Language awareness through re-use of NLP technology

Luiz Amaral, Vanessa Metcalf, Detmar Meurers

The Ohio State University

NLP in CALL Workshop
CALICO. May 16, 2006

NLP in ICALL

- The use of NLP in ICALL has primarily centered on diagnosing learner errors and, more recently, testing and assessment.
- Idea: Explore how NLP technology can support other aspects of second language learning.
- Our specific focus: What can NLP contribute to awareness of language forms and rules, an important component of adult second language acquisition.
  - (cf., e.g., Long 1991, 1996; Ellis 1994; Schmidt 1995; Lyster 1998; Lightbown and Spada 1999; Norris and Ortega 2000; Schulz 2002)

Language awareness through re-use of NLP technology
Luiz Amaral, Vanessa Metcalf and Detmar Meurers
Background
NLP in ICALL
Pedagogical grounding
Proposal
Modeling FLT practice
Progression in our approach
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Realizing proposal
1. Obtaining the text base
2. Selecting texts from base
3. Creating ex. progression
Some challenges
Related approaches
The MIRTO project
The VISL project
Generating cloze tests
Cognate exercises, FL reading support, FL text retrieval
Summary
References

Pedagogical grounding of our research

Awareness

Awareness (Schmidt 1995):

- Noticing
  - "conscious registration of an event"
  - low level of awareness
  - implicit learning

  E.g.: noticing that in some occasions speakers of Spanish omit the subject pronoun
- Understanding
  - "recognition of a general principle, rule or pattern"
  - higher level of awareness
  - explicit learning
  - it can be internally generated or externally provided

  E.g. understanding that Spanish is a pro-drop language

Language awareness through re-use of NLP technology
Luiz Amaral, Vanessa Metcalf and Detmar Meurers
Background
NLP in ICALL
Pedagogical grounding
Proposal
Modeling FLT practice
Progression in our approach
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Realizing proposal
1. Obtaining the text base
2. Selecting texts from base
3. Creating ex. progression
Some challenges
Related approaches
The MIRTO project
The VISL project
Generating cloze tests
Cognate exercises, FL reading support, FL text retrieval
Summary
References

Pedagogical grounding of our research

The role of awareness

- Research on awareness shows:
  - There is no learning without noticing.
  - Awareness without input is not sufficient.
  - "Learning takes place within the learner's mind and cannot be completely engineered by teachers or syllabus designers."
  - One can only provide opportunities for developing learner awareness.

⇒ Consequences:

- Learners have to be exposed to linguistic features to acquire them.
- Learners have to notice those features.
- Tools presenting such linguistic features in a contextualized way, allowing for student interaction, can be helpful.

Language awareness through re-use of NLP technology
Luiz Amaral, Vanessa Metcalf and Detmar Meurers
Background
NLP in ICALL
Pedagogical grounding
Proposal
Modeling FLT practice
Progression in our approach
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Realizing proposal
1. Obtaining the text base
2. Selecting texts from base
3. Creating ex. progression
Some challenges
Related approaches
The MIRTO project
The VISL project
Generating cloze tests
Cognate exercises, FL reading support, FL text retrieval
Summary
References

Pedagogical grounding of our research

Linguistic information and how it is conveyed

- A wide range of linguistic features can be relevant for awareness, incl. morphological, syntactic, semantic, and pragmatic information (cf. Schmidt 1995, p. 30).
- Linguistic information can be conveyed to the learner
  - using explicit linguistic terminology/representations, e.g.:
    - parts of speech
    - verbal tense, mood and aspect
    - sentence classification
    - syntactic analyses (shown as trees or sentence diagrams)
  - using implicit presentation, e.g.
    - coloring, underlining, moving, etc
    - pointing to correct or incorrect uses

  ⇒ Awareness activities can include both implicit and explicit presentation of linguistic features.

Modeling FLT practice

- A common pedagogical practice in FLT moves from target language presentation, to practice, on to production.
- Proposal: Create sequences of linguistic awareness activities following the initial stages of such a progression:
  I. Receptive presentation
  II. Productive presentation
  III. Controlled practice
- What makes this idea interesting?
  - NLP technology can identify certain relevant linguistic categories and forms in real-life texts.
  - The contents of these texts can be selected by the learners based on their interests.
  - The sentences turned into exercises can remain fully contextualized as part of the text selected by learner.
  - Automatic grading of the created exercises can be feasible since the original text is known.

The progression in our approach

Using texts chosen on the basis of learner interests and automated filtering, we provide a progression of activities:

Step 1. Receptive presentation

Ex. The system colors examples of targeted items.

Step 2. Productive presentation

Ex. The learner is asked to find and mouse-click all tokens of the targeted category. The system shows correct picks in green, incorrect ones in red.

Step 3. Controlled practice

Ex. The learner is asked to

- reorder words/phrases given (scrambled) list
- complete fill-in-the-blank (FIB) slots
  - created for tokens of targeted category
  - given some information, where needed (e.g., stems)

Example for an activity progression

1. Pronouns

Step 1. Receptive presentation

Ex. System colors different pronoun types.

(1) *Someone told me that he accidentally hit himself in the face with his car keys.*

Step 2. Productive presentation

Ex. Click on examples of a particular type of pronoun.

Step 3. Controlled practice

Ex. Fill in all pronouns in a text.

Ex. Find and correct incorrect pronoun choices in text.
E.g.: *That's him car.* → *That's his car.*

Example for an activity progression

2. Passive

Step 1. Receptive presentation

Ex. System colors passive verb forms.

(2) *Her purse was taken while she wasn't looking.*

Step 2. Productive presentation

Ex. Click on passive sentences

Step 3. Controlled practice

Ex. Given the main verb stem, fill in the passive verb string (i.e., the correct form of *be* and the past participle form of the main verb).

Ex. Given an active sentence, transform the sentence to a passive using a combination of click and drag, and FIB.

Example for an activity progression

3. Adverb placement

Step 1. Receptive presentation

Ex. System colors verbs and verb-modifying adverbs.

(3) *The house had already been damaged.*

Step 2. Productive presentation

Ex. Click on adverbs in a particular position:

- at the beginning of a sentence
- between a main verb and a prepositional phrase
- before an auxiliary verb

Step 3. Controlled practice

Ex. Given constituent chunks and an adverb, with instructions on where this adverb should go, put the sentence together.

Example for an activity progression

4. Tense and Aspect

Step 1. Receptive presentation

Ex. System colors examples of different aspectual meanings together with relevant contextual cues.

(4) a. *We are going to New York tomorrow.*

b. *We usually go to the grocery store on Fridays.*

Note: While the effect is semantic, the cues are lexical.

Step 2. Productive presentation

Ex. Click on sentences expressing a particular kind of meaning with the targeted verb forms, e.g., expressing future plans using present tense.

Step 3. Controlled practice

Ex. Given a main verb stem, provide the appropriate verb string using cues from context.

What is involved in realizing such an approach?

- Our work on this project so far has primarily focused on the pedagogical motivation and exercise setup.
- In the following, we sketch the components needed for its realization:
  1. Obtaining the text base
  2. Selecting appropriate texts from text base
  3. Identifying the targets in the selected texts and creating
     - receptive and productive presentations, and
     - controlled practice exercises using the texts.

Realizing the proposal

1. Obtaining the text base

- Text source: web
  - can be restricted to specific domains (e.g, `.edu`) or news sites (e.g., NYT, BBC) for high(er) quality texts
  - alternative: specific corpora
- Search for content of interest as specified by user
  - What? any string
    - e.g.: horseback-riding in Ireland
  - How? search engine API
    - REST developer interface of Yahoo! supports straightforward submission of web queries (http://developer.yahoo.com/search/web/V1/webSearch.html)
  - Result: a small set of web documents, the text base.

Realizing the proposal

2. Selecting text from text base

- Select text from text base using:
  - instructor model, e.g.:
    - target language
    - appropriate length
    - categories, forms, and constructions to be emphasized
  - learner model, e.g.:
    - level of learner
    - mastered vocabulary, constructions, sentence complexity
    - problem cases
- This requires the text base to be annotated with efficient and robust NLP tools performing
  - tokenization
  - lemmatization
  - part-of-speech tagging, morphological analysis
  - chunking, statistical parsing

Realizing the proposal

3. Creating activity sequence

- The receptive and productive presentation activities result from regular expression matching of target and contextual items in annotated texts from text base:
  - tokens → tokenized learner input
  - lexical categories → POS tagging
  - morphological properties → morphological analysis
  - phrasal categories → shallow parsing
- The creation of certain controlled practice activities requires additional information, e.g.:
  - Providing the user with a verb lemma as cue for a FIB activity requires lemmatization.
- The nature of the activity determines the complexity of the regular expressions required:
  - Pronoun activity: single instances of a lexical category
  - Tense and aspect: sequences of auxiliaries, inflected forms, and specific lexical items (contextual cues)

Realizing the proposal

Some challenges

- Annotation errors:
  - Statistical NLP tools are efficient and robust
  - Such tools make errors, e.g., 3–5% for POS tagging.
  - What impact do such errors have for the envisaged use?
  - It is known where errors are likely to arise (cf., e.g., Dickinson and Meurers 2003; Dickinson 2005), so one can avoid basing activities on likely error locations.
- The complexity of real life:
  - Real-life texts from the web often have
    - complex structure
    - mark-up and integrated multimedia
  - It's a challenge to preserve that structure and mark-up during linguistic annotation of the text base.
  - Receptive and productive presentation can be added modularly to an existing document (mark-up/javascript); inserting forms for controlled practice can be challenging.

Related approaches

The MIRTO project (Antoniadis et al. 2004)

- Similarities
  - Emphasizes pedagogical practice and integration
  - Automatic exercise generation:
    - Plans to support “gap-filling” and “lexical spotting” exercises in combination with a corpus database.
- Differences
  - Aims at creating a general toolbox architecture supporting instructor-determined activity design.
  - General toolbox = no explicit mention of language awareness or specific pedagogical progressions or aims

Related approaches

VISL: Visual Interactive Syntax Learning (Bick 2001, 2005a,b)

- Similarities
  - Emphasis on language awareness:
    - VISL offers games and visual presentations to foster knowledge of syntactic forms and functions.
  - Automatic exercise generation:
    - The “exercise building tool” KillerFiller automatically creates slot-filler exercises from texts.
- Differences
  - KillerFiller intended as evaluative tool, not for teaching.
  - Annotated corpora and databases used as text base.
  - Sentences presented in isolation, not in context.
  - Slots determined by general category (e.g., prepositions, verbs), not more specific or other linguistic features.

Related approaches

Generating cloze tests

Automatic generation of multiple choice “cloze tests” (FIB) for language testing and vocabulary drill

(cf., e.g. Coniam 1997; Irvine and Kyllonen 2002; Deane and Sheehan 2003; Huang et al. 2005; Liu et al. 2005a,b).

- Sumita et al. (2005): automatic generation of FIB questions for testing English proficiency
  - \+ selection of seed sentence mentioned as relevant issue
  - \+ uses web to test whether potential distractor items are indeed incorrect
  - − addresses testing, not pedagogical exercise progression
  - − sentences not selected by learner or contextualized

Related approaches

Cognate exercises, FL reading support, FL text retrieval

- False friend (cognate) exercise creation (Wagner 2004):
  - uses authentic corpus material
  - NLP use very limited: only identifies major part-of-speech tokens (those which potentially have cognates)
- Support tools for reading texts in a foreign language support awareness by highlighting linguistic features:
  - *Glosser-RuG* project (Nerbonne et al. 1998): supports reading of French texts for Dutch learners with on-line, context dependent dictionary, morphological analysis, and examples of word use in corpora.
  - *COMPASS* project (Breidt and Feldweg 1997): similar, but focuses on multi-word lexemes
- *REAP* project: Automatic retrieval of FL texts appropriate to learner level from the web (Brown et al. 2005).

Summary

- Fostering awareness is a well-motivated component of FLT, and the web-based activities can
  - follow the common pedagogical practice from presentation of the target language to controlled practice
  - be based on texts selected by the learner and remain fully contextualized as part of the text
- We discussed instances of such activity progressions for targets that are
  - lexical (pronouns),
  - structural and morphological (passive),
  - word order (adverb placement), or
  - semantic (tense/aspect)

  in nature.

Summary (cont.)

- The envisaged approach requires explicit learner and instructor models to select adequate texts from the user-determined text base.
- A range of robust NLP tools are required to
  - select text from text base, using requirements from learner and instructor model
  - identify target categories for presentation and exercise creation
- Creating sequences of activities fostering awareness of language forms and rules is an interesting opportunity for NLP-based FLT tools – one that we hope will attract a joint effort by our research community.

References

Antoniadis, G., S. Echinard, O. Kraif, T. Lebarbé, M. Loiseau and C. Ponton (2004). NLP-based scripting for CALL activities. In Erhard Hinrichs Lothar Lemnitzer, Detmar Meurers (ed.), *COLING 2004 eLearning for Computational Linguistics and Computational Linguistics for eLearning*. Geneva, Switzerland: COLING, pp. 18–25. http://acl.ldc.upenn.edu/coling2004/W6/pdf/3.pdf.

Bick, Eckhard (2001). The VISL System: Research and applicative aspects of IT-based learning. In *Proceedings of NoDaLiDa (Uppsala)*. .

Bick, Eckhard (2005a). Grammar for Fun: IT-based Grammar Learning with VISL. In Peter Juel (ed.), *CALL for the Nordic Languages*, Copenhagen: Samfundslitteratur, Copenhagen Studies in Language, pp. 49–64.

Bick, Eckhard (2005b). Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL. In Henrik Holboe (ed.), *Nordic Language Technology, Arbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (Yearbook 2004)*, Copenhagen: Museum Tusculanum, pp. 171–186.

Breidt, Elisabeth and Helmut Feldweg (1997). Accessing Foreign Languages with COMPASS. *Machine Translation* 12(1–2), 153–174. Special Issue on New Tools for Human Translators.

Brown, Jonathan and Maxine Eskenazi (2004). Retrieval of authentic documents for reader-specific lexical practice. In InSTIL/ICALL (2004).

Brown, Jonathan and Maxine Eskenazi (2005). Student, text and curriculum modeling for reader-specific document retrieval. In *Proceedings of the IASTED International Conference on Human-Computer Interaction*. Phoenix, Arizona.

Brown, Jonathan, Gwen Frishkoff and Maxine Eskenazi (2005). Automatic Question Generation for Vocabulary Assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 819–826.

Burstein, Jill and Claudia Leacock (eds.) (2005). *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, Ann Arbor, Michigan. Association for Computational Linguistics.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *Computer Assisted Language Instruction Consortium* 16(2–4), 15–33.

Deane, P. and K. Sheehan (2003). Automatic item generation via frame semantics. Education Testing Service: http://www.ets.org/research/dload/ncme03-deane.pdf.

Dickinson, Markus (2005). Error detection and correction in annotated corpora. Ph.D. thesis, The Ohio State University.

Dickinson, Markus and W. Detmar Meurers (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*. Budapest, Hungary, pp. 107–114. http://ling.osu.edu/~dm/papers/dickinson-meurers-03.html.

Ellis, Nick (1994). Implicit and Explicit Language Learning - An Overview. In *Implicit and Explicit Learning of Languages*, San Diego, CA: Academic Press, pp. 1–31.

Huang, Shang-Ming, Chao-Lin Liu and Zhao-Ming Gao (2005). Computer-assisted item generation for listening cloze tests and dictation practice in English. In *Proceedings of the 4th Int. Conf. on Web-based Learning*.

Language awareness through re-use of NLP technology
Luiz Amaral, Vanessa Metcalf and Detmar Meurers
Background
NLP in ICALL
Pedagogical grounding
Proposal
Modeling FLT practice
Progression in our approach
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect
Realizing proposal
1. Obtaining the text base
2. Selecting texts from base
3. Creating ex. progression
Some challenges
Related approaches
The MIRTO project
The VISL project
Generating cloze tests
Cognate exercises, FL reading support, FL text retrieval
Summary
References
THE OHIO STATE UNIVERSITY
22 / 22

InSTIL/ICALL (2004). *InSTIL/ICALL 2004 Symposium on Computer Assisted Learning, NLP and speech technologies in advanced language learning systems*. Venice, Italy: International Speech Communication Association (ISCA).

Irvine, S.H. and P.C. Kyllonen (eds.) (2002). *Item Generation for Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lightbown, P. M. and N Spada (1999). *How languages are learned*. Oxford: Oxford University Press.

Liu, Chao-Lin, Chun-Hung Wang and Zhao-Ming Gao (2005a). Using lexical constraints for enhancing computer-generated multiple-choice cloze items. *Int. Journal of Computational Linguistics and Chinese Language Processing* 10.

Liu, Chao-Lin, Chun-Hung Wang, Zhao-Ming Gao and Shang-Ming Huang (2005b). Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In Burstein and Leacock (2005), pp. 1–8.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. De Bot, C. Kramsch and R. Ginsberg (eds.), *Foreign language research in cross-cultural perspective*, Amsterdam: John Benjamins, pp. 39–52.

Long, M. H. (1996). The role of linguistic environment in second language acquisition. In W. C. Ritchie and T. K. Bhatia (eds.), *Handbook of second language acquisition*, New York: Academic Press, pp. 413–468.

Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classroom. *Language Learning* 48, 183–218.

Nerbonne, John, Duco Dokter and Petra Smit (1998). Morphological Processing and Computer-Assisted Language Learning. *Computer Assisted Language Learning* 11(5), 543–559.

Norris, J. and L. Ortega (2000). Effectivenss of L2 Instruction: A Research Synthesis and Quantitative Meta-Analysis. *Language Learning* 50(3).

Schmidt, R. (1995). Consciousness and foreign language: A tutorial on the role of attention and awareness in learning. In R. Schmidt (ed.), *Attention and awareness in foreign language learning*, Honolulu: University of Hawaii Press, pp. 1–63.

Schulz, Renate A. (2002). Hilft es die Regel zu wissen um sie anzuwenden? Das Verhältnis von metalinguistischem Bewusstsein und grammatischer Kompetenz in DaF. *Die Unterrichtspraxis—Teaching German* 35(1), 15–24.

Sumita, Eiichiro, Fumiaki Sugaya and Seiichi Yamamoto (2005). Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In Burstein and Leacock (2005), pp. 61–68.

Wagner, Joachim (2004). A false friends exercise with authentic material retrieved from a corpus. In InSTIL/ICALL (2004). .

Language awareness through re-use of NLP technology

Luiz Amaral, Vanessa Metcalf and Detmar Meurers

Background
NLP in ICALL
Pedagogical grounding

Proposal
Modeling FLT practice
Progression in our approach
Example 1: Pronouns
Example 2: Passive
Example 3: Adverb placement
Example 4: Tense and Aspect

Realizing proposal
1. Obtaining the text base
2. Selecting texts from base
3. Creating ex. progression
Some challenges

Related approaches
The MIRTO project
The VISL project
Generating cloze tests
Cognate exercises, FL reading support, FL text retrieval

Summary

References

22 / 22