

CALL: THE POTENTIAL OF LINGWARE AND THE USE OF EMPIRICAL LINGUISTIC DATA

Dan Tufis

Romanian Academy & Research Institute for Informatics

13, "13 Septembrie", 74311, Bucharest 5, RO

e-mail:tufis@ul.ici.ro

Language technology has significantly evolved during the last decade. However, the community of language learning seems to ignore this development, most of the existing language learning systems drawing their enhancements from other sources, such as hypertext, multimedia, interactive video, information retrieval. Despite some spectacular progress made at the level of interface, several fundamental language learning principles, are only partially met. Nevertheless, the hypermedia technology did solve one very important aspect of computer-assisted learning by putting the student in a visual environment. Minimizing cultural differences they've been able to draw on shared background knowledge (microworld immersiveness). Other important aspects of typical immersion-based approaches, i.e. natural learning, such as mixed-initiative, fault-tolerance, dialogue repair, cooperative behaviour, etc. are still in their infancy. In real settings learners freely interact with their environment (parents, tutors), taking turns, asking for explanations, shifting topics, etc. The language produced by the learner is more often than not agrammatical, yet this does not prevent the tutor to proceed with the dialog. Error correction is usually done contextually, by drawing either explicitly attention to the deviation, by producing a similar but correct sentence, or by simply ignoring the mistake leaving its correction for later.

There are many AI and CL programs solving various specific CALL-relevant problems. If assembled properly, these pieces could result in very powerful language learning systems.

Lexical thesauri Since word acquisition is a crucial part of language learning, a thesaurus such as WordNet is practically a must in a broader CALL system. Such a tool could provide, lists of synonyms, antonyms, hyper/hyponyms, meronyms and contexts in which these words are used.

Parsers While finding a freeware parser is not a problem anymore (if you don't know where to get hold of one, —just send me an e-mail,— we have developed different parsers), it is no easy to find the right kind of grammar for teaching purposes. Such a grammar should have at least the following qualities: parsing the student's input it should be error tolerant yet having a broad enough coverage for being useful both for beginners and for advanced students. In

order to deal with the student's errors in a principled way, the grammar should anticipate typical errors and annotate them for automatic recovery and explanation generation. While introspection and observation are a first step in determining typical errors, data gathered in a corpus are a much more reliable approach. Corpus linguistics has become a very promising and active area of investigation. However, few corpora (if any) have been gathered with respect to register. Such corpora should contain among other things: native tongue of the speaker, the complexity of the text under consideration; error/correction markup, etc.

Generators In order to communicate with the student, the CALL system should be able to produce natural language output. It is debatable whether the system should communicate with the student only in the target language, or, whether under specific circumstances (such as error correction mode) it should also be able to generate texts also in the student's mother tongue. Such a pedagogical decision has of course important consequences on the system's architecture: a bilingual approach, requiring several components of a MT system. Again, there are several NL generators (most of them head-driven) available in the public domain.

Semantic interpreters/generators Unlike the previous modules, the one in charge of the bidirectional mapping of the syntactic structures onto the knowledge structures of the microworld, is very sensitive to factors such as discourse universe, tutorial strategies, student profile etc. That's why it is not easy to find a ready made plug-in module for CALL systems. Yet, there are several generic programs that support the contextual interpretation of the student's input (linguistic or graphical), tracking his/her goals and providing cooperative responses. Intelligent planners (linear or nonlinear) could be used in plan-based tutorials, with the microworlds defining the possible limits of departure from the expectation-based tutorial plans. User modelling subsystems, tuned to the language learning problems could provide valuable support in dealing with notorious difficult problems (discovering student's misconceptions, tailoring explanations to the level of the student's expertise, etc.)

Speech synthesizers and prosody processors Speech technology is definitely a valuable candidate

for CALL tools. In spite of the current gap between speech technology and natural language processing, language learning is a very promising area where the two fields could meet. One could easily imagine a scenario where the student is asked to utter a word or a sentence, which are then compared and corrected against the tutor's pronunciation. With a graphical representation of the two pronunciations (waveform, pitch, duration etc.) and a means to operate on them (e.g. mouse dragging the waveform, followed by the synthesized result) the pedagogical value and user acceptability of a CALL system would certainly be greatly enhanced.