Towards Deeper Understanding and Personalisation in CALL

Galia Angelova, Albena Strupchanska, Ognyan Kalaydjiev, Milena Yankova
Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria
{galia, albena, ogi, myankova}@lml.bas.bg
Svetla Boytcheva, Irena Vitanova
Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
svetla@fmi.uni-sofia.bg, itv@gmx.co.uk
Preslav Nakov
University of California at Berkeley, USA, nakov@eecs.berkeley.edu

Abstract

We consider in depth the semantic analysis in learning systems as well as some information retrieval techniques applied for measuring the document similarity in eLearning. These results are obtained in a CALL project, which ended by extensive user evaluation. After several years spent in the development of CALL modules and prototypes, we think that much closer cooperation with real teaching experts is necessary, to find the proper learning niches and suitable wrappings of the language technologies, which could give birth to useful eLearning solutions.

1 Introduction

The tendency to develop natural interfaces for all users implies man-machine interaction in a natural way, including natural language too, both as speech and as free text. Many recent eLearning research prototypes try to cope with the unrestricted text input as it is considered old-fashioned and even obsolete to offer interfaces based on menu-buttons and mouse-clicking communication only. On the other hand, the available eLearning platforms such as WebCT [1], CISCO [2], and the freeware HotPotatoes [3], are far from the application of advanced language technologies that might provide interfaces based on speech and language processing. They represent complex communication environments and/or empty shells where the teacher uploads training materials, drills, etc. using specialised authoring tools. Recently on-line voice communication between teachers and students has been made available as well, via fast Internet in virtual classrooms, but no speech or language processing has been considered. So there is a deep, principle gap between the advanced research on tutoring systems and the typical market eLearning environments addressing primarily the communication needs of the mass user.

In what follows we will concentrate on research prototypes integrating language technologies in eLearning environments. In general, such prototypes might be called Intelligent Tutoring Systems (ITS) and we will stick to this notion here. Most of the systems discussed below address Computer-Aided Language Learning (CALL) but language technologies are applied for automatic analysis of user utterances in other domains too. A review of forty Intelligent CALL systems (Gamper, 2002) summarises the current trends to embed "intelligence" in CALL. What we developed (and report here) might be considered intelligent because of the integration of reasoning and the orientation to adaptivity and personalisation.

This paper is structured as follows. In section 2 we consider the task of semantic analysis of the learner's input, which is an obligatory element when the student is given the opportunity to type in freely in response to ITS's questions and/or drills. Section 3 deals with Information Retrieval (IR) approaches for measuring document similarity, which are integrated in ITS as techniques for e.g. assessing the content of student essays or choosing the most relevant text to be shown to the learner. Section 4 discusses how the language technologies in question can provide some adaptivity of the ITS, as a step towards personalisation. In section 5 we summarise the current results regarding the evaluation of our prototypes with real users. Section 6 contains the conclusion.

2 Semantic Analysis in ITS

Although the automatic analysis of user utterances is a hot research topic, it achieved only partial success so far. The review (Nerbonne, 2002) shows that Natural Language Processing (NLP) is often integrated in CALL, as the domain of language learning is the first "candidate" for the application of computational linguistics tools. Different language technologies are applied in "programs designed to help people learn foreign languages": morphology and lemmatisation, syntax, corpus-based language acquisition, speech processing, etc. Attempts to implement automatic semantic analysis of free text input are relatively rare, due to the sophisticated paradigm and the default assumption that it will have a very limited success (i.e. will be the next failure).

The famous collection of papers (Holland, 1995) presents several systems, which integrate NLP modules in different ways. The most advanced one regarding semantic analysis is MILT (Dorr, 1995), where the correctness as well as the appropriateness of the student's answer are checked by matching them against expectations. This is performed in the context of a question-answering session, where expected answers are predefined by the foreign language tutor. The language-independent internal semantic representation is based on lexical conceptual structures, which (following Jackendoff) have types, with primitives and propositional descriptions along different dimensions and fields etc. Consider as an example that the teacher has specified that "*John ran to the house*" is a correct answer. This sentence is processed by the system and the following lexical conceptual structure is obtained:

```
[Event GO Loc
  ([Thing JOHN],
   [Path TO Loc ([Position AT Loc ([Thing JOHN],
                  [Property HOUSE])])],
   [Manner RUNNINGLY])]
```

which is stored by the tutoring system and later matched against the student's answer. If the student types "*John went to the house*", the system must determine whether this matches the teacher-specified answer. The student's sentence is processed and respresented as:

```
[Event GO Loc
  ([Thing JOHN],
   [Path TO Loc ([Position AT Loc ([Thing JOHN],
                  [Property HOUSE])])])]
```

The matcher compares the two lexical conceptual structures and produces the output:

```
Missing: MANNER RUNNINGLY
         INCORRECT ANSWER
```

Put another way, the comparison of internal representations helps in the diagnostics of *semantic errors* and *appropriateness,* which are two different notions. For instance "*John loves Marry*" is a semantically correct sentence, but it is not an appropriate answer when the system expects "*John ran to the house*". Further discussions in (Dorr, 1995) show that the matching scenario is very useful in question-answering lessons, which are formulated as sets of free response questions associated with a picture or text in the target language. In an authoring session, the lesson designer enters the texts, the questions and a sample appropriate answer to each question. At lesson time, the questions are presented to the student who answers them. If the predefined answers are general enough, the system will flexibly recognise a set of possible answers. For instance, the student might answer:

Juan died or *Carlos killed Juan* or
Carlos murdered Juan

to the question "*What happened to Juan*", which checks the comprehension of a simple newspaper article. The matching technique can be extended to check whether the translations of sentences into the target language are correct etc. Even as an earlier implementation at the "concept demonstration stage", this prototype identifies possible solutions for the integration of semantic analysis in CALL.

A recent system Why2-Atlas (VanLehn, 2002), based on deep syntactic analysis and compositional semantics, aims at the understanding of student essays in the domain of physics. Why2-Atlas is developed within the project Why2 where several different NL processing techniques are compared (Rose, 2002). The sentence-level understander converts each sentence of the student's essay into a set of propositions. For instance, the sentence

"Should the arrow have been drawn to point down?"

is to be (roughly speaking) converted to
∃e∈events, ∃v∈vectors, ∃s∈draw(e,s,v) & tense (e, past)&mood(e,interrog)&direction(v,down).

As the authors note in (VanLehn, 2002), this is just an approximation of the real output, which illustrates the challenge of converting words into the appropriate domain-specific predicates. The left-corner parser LCFlex copes with ungrammatical input by skipping words, inserting missing categories and relaxing grammatical constraints as necessary in order to parse the sentence. For instance, "*Should the arrow have been drawn point down?*" would parse. In case of too many analyses, the parser uses statistical information about the word roots frequency and the grammatical analyses in order to determine the most likely parse. If no complete analysis can be produced, a fragmentary analysis will be passed for further processing. The fragments present "domain-specific predicates that are looking for argument fillers, and domain-specific typed variables that are looking for arguments to fill". If the symbolic approach for input analysis via logical forms fails, a probabilistic one will be used as an alternative.

What is particularly interesting for us here, is the discourse-level understander (VanLehn, 2002) which, given logical forms, outputs a proof. Topologically, this is a forest of interwoven trees, where the leaves are facts from the problem statement or assumptions made during the proof construction. The roots (conclusions) are student's propositions. Consider the example:

Question: Suppose you are in a free-falling elevator and you hold your keys motionless in front of your face and then let go. What will happen to them? Explain.

Answer: The keys will fall parallel to the person face because of the constant acceleration caused by gravity but later the keys may go over your head because the mass of the keys is less.

The essay answer will be translated into four propositions, which will be passed to the discourse understander. The first one (*keys fall parallel to the person's face*) is correct and becomes the root of the proof. The second one (*gravitation acceleration is constant*) corresponds to facts from the knowledge base. The third proposition (*keys go over the person's head*) is based on the common misconception that heavier objects fall faster, which is pre-stored in the knowledge base as well, it becomes the root of the proof. The last one (*the mass of the keys is less*) corresponds to a node of the interior of the proof of the third proposition. Once a proof has been constructed, a tutorial strategist performs an analysis in order to find flaws and to discuss them. Here the major one is the misconception "heavier objects fall faster". The tutoring goals have priorities as follows: fix misconceptions, then fix self-contradictions, errors and incorrect assumptions, and lastly elicit missing mandatory points. The Why2 project in general, and Why2-Atlas in particular, illustrate the recent trends in the ITS development:

(i) mixture of symbolic and stochastic approaches in order to cope with the free NL input;
(ii) application of shallow and partial analysis as an alternative to the deep understanding;
(iii) integration of AI techniques (esp. reasoning and personalisation);
(iv) organisation of bigger projects with considerable duration to attack the whole spectre of problems together (incl. development of authoring tools, systematic user evaluation at all stages, several development cycles and so on).

We are experienced in the application of semantic analysis to CALL in two scenarios. The first one[1], in 1999-2002, deals with deep understanding of the correct sentences and proving the domain correctness and the appropriateness of the logical form of each one. The second one focuses on the integration of shallow analysis and partial understanding in CALL (Boytcheva, 2004). The system described in (Angelova, 2002) is a learning environment for teaching English financial terminology to adults, foreigners, with intermediate level of English proficiency. The prototype is a Web-based learning environment where the student accomplishes three basic tasks: *(i)* reading teaching materials, *(ii)* performing test exercises and *(iii)* discussing his/her own learner model with the system. The project is oriented to learners who need English language competence as well as expertise in correct usage of English financial terms. This ambitiously formulated paradigm required the integration of some formal techniques for NL understanding, allowing for analysis of the user's answers to drills where the student is given the opportunity to enter free natural language text (normally short discourse of 2-3 sentences). The morphological, syntax and semantic analysis is performed by the system Parasite (Ramsay, 2000), developed in UMIST. After the logical form has been produced for each correct sentence, the CALL environment has to determine whether the student's utterance matches the expected appropriate answer in the current learning situation. A special prover has been developed, which checks whether the logical form of the answer is "between" the minimum and maximum predefined answers (Angelova, 2002). Unlike MILT (Dorr, 1995), we think that the correct answer has to be subsumed by the maximum expected one, i.e. there is not only a lower but also an upper limit on the correctness. Table 1 lists examples for all diagnostic cases from user's perspective, by sentences in natural language. Please note that nowadays the deductive approach can be relatively efficient, as our prover (in Sicstus Prolog) works on-line, integrated in a Web-based environment, in real time with several hundred meaning postulates. Proofs are certainly based on a predefined ontology of the domain terms, which in this case is a lexical one since the terms are treated as words with special lexical meaning encoded in the meaning postulates thus forming a hidden hierarchy of meanings. The conceptual and lexical hierarchy of meanings are further discussed in (Angelova, 2004).

However, we discovered that deep semantic analysis is difficult to integrate in CALL. First, this requires enormous amount of efforts for the meaning postulates acquisition. While hierarchy of terms is reusable, as it is in fact the domain model, the propositions, which encode the lexical semantics are somewhat application and domain specific and therefore difficult to reuse or to transfer to another domain (moreover they are bound to the domain words). Implementing the prover and testing the definitions and the inference

[1] In Larflast, a Copernicus Joint Research Project.

Case	Sample of learner's utterance	Discussion
Kernel (predefined minimum answer)	Primary market is a financial market that operates with newly issued debt instruments and securities.	The logical form is pre-stored in the system as a Kernel.
Cover (predefined maximum answer)	Primary market is a financial market that operates with newly issued debt instruments and securities and provides new investments and its goal is to raise capital.	The logical form is pre-stored in the system as a Cover.
1.Correct answer	Primary market is a financial market that operates with newly issued debt instruments and securities *and provides new investments.*	This logical form is between the Kernel and the Cover.
2a) Incomplete answer	Primary market is a financial market that operates with newly issued securities.	Missing Kernel term: debt instruments.
2b) Specialisation of concepts from the definition	Primary market is a financial market that operates with newly issued bonds.	Bond is a specialisation of security; Missing: debt instruments.
2c) Paraphrase using the concept definition	Primary market is a financial market that operates with new emissions of stocks, bonds and other financial assets.	New emissions = newly issued; stocks, bonds and other financial assets = debt instruments and securities.
3a) Partially correct	Primary market is a financial market that operates with newly issued debt instruments and securities for instant delivery.	Wrong: for instant delivery.
3b) Generalisation of concepts from the definition	Primary market is a market that operates with newly issued financial instruments.	Market is a generalisation of financial market**;** Financial instruments are generalisation of debt instruments and securities.
4. Partially correct	Primary market is a financial market that operates with newly issued securities for instant delivery *and provides new investments.*	Wrong: for instant delivery; Missing: debt instruments.
5. Wrong answer	Primary market is an organisation in which the total worth is divided into commercial papers.	Wrong: an organisation in which the total worth is divided into commercial papers; Missing: financial market that operates with newly issued debt instruments and securities.
6. Wrong answer	Primary market *provides new investments* for instant delivery.	Wrong: for instant delivery; Missing: financial market that operates with newly issued debt instruments and securities;
7. Partially correct	Primary market is a financial market that operates with newly issued securities *and provides new investments.*	Missing: debt instruments.
8. Wrong answer	Primary market *provides new investments.*	Missing: financial market that operates with newly issued debt instruments and securities.

Table 1: Decisions about erroneous answers according to the configuration of the logical forms of the predefined minimal, maximal and the current learner's answer (see also Angelova, 2002).

procedures with several hundred predicates required approximately one man-year for an AI expert who worked closely with domain experts. Second, the result is not perfect from the perspective of the user who has to answer with correct and full sentences (see section 5 for details). Thus our recent work (Boytcheva, 2004) is directed towards integration of shallow and deep semantic techniques in CALL systems. We use shallow parsing, which allows for the processing of both syntactically incorrect and incomplete answers. However, during the user's utterances evaluation we use deep semantic analysis concerning the concepts and the relations that are important for the domain only. Users' answers are represented as logical forms, convenient for the inference mechanism, which takes into account the type hierarchy and is elaborated in domain-specific points only. Thus the combination of shallow and deep techniques gives the users more freedom in answering, i.e. various utterances to express themselves without impeding the evaluation process. The idea to apply the shallow NLP techniques in CALL was inspired by their successful application in IE for template filling. The assessment of user knowledge in a specific domain can be viewed as a kind of template filling, where the templates correspond to concepts and relations relevant to the tested domain.

3 Exploiting Document Proximity in ITS

There is a huge demand for intelligent systems that can handle free texts produced by the learners in eLearning mode. As most of the courses being taught are represented as texts, the challenge is to compare one text to another. Since the phrasing will not be the same in both texts, the comparison needs to be performed at the semantic level. One solution is sketched above: translate the student's text to a set of logical forms and then apply symbolic approaches for their assessment. Unfortunately, there are only few research prototypes that address the problem from this perspective, which are very expensive and have delivered only partially applicable results so far. Another option is to try to exploit the IR techniques we have at hand in order to check for instance whether the student's answer contains the "right words" (in which case it would be a good writing, since it would be similar to the expectation). A natural choice for assessing the usage of the "right words" is the so-called Latent Semantic Analysis (LSA) as it reveals the latent links between the words and phrases, especially when it is trained with enough samples. Below we briefly overview the application of LSA in eLearning and our experiments in this direction.

The classical LSA method, as proposed in (Deerwester, 1990) is a bag-of-words technique, which represents the text semantics by assigning vectors to words and texts (or text fragments). Indeed, knowing how words are combined to encode the document knowledge is a kind of semantic representation of the word meaning and text semantics. The underlying idea is that words are semantically similar, if they appear in similar texts, and texts are semantically similar, if they contain similar words. This mutual word-text dependency is investigated by building a word-text matrix, where each cell contains the number of occurrences of word X in document Y, after which the original matrix is submitted to Singular Value Decomposition – a transformation that is meant to reveal the hidden (latent) similarity between words and texts. This produces a vector of low dimensionality (the claim is that 300 is near optimal) for each word and for each text. The similarity between two words, two texts, or a word and a text, is given by the cosine of the angle between their corresponding vectors (the cosine is the most popular similarity measure). Therefore, the similarity between two words or two sets of words is a number between –1 (lowest similarity) and 1 (highest similarity). Without morphology and grammar rules, syntactical analysis, and manually encoded implicit knowledge, LSA is considered successful in various experiments including assessment of student essays.

For the purposes of assessment, usually a high-dimensional space is computed from texts describing the domain (most often the available electronic version of the course). Each word from the domain as well as the student's essay are juxtaposed a vector, usually a 300-dimensional one. The student gets as feedback an assessment score and/or an indication of the topics/aspects that are not covered well by the essay. The Intelligent Essay Assessor (IEA) (Foltz 1999a, Foltz 1999b) is based on reference texts (manually pre-graded essays) and assigns *a holistic score* and *a gold standard score*. The former is computed by seeking the closest pre-graded essay and returning its grade (i.e. the current one is scored as the closest pre-graded one), while the latter is based on a standard essay written by an expert. It returns the proximity between the student's essay and the expert's one. An experiment with 188 student essays showed a correlation of 0.80 between the IEA scores and teacher's ones, which is a very high similarity. However, IEA outputs no comments or advice regarding the student essay. The Apex system (Lemaire, 2001) performs a semantic comparison between the essay and the parts of the course previously marked as relevant by the teacher. The whole student essay is to be compared to each of these text fragments. For instance, if the student has to write an answer to the question "*What were the consequences of the financial crash of 1929*?", the essay is compared to the following sections of the teaching course: *The political consequences in Europ*e, *Unemployment and povert*y, *The economical effect*s, *The consequences until 194*0. An experiment with 31 student essays in the domain of Sociology of Education exhibited a correlation of 0.51 between Apex grades and teacher's ones, which is close to the correlation agreement between two human graders in this literary domain. Select-a-Kibitzer (Wiemer-Hastings, 2000) aims at the assessment of essay composition. Students are required to write on topics like: "*If you could change something about school, what would you change*?". The assessment module is based on reference sentences of what students usually discuss about school (food, teachers, school hours, etc.). Several kinds of feedback are delivered to the student, concerning the text coherence, the kind of sentences or the topic of the composition. For example, the advice regarding coherence can be: "*I couldn't quite understand the connection between the first sentence and the second one. Could you make it a bit clearer? Or maybe make a new paragraph*." (Here the underlying idea is that totally new words in the subsequent sentence normally concern another topic, i.e. this fits to a new paragraph). A principled criticism of these three recent systems is that the bag-of-words model does not take into consideration the grammar correctness and the discourse structure, i.e. two essays with the same sentences structured in a different order would be scored identically (which is a funny idea from an NLP perspective). A further example illustrates attempts to combine the strengths of the bag-of-words and the symbolic approaches, while trying to avoid some of their weaknesses. CarmelTC (Rose, 2002), a recent system which analyses essay answers to qualitative physics questions, learns to classify units of text based on features extracted from a syntactic analysis of that text. The system was developed inside the Why2-Atlas conceptual physics tutoring environment for the purpose of grading short essays written in response to questions such as "*Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain*". CarmelTC's goal is not to assign a letter grade to student essays, but to tally which set of 'correct answer' aspects are present in student essays (e.g. a satisfactory answer to the example question above should include a detailed explanation of how the Newton's 1st law applies to this scenario. Then the student should infer that the pumpkin and the man will continue at the same constant horizontal velocity that they both had before the release. Thus, they will always have the same displacement from the point of release. Therefore, after the pumpkin rises and falls, it will land back in the man's hands. The "presence" of certain sentences is checked by word classification). The evaluation shows that the hybrid CarmelTC approach achieves 90% precision, 80% recall and 85% F-measure, and thus outperforms the pure bag-of-words run of LSA, which scores 93% precision, 54% recall and 70% F-measure (Rose, 2002).

Our experiments with LSA (Angelova, 2002) were focused on finding financial texts, which are appropriate to be shown as teaching materials in a particular learning situation. Given a set of key-words, agents retrieve texts from well-known financial sites and store them to the servers of our environment for further assignment of appropriateness. We implemented the classical LSA scenario and applied it as a filtering procedure, which assigns off-line a similarity score to each new text. The text archive consisted of 800 *most relevant* readings, which represent HTML-pages with textual information (elements signaling prevailing technical content, e.g. tables, have been excluded). These texts are offered as suggested readings but are also used for building dynamic concordances, which show samples of terms usages to the learner. The latter may be displayed in cases of language errors to drills where the student makes linguistic mistakes. Choosing this option (*view samples*) is up to the student. The dynamic nature of the text collection ensures the appearance of new samples, which makes the browsing interesting at every run.

4 Personalisation

Our learning environment supports personalisation as follows:

- as a step towards instructional as well as content planning: a planner (the so-called pedagogical agent) plans the next learner's moves across the hypertext pages which, technically, constitute the Web-site of our tutoring system; these moves are between *(i)* performing drills and *(ii)* choices for suggestion of readings, which may be either texts from Internet or especially generated Web-pages. The pedagogical agent deals with both presentational and educational issues. The local planning strategy aims at creating a complete view of the learner's knowledge of the current concept. It supports movements between drills with increasing complexity, when the student answers correctly. The global planning strategy determines movements between drills testing different concepts, going from the simple and general concepts to the more specific and complex notions.
- as a step towards personalised IR: an LSA-filter assigns proximity score to constantly updated texts, which are stored as suggested readings. This allows for constant update of the system's

text archive and, following the practice at the main financial sites, provides up-to-date news and readings, which may be used as texts for different teaching purposes. As key words for initial collection of texts, the **not_known** and **wrongly_known** terms from the learner's models are chosen, so the CALL system always stores the proper relevant text for each student.

The adaptivity is provided using an ontology of financial terms as a backbone of all system's resources. No matter whether these are *conceptual* (e.g. knowledge base), *linguistic* (e.g. lexicons, meaning postulates, etc) or *pedagogical* resources (e.g. set of preliminary given drills or learner model, which is dynamically constructed at run-time), the ontology always represents the unifying skeleton as all chunks of knowledge are organised around the terms–labels. In addition to the *is-a* partition, we support in the knowledge base explicit declarations of the perspectives or viewpoints. E.g., the **isa_kind/4** clause:

```
isa_kind(security, [bond, hybrid_security, stock],
[exhaustive, disjoint],
'status of security holder: creditor or owner')
```

means that the securities are disjoint and exhaustively classified into bonds, stocks and hybrid securities depending on the status of their owner. These comments provide nice visualisation (Angelova, 2004).

5 User Study and User Evaluation

Larflast started with a user study of how foreigners – adults acquire domain terminology in their second language. In fact the acquisition is closely related to the elicitation of domain knowledge, especially in a relatively new domain (students have to learn simultaneously a subject with its terminology and its specific language utterances). Mistakes are linguistically-motivated but wrong domain conceptualisations contribute to the erroneous answers as well. Erroneous answers appear in terminology learning due to the following reasons:

- **Language errors** (spelling, morphology, syntax);
- **Question misunderstanding,** which causes wrong answer;
- **Correct question understanding,** but **absent knowledge of the correct term**, which implies usage of paraphrases and generalisation instead of the expected answer;
- **Correct question understanding,** but **absent domain knowledge,** which implies specialisation, partially correct answers, incomplete answers and wrong answers.

This classification influenced considerably the design of the prover's algorithms, i.e. the decision how to check of the appropriateness of the student answer. The diagnostics shown in Table 1 follows closely the four reasons above.

Our learning prototype was tested by *(i)* two groups of university students in finance with intermediate knowledge of English, *(ii)* their university lecturers in English, and *(iii)* a group of students of English philology. The system was evaluated as a CALL-tool for self-tuition and other autonomous classroom activities, i.e. as an integral part of a course in "English for Special Purposes". The learners could test their knowledge through the specially designed exercises, compare their answers to the correct ones using the generated feedback (immediate, concrete and time-saving, it comes in summary form, which is crucial in order to accomplish the system's use autonomously) and extract additional information from the suggested readings and concordancers.

The users liked the feedback after performing drills, immediately after they prompted erroneous answers to exercises where this term appears. They evaluated positively the visualisation of the hierarchy as well as the surrounding context of texts and terms usages organised in a concordancer, which is dynamically built and centred on the terms discussed at the particular learning situation. The teachers were very pleased to have concordancers with contiguously updated term usages; they would gladly see such a language resource integrated in a further authoring tool, because searching suitable texts in Internet is a difficult and time-consuming task.

Unfortunately the learners were not very enthusiastic regarding the free NL input, as it permits relatively restricted simple answers and does not go beyond the human capacity of the teacher. The main disappointment of both learners and teachers is the system's inability to answer *why*, i.e. while the formal semantics and reasoning tools provide extremely comprehensive diagnostic about the error type, they tell nothing about the reason. Fortunately, all users liked the fact that there were numerous examples of terms usages from real texts whenever morphological or syntax errors were encountered in the free NL input. Thus we conclude with a certain pessimism concerning the appropriateness of today's formal semantic approaches in ITS and much optimism that data-driven corpus techniques, if properly applied, fit quite well to the adaptive ITS. What is still desirable regarding the filtering module is to restrict the genre of the suggested readings, since the current texts are freely collected from the

Internet and some of them should be used as teaching materials (LSA cannot recognise the text educational appropriateness since it considers the terms' occurrences only; other supervised techniques such as text categorisation might improve the filtering, if properly integrated).

As a possible improvement of the current paradigm for formal analysis, we turned recently to partial analysis, which gives more flexibility to the students to enter phrases instead of full sentences (Boytcheva, 2004).

6 Conclusion

The conclusion is that teachers as well as learners like CALL systems that are easy to integrate in the typical educational tasks, i.e. the area of language learning has well-established traditions and the experimental software is well-accepted, only if it is really useful and facilitates the learning process. Our feeling is that all attempts to integrate language technologies in CALL should be closely related to testing the laboratory software with real students. At the same time, cooperation with teachers is an obligatory condition as the necessary pedagogical background is often missing in the research environments where normally the NLP applications and language resources appear. Language technologies have a long way to go, until they find the proper wrappings for integration of advanced applications and the necessary resources into useful CALL systems.

References

[1] WebCT, http://www.webct.com/

[2] CISCO, http://cisco.netacad.net/

[3] HotPotatoes, http://web.uvic.ca/hrd/hotpot/

Angelova G., Boytcheva, Sv., Kalaydjiev, O. Trausan-Matu, St., Nakov, P. and A. Strupchanska. *2002. Adaptivity in Web-based CALL* In Proc. of ECAI'02, the 15th European Conference on AI, IOS Press, pp. 445-449.

Angelova G., Strupchanska, A., Kalaydjiev, O., Boytcheva, Sv. and I. Vitanova. 2004. *Terminological Grid and Free Text Repositories in Computer-Aided Teaching of Foreign Language Terminology*. Proc. "Language Resources: Integration & Development in e-learning & in Teaching Computational Linguistics", Workshop at LREC 2004, 35-40.

Boytcheva Sv., Vitanova, I., Strupchanska, A., Yankova, M. and G. Angelova. 2004. *Towards the assessment of free learner's utterances in CALL*. Proc. "NLP and Speech Technologies in Advanced Language Learning Systems", InSTIL/ICALL Symposium, Venice,17-19 June.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R. 1990. *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, 41(6), pp. 391–407.

Dorr, B., Hendler, J., Blanksteen, S. and B. Migdaloff. 1995. *On Beyond Syntax: Use of Lexical Conceptual Structure for Intelligent Tutoring*. In (Holland, 1995), pp. 289-311.

Foltz P.W., Laham D., and Landauer T.K. 1999. *Automated essay scoring: Applications to educational technology,* In Proceedings of the ED-MEDIA Conference, Seattle.

Foltz P.W., Laham D., and Landauer T.K. 1999. *The intelligent essay assessor: Applications to educational technology*, Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2).

Gamper, J. and J. Knapp. 2002. *Review of intelligent CALL systems*. Computer Assisted Language Learning 15/4, pp. 329-342.

Holland, M., Kaplan, J. and R. Sams (eds.) 1995. *Intelligent Language Tutors: Theory Shaping Technology*. Lawrence Erlbaum Associates, Inc.

Lemaire B. and Dessus P. 2001. *A system to assess the semantic content of student essays*, J. of Educ. Computing Research, 24(3), 305–320.

Nerbonne, J. 2002. *Computer-Assisted Language Learning and Natural Language Processing*. In: R. Mitkov (Ed.) Handbook of Computational Linguistics, Oxford Univ. Press, pp. 670-698.

Ramsay, A. and H. Seville. 2000. *What did he mean by that*? Proc. Int. Conf. AIMSA-2000, Springer, LNAI 1904, pp. 199-209.

Rose, C.P., Bhembe, D., Roque, A., Siler, S., Srivastava, R. and K. van Lehn. 2002. *A hybrid language understanding approach for robust selection of tutoring goals.* . In Proc. of the Int. Conf. Intelligent Tutoring Systems, Springer, LNCS, 2363: 552-561

VanLehn, K., Jordan, P., Rose, C., Bhembe, D., Boettner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Rindenberg, M., Roque, A., Siler, A. and Srivastava, R. 2002. *The Architecture of Why2-Atlas: A Coach for Qualitative Physics Essay Writing*. In Proc. of the Int. Conf. Intelligent Tutoring Systems, Springer, Lecture Notes in CS, 2363: 158-162.

Wiemer-Hastings P. and Graesser A. 2000. *Select-a-kibitzer: A computer tool that gives meaningful feedback on student compositions*, Interactive Learning Environments, 8(2), pp. 149–169.