- November, the sixth
- Nov., the sixth
- the sixth
- 6. Nov.

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

⇒ We need linguistic generalizations
- named entity recognition = identify special expressions, e.g., dates, addresses, names

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 into 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

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

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 & vehicle
- 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

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
Adding linguistic analysis

Tokenization (2)

2. Overlapping ambiguity: a given character may either combine with the previous or with the next word

- Bush at talk middle-finger out
- Bush at talk middle point-out

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

Adding linguistic analysis

Tokenization (3)

Even for English, spaces are not exact:

- e.g., inasmuch as and insofar as, in spite of should be single tokens

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 pair 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

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 3 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 distributional slot of a determiner such as the or a
   - For automatic POS taggers, distributional information encoded as statistics about POS (n-gram) sequences

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: Pass me that can of beer, please!

   - Words not in the lexicon: a big problem for computers

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

Adding linguistic analysis

Tokenization

To get lemmas (or anything else), we need to 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., 求真在 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 will hurt
   - Context determines the segmentation
Adding linguistic analysis

POS tagging (3)

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 indicated 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

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
  - 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

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:
- Heit'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

TAGARELA

TAGARELA is a system 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

Demand-driven architecture

Different from the e-Tutor, TAGARELA works in a demand-driven fashion; the analysis manager:
- receives input from the student
- gather 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 linearily)
- 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

Generally, we have 3 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

**Learner modeling** includes 2 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 & 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

---

The WERTi System

**REAP**

- Obtaining learner information
  - 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”

---

The REAP Project

**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

**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

---

The WERTi System

**Working with English Real-Texts: An Intelligent Workbook for English**

WERTi is an “intelligent automatic workbook, providing an unlimited number of activities designed to foster awareness of English grammatical forms and functions”

- 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