# Proceedings of the first Workshop on Supporting eLearning with Language Resources and Semantic Data

**May 22nd, 2010**

**Valletta, Malta**

In conjunction with LREC 2010

# Foreword

Language resources are of crucial importance not only for research and development in language and speech technology but also for eLearning applications. In addition, the increasingly availability of semantically interpreted data in the WEB 3.0 is creating a huge impact in semantic technology. Social media applications such as Delicious, Flickr, YouTube, and Facebook, provide us with data in the form of tags and interactions among users. We believe that the exploitation of semantic data (emerging both from the Semantic Web and from social media) and language resources will drive the next generation eLearning platforms. The integration of these technologies within eLearning applications should also facilitate access to learning material in developing economies.

The workshop aims at bringing together computational linguists, language resources developers, knowledge engineers, social media researchers and researchers involved in technology-enhanced learning as well as developers of eLearning material, ePublishers and eLearning practitioners. It will provide a forum for interaction among members of different research communities, and a means for attendees to increase their knowledge and understanding of the potential of language resources in eLearning. We will especially target eLearning practitioners in the Mediterranean Partner Countries.

The proceedings of the workshop contain 10 papers discussing the integration of language resources, natural language processing techniques, ontologies and social media in eLearning. The organizers hope that the selection of papers presented here will be of interest to a broad audience, and will be a starting point for further discussion and cooperation.

*Paola Monachesi, Alfio Massimiliano Gliozzo and Eline Westerhout*

# The Workshop Programme

| | |
|---|---|
| **14.30-14.45** | **Introduction** |
| | **Session on Language Resources, NLP and eLearning** |
| **14.45-15.05** | **Language resources and CALL applications**<br>*Helmer Strik, Jozef Colpaert, Joost van Doremalen and Catia Cucchiarini* |
| **15.05-15.15** | **Challenges for Discontiguous Phrase Extraction**<br>*Dale Gerdemann and Gaston Burek* |
| **15.15-15.25** | **Towards Resolving Morphological Ambiguity in Arabic Intelligent Language Tutoring Framework**<br>*Khaled Shalaan, Doaa Samy and Marwa Magdi* |
| **15.25-15.45** | **Language Resources and Visual Communication in a Deaf-Centered Multimodal E-Learning Environment: Issues to be Addressed**<br>*Elena Antinoro Pizzuto, Claudia S. Bianchini, Daniele Capuano, Gabriele Gianfreda and Paolo Rossini* |
| **15.45-15.55** | **Deaf People Education: crossing linguistic borders through e-learning**<br>*Giuseppe Nuccetelli and Maria Tagarelli De Monte* |
| **16.00- 16.30** | **Break** |
| | **Session on ontologies, social media and learning** |
| **16.30-16.50** | **BONy: a knowledge centric collaborative learning platform**<br>*Alfio Massimiliano Gliozzo, Concetto Elvio Bonafede and Aldo Gangemi* |
| **16.50-17.00** | **Social E-SPACES; socio-collaborative spaces within the virtual world ecosystem**<br>*Vanessa Camilleri and Matthew Montebello* |
| **17.00-17.20** | **A Semantic Knowledge Base for Personal Learning and Cloud Learning Environments**<br>*Alexander Mikroyannidis, Paul Lefrere and Peter Scott* |
| **17.20-17.30** | **Semantic Annotation for Semi-Automatic Positioning of the Learner**<br>*Petya Osenova and Kiril Simov* |
| **17.30-17.50** | **Facilitating cross-language retrieval and machine translation by multilingual domain ontologies**<br>*Petr Knoth, Trevor Collins, Elsa Sklavounou and Zdenek Zdrahal* |
| **18.00-19.00** | **Wrap up, discussion, plans for common projects** |

# Workshop Organisers

**Paola Monachesi**
**University of Malta, Malta**
**and Utrecht University, The Netherlands**
**P.Monachesi@uu.nl**

**Alfio Massimiliano Gliozzo**
**ISTC-CNR, Italy**
**alfio.gliozzo@istc.cnr.it**

**Eline Westerhout**
**Utrecht University, The Netherlands**
**e.n.westerhout@uu.nl**

# Workshop Programme Committee

**Claudio Baldassarre (Open University)**
**Roberto Basili (University of Rome Tor Vergata)**
**Eva Blomqvist (ISTC–CNR)**
**Antonio Branco (University of Lisbon)**
**Dan Cristea (University of Iaşi)**
**Ernesto William De Luca (TU Berlin)**
**Philippe Dessus (University Pierre-Mendès-France, Grenoble)**
**Claudio Giuliano (FBK-irst)**
**Wolfgang Greller (Open University of the Netherlands)**
**Alessio Gugliotta (Innova spa)**
**Jamil Itmazi (Palestine Ahliya University)**
**Susanne Jekat (Zürich Winterthur Hochschule)**
**Vladislav Kubon (Charles University Prague)**
**Lothar Lemnitzer (Berlin-Brandenburgische Akademie der Wissenschaften)**
**Stefanie Lindstaedt (Know-Center)**
**Angelo Marco Luccini (INSEAD)**
**Manuele Manente (JOGroup)**
**Dunja Mladenic (J. Stefan Institute)**
**Mattew Montebello (University of Malta)**
**Jad Najjar (WU Vienna)**
**Valentina Presutti (ISTC–CNR)**
**Adam Przepiorkowski (Polish Academy of Sciences)**
**Mike Rosner (University of Malta)**
**Doaa Samy (Cairo University)**
**Khaled Shaalan (Cairo University)**
**Kiril Simov (Bulgarian Academy of Sciences)**
**Stefan Trausan-Matu (University of Bucarest)**
**Cristina Vertan (University of Hamburg)**
**Fridolin Wild (Open University)**

# Table of Contents

**Foreword** ii

**Programme** iii

**Organizers** iv

**Programme Committee** v

**Table of Contents** vi

**Author Index** vii

**Language resources and CALL applications** 1
*Helmer Strik, Jozef Colpaert, Joost van Doremalen and Catia Cucchiarini*

**Challenges for Discontiguous Phrase Extraction** 7
*Dale Gerdemann and Gaston Burek*

**Towards Resolving Morphological Ambiguity in Arabic Intelligent Language Tutoring Framework** 12
*Khaled Shalaan, Doaa Samy and Marwa Magdi*

**Language Resources and Visual Communication in a Deaf-Centered Multimodal E-Learning Environment: Issues to be Addressed** 18
*Elena Antinoro Pizzuto, Claudia S. Bianchini, Daniele Capuano, Gabriele Gianfreda and Paolo Rossini*

**Deaf People Education: crossing linguistic borders through e-learning** 24
*Giuseppe Nuccetelli and Maria Tagarelli De Monte*

**BONy: a knowledge centric collaborative learning platform** 29
*Alfio Massimiliano Gliozzo, Concetto Elvio Bonafede and Aldo Gangemi*

**Social E-SPACES; socio-collaborative spaces within the virtual world ecosystem** 35
*Vanessa Camilleri and Matthew Montebello*

**A Semantic Knowledge Base for Personal Learning and Cloud Learning Environments** 40
*Alexander Mikroyannidis, Paul Lefrere and Peter Scott*

**Semantic Annotation for Semi-Automatic Positioning of the Learner** 46
*Petya Osenova and Kiril Simov*

**Facilitating cross-language retrieval and machine translation by multilingual domain ontologies** 51
*Petr Knoth, Trevor Collins, Elsa Sklavounou and Zdenek Zdrahal*

# Author Index

| | |
|---|---|
| Antinoro Pizzuto, E. | **18** |
| Bianchini, C.S. | **18** |
| Bonafede, C.E. | **29** |
| Burek, G. | **7** |
| Camilleri, V. | **35** |
| Capuano, D. | **18** |
| Collins, T. | **51** |
| Colpaert, J. | **1** |
| Cucchiarini, C. | **1** |
| van Doremalen, J. | **1** |
| Gangemi, A. | **29** |
| Gerdemann, D. | **7** |
| Gianfreda, G. | **18** |
| Gliozzo, A. M. | **29** |
| Knoth, P. | **51** |
| Lefrere, P. | **40** |
| Magdi, M. | **12** |
| Mikroyannidis, A. | **40** |
| Montebello, M. | **35** |
| Nuccetelli, G. | **24** |
| Osenova, P. | **46** |
| Rossini, P. | **18** |
| Samy, D. | **12** |
| Scott, P. | **40** |
| Shalaan, K. | **12** |
| Simov, K. | **46** |
| Sklavounou, E. | **51** |
| Strik, H. | **1** |
| Tagarelli De Monte, M. | **24** |
| Zdrahal, Z. | **51** |

# Language resources and CALL applications: speech data and speech technology in the DISCO project

**Helmer Strik [a], Jozef Colpaert [b], Joost van Doremalen [a], Catia Cucchiarini [a]**
[a] CLST, Department of Linguistics, Radboud University, Nijmegen, The Netherlands
[b] Linguapolis, Institute for Education and Information Sciences, University of Antwerp, Antwerp, Belgium
E-mail: H.Strik | J.vanDoremalen | C.Cucchiarini@let.ru.nl; Jozef.Colpaert@ua.ac.be

**Abstract**

The current paper deals with the relation between language resources and Computer Assisted Language Learning (CALL) systems: language resources are essential in the development of CALL applications, during the development of the system resources are created, and finally the CALL system itself can be used to generate additional resources that are useful for research and development of new (CALL) systems.
We focus on the system developed in the project DISCO (Development and Integration of Speech technology into COurseware for language learning): we describe the language resources employed for developing the DISCO system and present the DISCO system paying attention to the design, the automatic speech recognition modules, and the resources produced within the project. Finally, we discuss how additional language resources can be generated through the DISCO system.

## 1. Introduction

In the last few years the interest in applying Automatic Speech Recognition (ASR) technology to second language (L2) learning has been growing considerably (Eskenazi, 2009). The addition of ASR technology to Computer Assisted Language Learning (CALL) systems makes it possible to assess oral skills in a second language and to provide corrective feedback automatically. The latter feature appears particularly appealing, since research has shown that usage-based acquisition in the L2 is not as successful as in the L1 (Ellis and Larsen-Freeman, 2006: 571), that L2 learners have difficulty identifying their own errors (Dlaska and Krekeler, 2008), and that they indeed need guidance to improve their language skills (Ellis and Bogart, 2007). Since providing practice and feedback for speaking proficiency is particularly time-consuming, the necessary amount of practice is almost never achieved in traditional teacher-fronted lessons. Against this background, ASR-based CALL systems would seem to make for an interesting supplement to traditional L2 classes.

However, developing ASR-based CALL systems that can provide accurate and useful feedback on oral proficiency is not trivial, because the speech of L2 learners poses special difficulties to ASR technology (Compernolle 2001; Benzeghiba et al. 2007; Doremalen et al. 2009a; Doremalen et al. 2009b). In addition, existing systems in general fail to provide corrective feedback that is detailed enough and accurate, especially on L2 pronunciation which is considered a particularly challenging skill, both for L2 learners (Flege, 1995) and CALL systems (Menzel et al. 2000: 54; Morton and Jack, 2005).

Another problem that has hampered the realization of ASR-based CALL systems, especially for the smaller languages, is that although companies, esp. publishers, are willing to use the technology, many companies do not have the means to finance the development of such technology. For these and other reasons, in the Netherlands and Flanders a programme was started, called STEVIN (a Dutch acronym that stands for Essential Language Resources in Dutch), which is funded by the Flemish and Dutch governments and aims at stimulating the development of basic language and speech technology for the Dutch language.

Within the framework of the STEVIN programme a project called DISCO (Development and Integration of Speech technology into COurseware for language learning, http://lands.let.kun.nl/~strik/research/DISCO) was started that aims at developing a prototype of an ASR-based CALL system for practicing oral skills in Dutch L2. The system addresses different aspects of speaking proficiency (syntax, morphology and phonology), detects errors in speaking performance, points them out to the learners and gives them the opportunity to try again until they manage to produce the correct form.

One of the interesting things about this project is that since it is carried within the STEVIN programme, the technology that is developed for the present project will be publicly made available to interested users (researchers, HLT companies and publishers) through the Dutch HLT Agency.

In the current paper we discuss the relation between language resources and CALL systems: language resources are essential in the development of CALL applications, during R&D resources are created, and finally the CALL system itself can be used to generate additional resources that are useful for research and development of new (CALL) systems.

In section 2 we describe which language resources were employed in the DISCO project. In section 3 we present the DISCO system paying attention to the design, the automatic speech recognition modules, some preliminary results and the resources produced within the project. In section 4 we discuss how additional language resources can be generated through the DISCO system.

## 2. CALL applications and the need for language resources

An important requirement for developing ASR-based CALL applications is the availability of language resources such as language and speech corpora and speech technology toolkits.

In order to develop technology that is able to identify errors in oral proficiency we need to know which errors are made by L2 learners in the first place. Part of this information can be found in the literature, but, in general, the information provided in the literature is not complete and not sufficiently quantified to be suitable for developing CALL applications.

In our previous research on developing a computer assisted pronunciation training (CAPT) for Dutch, Dutch-CAPT (Cucchiarini et al., 2009), we needed to draw up an inventory of pronunciation errors. We discovered that the information on L2 errors provided in the literature was mostly based on observational studies, was often incomplete, and not quantitative in nature. For this reason we had no other choice than conducting L2 error studies ourselves (Neri et al., 2006). However, since a speech corpus of non-native Dutch was not available at the time, we had to resort to the auditory analysis of Dutch L2 speech recordings that had been collected in the framework of previous projects (Neri et al., 2006).

For the DISCO project we had the opportunity of using the results of another STEVIN project that had been completed in the meantime, the JASMIN corpus (Cucchiarini et al., 2008).

### 2.2.1. The JASMIN speech corpus

The JASMIN corpus is an extension of the large Spoken Dutch Corpus (CGN; Oostdijk, 2002). JASMIN contains speech by children of different age groups, elderly people and non-natives with different mother tongues. The JASMIN corpus was collected in the Netherlands and Flanders and is specifically aimed at facilitating the development of speech-based applications for children, non-natives and elderly people. In the case of non-native speakers the applications envisaged were especially language learning applications because there is considerable demand for CALL products that can help making Dutch L2 teaching more efficient.

In selecting the non-native speakers for this corpus, mother tongue constituted an important variable. For the Flemish part, Francophone speakers were selected because they form a significant proportion of the Dutch learning population. In the Netherlands, on the other hand, a miscellaneous group of L2 learners with various mother tongues was selected because this more realistically reflects the situation in Dutch L2 classes.

Since an important aim in collecting non-native speech material was that of developing language learning applications for education in Dutch L2, various experts were consulted to determine for which proficiency level such applications are most needed. It turned out that for the lowest levels of the Common European Framework (CEF), namely A1, A2 or B1, there is relatively little material and that ASR-based applications would be very welcome. For this reason, speech from adult Dutch L2 learners at these lower proficiency levels was recorded.

The speech collected in the JASMIN corpus was recorded in two different modalities: about 50% of the material consists of read speech material while the other 50% is made up of extemporaneous speech produced in human-machine dialogues. The JASMIN dialogues were collected through a Wizard-of-Oz-based platform and were designed such that the wizard was in control of the dialogue and could intervene when necessary. In addition, recognition errors were simulated and difficult questions were asked to elicit some typical phenomena of human-machine interaction that are known to be problematic in the development of spoken dialogue systems, such as hyperarticulation, restarts, filled pauses, self talk and repetitions.

The speech recordings were annotated at different levels. For the DISCO project, the verbatim transcription and the automatically generated phonemic transcription are particularly relevant.

For all the reasons mentioned above the JASMIN speech material turned out to be extremely useful and appropriate for the development of the DISCO system.

Both read and extemporaneous speech were analyzed to study which errors are made at the level of pronunciation, morphology and syntax. For this purpose the annotations contained in JASMIN were supplemented with extra annotations of the morphological and syntactical errors made by the speakers. The automatically generated phonemic transcriptions were manually verified by trained students and where necessary improved. Subsequently they were used to study which pronunciation errors are made by L2 learners of Dutch with different mother tongues.

The human-machine dialogues were used for conducting experiments for the DISCO system because they closely resemble the situation we will encounter in this CALL application.

### 2.2.2. The SPRAAK speech recognizer

The speech recognizer adopted in the DISCO project is SPRAAK (Demuynck et al., 2008), a hidden Markov model (HMM)-based ASR package developed for over 15 years by ESAT at the University of Leuven and later enriched with knowledge and code from other partners through the STEVIN project SPRAAK. The availability of a speech recognition system for Dutch was considered to be an important requirement by the whole language and speech technology (LST) community in the Netherlands and Flanders. For this reason a project was started within the STEVIN programme for this specific purpose: the SPRAAK project. The aim of SPRAAK was twofold: a) developing a highly modular toolkit for research into speech recognition algorithms and b) providing a state-of-the art recogniser for Dutch with a simple interface that could be used by non-specialists. SPRAAK is distributed as open source for academic usage and at moderate cost for commercial exploitation (for further details, see http://www.spraak.org/).

# 3. The DISCO system

## 3.1 Design of the DISCO system

Within the STEVIN programme a project called DISCO was started on 01-02-2008, in which a CALL system will be developed. The target user group for the DISCO system are immigrants who want to learn Dutch as L2 to be able to work in the Netherlands or Flanders.

The model adopted for designing the system is Distributed Language Learning (DLL), a methodological and conceptual framework for designing competency-oriented and effective language education (Colpaert, 2004). Its starting point is the design of a language learning environment for a specific language learning situation. The design is based on a thorough analysis of all factors and actors in the language learning situation, and on the identification of aspects amenable to change or improvement. The main phases of the design are goal-oriented conceptualization and ontological specification. Goal-oriented conceptualization stands for the formulation of a solution based on the realization of 'practical goals' as a hypothetical compromise between (often conflicting) personal and pedagogical goals, both for teachers and learners. Ontological specification is a detailed description of the architecture of the language learning environment, defined as the network of interactions between learner, co-learner, teacher, content, native, etc. inside or outside the learning place.

In DISCO, we limit our general design space to closed response conversation simulation courseware and interactive participatory drama (IPD), a genre in which learners play an active role in a pre-programmed scenario by interacting with computerized characters or "agents". The use of drama is beneficial for various reasons:

1. it "reduces inhibition, increases spontaneity, and enhances motivation, self-esteem and empathy" (Hubbard, 2002),
2. it casts language in a social context, and
3. its notion implies a form of planning, scenario-writing and fixed roles, which is consistent with the limitations we set for the role of speech technology in DISCO.

To summarize, this framework allows us to create a rich and communicative CALL application that stimulates Dutch L2 learners to produce speech and experience the social context. On the other hand, these choices are appropriate from a technological perspective, since they make it possible to successfully deploy speech technology while taking into account its limitations (Strik et al., 2009).

To gain more insight into appropriate feedback strategies, pedagogical goals, and personal goals a number of preparatory studies were carried out, such as exploratory in-depth interviews with Dutch L2 teachers and experts, focus group discussions to elicit the personal goals of learners, and pilot studies through partial systems with limited functionality (e.g. no speech technology). The functions of the system that were not implemented (play prompts, give feedback, etc.) were simulated. The results of these preparatory studies were taken into account in finalizing the design of the DISCO system.

The learning process starts with a relatively free conversation simulation, taking well into account what is (not) possible with speech technology: learners are given the opportunity to choose from a number of prompts at every turn (branching, decision tree). Based on the errors they make in this conversation they will be offered remedial exercises, which are very specific exercises with little freedom.

Feedback depends on individual learning preferences: the default feedback strategy is immediate corrective feedback, which is visually implemented through highlighting, and from an interaction perspective by putting the conversation on hold and focusing on the mistakes. Learners that wish to have more conversational freedom can choose to receive communicative recasts as feedback, which let the conversation go on while highlighting mistakes for a short period of time.

The final system will have several parameters that can be changed by the learner or teacher. During development and implementation, we will try to have these parameters behave intelligently (based on error analysis and learner behavior), so that the system can adapt itself to the learner. For future research these parameters offer the possibility of studying different modes of behavior of the CALL system and their effect on language learners.

## 3.2 The speech recognition modules

First, we provide some technical details about our system. As mentioned above, the human-machine dialogues were

used for conducting experiments for the DISCO system. The material used consisted of speech from 45 speakers who each give answers to 39 questions about a journey.
The input speech, sampled at 16kHz, is divided into overlapping 32ms Hamming windows with a 10ms shift and pre-emphasis factor of 0.95. 12 Mel-frequency cepstral coefficients (MFCCs: C1-C12) plus C0 (energy), and their first and second order derivatives were calculated and cepstral mean subtraction (CMS) was applied. The constrained language models and pronunciation lexicons are implemented as finite state machines (FSM).

In the DISCO system feedback on speaking performance is given on three levels: syntax, morphology and phonology. To give feedback, errors on these levels have to be detected automatically. In our system architecture, this task is divided in two modules: (1) the speech recognition module and (2) the error detection module. The first module, speech recognition, determines the sequence of words the student uttered. For each prompt a list of predicted correct and (grammatically) incorrect responses is created beforehand based on errors that are expected on empiric grounds. This list is the basis for a Finite State Grammar (FSG) language model, which is used by an hidden Markov model (HMM)-based speech recognition system. The recognition system is forced to choose among the predicted response from the list.

To avoid false accepts, for example when an utterance is uttered that is not in the list of predicted responses, utterance verification (UV) is carried out. Using a combination of acoustic and durational similarity measures it is determined whether the response chosen by the speech recognizer reflects what has been said. If it is rejected the user is asked to try again; if it is accepted, the system will proceed to error detection (Van Doremalen et al. 2009a, b).

Note that once the chosen response is accepted by the utterance verifier we can already detect errors on the syntactic level because the system is confident enough that the student uttered a specific sequence of words and it also knows what the student was supposed to say.

Detecting errors on the morphological and phonological levels requires another, more detailed analysis of the speech signal. The starting point of this analysis is a segmentation of the speech signal into a sequence of phones obtained from the speech recognition module. Using a variety of spectral and temporal features a confidence measure (CM) is calculated for each of these phones. Based on this CM the system decides to mark the hypothesized phone in the segmentation as correctly pronounced or incorrectly pronounced (Van Doremalen et al. 2009c).

In the way described above, phonological errors can be detected. Since some phonemes are critical for certain morphological constructions, the approach used for detecting phonological errors will be used also for detecting some of the morphological errors, for instance those concerning regular verb forms. Irregular verbs, on the other hand, may require an approach that is more similar to that adopted for detecting syntactic errors. Once the system arrives at this final stage, the system has a detailed overview of all the errors on the different levels and based on this overview the system can provide feedback to the student.

### 3.3 The resources produced in the project

The resources mentioned above are employed to develop the DISCO system which consists of various parts. First of all, a blue-print of the design and the speech technology modules for recognition (i.e. for selecting an utterance from the predicted list, and verifying the selected utterance) and for error detection (errors in pronunciation, morphology, and syntax). In addition, the following resources have been developed: an inventory of errors at all these three levels, a prototype of the DISCO system with content, specifications for exercises and feedback strategies, and a list of predicted correct and incorrect utterances.

The fact that DISCO is being carried out within the STEVIN programme implies that its results, all the resources mentioned above, will become available for research and development through the Dutch Flemish Human Language Technology (HLT) Agency (TST-Centrale; www.inl.nl/tst-centrale). This makes it possible to reuse these resources for conducting research and for developing specific applications for ASR-based language learning.

### 3.4 Evaluation

A system that gives meaningful feedback must operate in a manner that is similar to what a competent teacher would do. Therefore, for the final evaluation of the whole system we intend to use a design in which different groups of students of Dutch as a second language (DL2) at the University of Antwerp and at the Radboud University in Nijmegen use the system and fill in a questionnaire with which we can measure the students' satisfaction in working with the system.
Teachers of DL2 will then assess all sets of system prompt, student response and system feedback for the quality of the feedback on the level of pronunciation, morphology and syntax. For this purpose, recordings will be made of students who complete the exercises developed to test the DISCO system.
Given the evaluation design sketched above, we consider the project successful from a scientific point of view if the DL2 teachers agree that the system behaves in a way that makes it as useful for the students as a teacher is, and if the students rate the system positively on its most important aspects.

## 4. Generating additional language resources

Above we described which resources we used in developing our CALL system, and which resources become available during development of the system. In this section, we describe which additional resources can be collected by using the CALL system.

After the CALL system has been developed, language learners can use it to practice oral skills. The system has been designed and developed in such a way that it is possible to log details regarding the interactions with the users. This logbook can contain, e.g., the following information: what appeared on the screen, how the user responded, how long the user waited, what was done (speak an utterance, move the mouse and click on an item, use the keyboard, etc.), the feedback provided by the system, how the user reacted on this feedback (listen to example (or not), try again, ask for additional, e.g. meta-linguistic, feedback, etc.).

Finally, all the utterances spoken by the users can be recorded in such a way that it is possible to know exactly in which context the utterance was spoken, i.e. it can be related to all the information in the logbook mentioned above. An ASR-based CALL system, like DISCO, can thus be used for acquiring additional non-native speech data, for extending already existing corpora like JASMIN, or for creating new ones. This could be done within the framework of already ongoing research without necessarily having to start corpus collection projects.

Such a corpus and the log-files can be useful for various purposes: for research on language acquisition and second language learning, studying the effect of various types of feedback, research on various aspects of man-machine interaction, and of course for developing new, improved CALL systems. Such a CALL system will also make it possible to create research conditions that were hitherto impossible to create, thus opening up possibilities for new lines of research.

For instance, at the moment a project is being carried out at the Radboud University of Nijmegen, which is aimed at studying the impact of corrective feedback on the acquisition of syntax in oral proficiency (http://lands.let.kun.nl/~strik/research/FASOP). Within this project the availability of an ASR-based CALL system makes it possible to study how corrective feedback on oral skills is processed on-line, whether it leads to uptake in the short term and to actual acquisition in the long term.

This has several advantages compared to other studies that were necessarily limited to investigating interaction in the written modality: the learner's oral production can be assessed on line, corrective feedback can be provided immediately under near-optimal conditions, all interactions between learner and system can be logged so that data on input, output and feedback are readily available for research.

## 5. Conclusions

In this paper we have discussed the importance of language resources for CALL application development on the basis of our experiences in the DISCO project in which speech data and speech technology are employed to develop a system for practicing oral skills in a second language.. We have seen that language resources are actually indispensable for developing sound CALL applications. Once developed, such applications can also be employed to produce new valuable language resources which can in turn be used to develop new, improved CALL systems.

## 6. Acknowledgements

The DISCO project is carried out within the STEVIN programme funded by the Dutch and Flemish Governments (http://taalunieversum.org/taal/technologie/stevin/).

## 7. References

Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C. (2007). Automatic speech recognition and speech variability: a review. *Speech Communication*, 49, 763–786.

Colpaert, J. (2004). Design of Online Interactive Language Courseware: Conceptualization, Specification and Prototyping. Research into the impact of linguistic-didactic functionality on software architecture. (Doctoral dissertation). University of Antwerp, 2004.

Cucchiarini, C., Neri, A., and Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication*, 51, 853-863.

Cucchiarini, C., Driesen, J., Van hamme, H., and Sanders, E., (2008). Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. In *Proceedings of LREC-2008*.

Demuynck, K., Roelens, J., van Compernolle, D., and Wambacq, P. (2008) SPRAAK: an open source SPeech Recognition and Automatic Annotation Kit. In *Proceedings of ICSLP-2008*, p. 495.

Dlaska, A. and Krekeler, C. (2008). Self-assessment of pronunciation. *System*, 36, pp. 506-516.

Ellis, N.C., Bogart, P.S.H. (2007). Speech and Language Technology in Education: the perspective from SLA research and practice. In *Proc. SLaTE*, Farmington PA, pp. 1-8.

Ellis, N. and Larsen-Freeman, D. (2006). Language emergence: implications for applied. *Linguistics, Applied Linguistics*, 27.4: 558–89.

Eskenazi, M. (2009). An overview of Spoken Language Technology for Education, *Speech Communication*.

Flege, J. (1995). Second language speech learning: theory, findings and problems. In W. Strange (Ed.) *Speech perception and linguistic experience*, Baltimore: York Press, pp. 233-272.

Hubbard, P. (2002). Interactive Participatory Dramas for Language Learning. *Simulation and Gaming*, vol. 33, pp. 210-216.

Morton, H., Jack, M. (2005). Scenario-Based Spoken Interaction with Virtual Agents. *Computer Assisted Language Learning*, 18, 171-191.

Oostdijk, N. (2002). The design of the spoken dutch corpus. In N*ew Frontiers of Corpus Research*, P. Peters, P. Collins, and A. Smith, Eds. Rodopi, pp. 105–112.

H. Strik, Cornillie, F., van Doremalen, J., Cucchiarini, C. (2009). Developing a CALL System for Practicing Oral Proficiency: How to Match Design and Speech Technology. In *Proc. SLATE*, Wroxall Abbey.

Van Compernolle, D. (2001). Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communiciation*, 35, 71-79.

Van Doremalen, J., Cucchiarini, C., Strik, H. (2009a). Optimizing automatic speech recognition for low-proficient non-native speakers. Accepted for publication in *EURASIP Journal on Audio, Speech, and Music Processing*, to appear.

Van Doremalen, J., Strik, H., Cucchiarini, C. (2009b). Utterance Verification in Language Learning Applications. In *Proc. SLATE*, Wroxall Abbey.

Van Doremalen, J., Cucchiarini, C., Strik, H. (2009c). Automatic Detection of Vowel Pronunciation Errors Using Multiple Information Sources. *Proceedings of the biannual IEEE workshop on Automatic Speech Recognition and Understanding (ASRU).*

# Challenges for Discontiguous Phrase Extraction

**Dale Gerdemann, Gaston Burek**

Dept. Linguistics
University of Tübingen
dg@sfs.uni-tuebingen.de, gaston.burek@gmail.com

### Abstract

Suggestions are made as to how phrase extraction algorithms should be adapted to handle gapped phrases. Such variable phrases are useful for many purposes, including the characterization of learner texts. The basic problem is that there is a combinatorial explosion of such phrases. Any reasonable program must start by putting the exponentially many phrases into equivalence classes (Yamamoto and Church, 2001). This paper discusses the proper characterization of gappy phrases and sketches a suffix-array algorithm for discovering these phrases.

## 1. Introduction

Writing is an essential part of learning and evaluating written texts is an essential part of teaching. A good teacher must attempt to understand the ideas presented in a learner text and evaluate whether or not these ideas make sense. Such evaluation can obviously not be performed by a computer. But on the other hand, computers are good at evaluating other aspects of texts. Computers are, for example, very good at picking out patterns of linguistic usage, in particular terms and phrases[1] that are used repeatedly. It is often the case that choice of terminology can be surprisingly effective in characterizing texts. For example, the terms "Latent Semantic Analysis" and "Latent Semantic Indexing" mean essentially the same thing, but the former is more characteristic of the educational and psychological communities whereas the latter is more characteristic of the information retrieval community. In a similar vein, Biber (2009) uses characteristic phrases to distinguish between written and spoken English. Up to now, in the eLearning community, bag-of-words based approaches have been most popular for evaluating student essays (Landauer and Dumais, 1997). It is the contention of this paper that the next step of considering phrases will not be possible until eLearning practitioners immerse themselves into the somewhat technical combinatorial pattern matching literature.

This paper is concerned with extracting phrases with gaps. This is an important topic since many phrases occur in alternative forms. For example, the English phrase *one and the same* has an essentially verbatim counterpart in Bulgarian, but the Bulgarian phrase occurs in a variety of forms depending on gender and number of the following noun. The following forms were extracted from a few Bulgarian texts: един и същи, една и съща, едно и също, едни и същи. In this simple Bulgarian phrase, there are three different alternations. First един ('one') occurs with inflections $-\emptyset$, -а, -о and -и. Second, същ- ('same') occurs with inflections -и, -а, and -о. And third, един also contains the "fleeting" or "ghost" vowel и, which alternates with $\emptyset$.[2] If we consider this Bulgarian expression as a sequence of letters. Then the inflection on един is in the middle, whereas the inflection on същ- is on the right periphery. Both of these instances of variation are problematic. The variation in the middle, however, is somewhat more problematic, and is the main focus of this paper.

Most phrase extraction programs are based on pattern matching algorithms developed for computational molecular biology. To adapt such algorithms for natural language, with worst case examples such as the Bulgarian phrase above will require a great deal of thought. In particular, cooperation between language researchers and computer scientists is required. Too often language researchers use off-the-shelf software packages, and apply no particular programming skills at all.[3] Hence, the goal of the present paper is not to present a new algorithm for gapped phrase extraction, but rather to present some features of what such a phrase extraction program ought to provide. Some technical literature is presented, but the intended readership of this paper is non-technical.

### 1.1. Algorithmic Introduction

Efficient algorithms for phrase (or n-gram) extraction were introduced into the computational linguistics literature by Yamamoto and Church (2001) and have subsequently been used for a wide variety of applications such as lexicography, phrase-based machine translation and bag-of-phrases based text categorization (Burek and Gerdemann, 2009).[4] Ultimately, the goal of such algorithms is to discover repetitive structure as represented by frequently recurring sequences of symbols. Unfortunately, the approach of Yamamoto and Church often misses repetitive structure since phrases often occur with slight variations. For example, the middle term of a phrase might occur in different morphological variants:

[1] We use the term "phrase" to mean repeated sequence of tokens. This is quite flexible, allowing any kind of tokenizer and phrases of any non-negative length.

[2] Ghost vowels are a characteristic of Bulgarian and Slavic languages in general (Jetchev, 1997). The vowel и (IPA: /i/) is, however, idiosyncratic as a ghost vowel.

[3] For language researchers wishing to acquire some programming skills, there is probably no better starting point than Sedgewick and Wayne (2010 forthcoming).

[4] Similar algorithms are also used by Dickinson and Meurers (2005) for detecting inconsistencies in annotated corpora. This is particularly relevant, since they are specifically interested in discontinuous (or gapped) annotations.

*all join in* vs *all joined in*; or the middle term may vary in other ways: *give me a* vs *give him a*.

Recently, an algorithm for finding such paired repeats was presented by Apostolico and Satta (2009). This algorithm is quite efficient, as it is shown to run in linear time with respect to the output size. Unfortunately, however, the algorithm is designed to extract "tandem repeats," which are defined in a way that may not be entirely appropriate for the researcher interested in extracting gapped phrasal expressions. The goal of this paper is, then, to specify the requirements of such researchers. The hope is that this paper will provide a challenge for algorithm designers who may either want to adapt the Apostolico and Satta algorithm or design a new competing algorithm.

One difference between the Yamamoto-Church algorithm and the Apostolico-Satta algorithm is the former is based on suffix arrays, whereas the latter is based on suffix trees. This should, however, not be seen as a major distinction, since recent developments with suffix arrays have tended to blur the distinction (Abouelhoda et al., 2004; Kim et al., 2008).[5] To some extent, one may think of suffix arrays simply as a data structure for implementing suffix trees. Further implementation issues will be discussed below.

## 2. Some Terminology

To start with, let us consider a typical gapped expression: *from one X to the other*.[6] The goal of gapped phrase extraction is to discover gapped expressions such as this. Once such a pattern is discovered, a researcher can easily find further instances of the pattern by searching with regular expressions in other corpora. Initially however, the phrase extraction may discover just a couple of instantiations for *X*, which may be expressed as a simple regular expression using only alternation: $from one [shore|edge] to the other$. In referring to patterns such as this, we will use $\alpha$ to refer to the left part $from\ one$ and $\beta$ to refer to the right part $to\ the\ other$. It will generally be assumed that the left and right parts are non-empty. For the alternation in the middle, We will use the letter $m$. It will generally be assumed that the middle consists of at least two alternatives.

As usual, we will use letters from the beginning of the alphabet $a, b, c$ to represent single symbols, and letters from the end of the alphabet $w, x, y$ to represent sequences. The reader should keep in mind, however, that what counts as a symbol depends on the tokenization. The two obvious approaches are character-based and word-based tokenization, with the latter in particular requiring algorithms adapted to a large alphabet. In some sense, word-based tokenization is more natural, though the character-based approach has the advantage of avoiding some difficult problems such as compound nouns in German and word segmentation in Chinese Zhang and Lee (2006). In this paper, we assume that some tokenization (and also possibly normalization) is performed on the corpus, and that tokens are replaced by integers.

## 3. Desiderata

We now present a rather incomplete list of desirable features for gapped phrase extraction.

### 3.1. Main Parameters

By default an extracted gapped phrase $\alpha m \beta$ should have $|\alpha| \geq 1$, $|\beta| \geq 1$ and $m = [a_1 | \dots | a_n]$ where $n \geq 2$. These are minimal values, and may be set to larger values to extract possibly more interesting phrases. If the length of $\alpha$ or $\beta$ is set to 0, then the gap will be on the periphery.

The length of $\alpha$ may also be seen as an efficiency consideration. The central idea of the Apostolico and Satta algorithm, for example, picks out candidate left parts first, and then for each of these, a recursive call is made to find a corresponding right part.[7] Putting a length restriction on $\alpha$ means that there are fewer candidates, and therefore fewer recursive calls. Clearly, an alternative approach would be to start with the right piece and recursively search for corresponding left pieces.

### 3.2. Conditions on the Gap

A language researcher studying gapped phrases may find a gap of length 4 interesting (***from one*** *end of the Earth* ***to the other***) but a gap of length 7 uninteresting (*Medical bills* ***from one*** *puppy catching something and passing it on* ***to the other*** *puppy*). With character-based tokenization, however, a gap of length 6 or more may well be interesting: $and\ half - [believ|form|melt|slouch]ed$.[8]

In addition to specifying the maximum length of the gap, it may be desirable to be able to specify a minimum length. An alternation like $b[|o]ut$ for 'boat' and 'but' seems particularly perverse, though perhaps there are other ways to filter out such uninteresting cases. Biber (2009) limits the gap to be of length exactly one. But this seems to merely reflect the limitations of a particular software package since in the context ***from one*** *X* ***to the other***, there is very little difference between the single word 'extreme' and the four word phrase 'end of the Earth'. It may also be possible for the gap to have negative length, effectively meaning that the left and right parts overlap. This is allowed, for example, in the Apostolico-Satta algorithm, though it is unclear what advantages this "feature" has for natural language texts.[9]

[5] Kim et al. (2008) is of particular interest for NLP, since their approach is optimized for a large alphabet, as opposed to most of the bioinformatics literature which uses a four-letter alphabet. With a large alphabet, it becomes possible to tokenize a text by words, and treat each word as a "letter."

[6] Perhaps eLearning practitioners who are interested in ontologies will find this example interesting. There is clearly a class of "polarized entities" that can serve as good instantiations for X. Paired, but non-polarized entities like *sock* and *shoe* are not very felicitous. Is there a WordNet synset for this?

[7] We're simplifying quite a bit here. The "recursive call" is, in fact, rather different from the original call.

[8] This pattern is found in Moby Dick. A language researcher might be interested in such an example since it seems to pick out a semantic class of actions that occur or can be performed in a partial manner.

[9] In fact, the the Apostolico-Satta algorithm has a parameter $d$ not for the length of the gap, but rather for the maximum distance between the beginning of the left part and the beginning of the right part. If $d < |\alpha|$, then there could be overlap. This, however, does not seem to be a serious limitation, since it would be easy enough to adapt the Apostolico-Satta allgorithm to let $d$ be some function of $|\alpha|$.

More sophisticated possibilities also exist. For example, one could specify the the gap length conditions as a function of the lengths of the left and right pieces. Or perhaps a function of the contents of the left and right parts and the gap could be used. Another possibility would be to measure the gap length as number of syllables or number of some other kind of linguistic unit. Probably, it would not be possible to incorporate such conditions directly into the extraction algorithm. Most likely, a secondary filter would be the required approach.

### 3.3. Principle of Maximal Extension

A fundamental notion in the pattern recognition literature is that of *saturation*, which Apostolico (2009) defines as follows:

> . . . a pattern is *saturated* relative to its subject text, if it cannot be made more specific without losing some of its occurrences.

This is stated in a rather imprecise way, but the intention should be clear. Suppose that the pattern *mumbo* has occurrences at $(i, i)$, $(j, j)$ and $(k, k)$. Suppose further that the pattern is extended (made more specific) to *mumbo jumbo* and that occurrences are now found at $(i, i + 1)$, $(j, j + 1)$ and $(k, k + 1)$. Then the 3 old occurrences should not be seen as lost, but rather as replaced by 3 corresponding longer occurrences. So the pattern for the incomplete phrase *mumbo* is unsaturated.

Suffix trees and suffix arrays are a kind of asymmetrical data structure that make extensions to the right easier to find than extensions to the left. So given *mumbo*, it is easy to extend this to the right, but given *jumbo*, it is much harder to extend this to the left. For left extensions, Abouelhoda et al. (2004) advocate the use of a Burrows and Wheeler transformation table.

For gapped phrases, the issue of extension to the left and right becomes even more complex. Given a pattern $\alpha[ax_1 \mid \cdots \mid ax_n]\beta$, it seems reasonable to extract the $a$, turning the pattern into $\alpha a[x_1 \mid \cdots \mid x_n]\beta$, capturing the generalization that the middle part always starts with $a$.

If the left and right parts are both extended, then one can find patterns like $Ahab\ r[each|emain|etir|ush]ed$ (from Moby Dick), where extension of the left part represents the linguistically interesting fact that all the verbs are in the past tense. The extension of the left part, on the other hand, captures the rather uninteresting fact that all the verbs happen to start with *r*. If the left part is now further extended, then the pattern becomes more specific, and loses some of its occurrences: $Ahab\ re[ach|main|tir]ed$. It is unclear how a gapped phrase extraction program should be designed to rule out such uninteresting extensions.[10]

It is interesting to think about the example in the previous paragraph in terms of *saturation*. Suppose we think of the patterns as $Ahab\, r \ldots ed$ and $Ahab\, re \ldots ed$. That is, think of the middle part as not really part of the pattern, but rather as providing information about occurrences of the pattern. In this sense, $Ahab\ re \ldots ed$ appears to be more specific, since the occurrence with *rushed* is lost. But there is a problem here. Recall that the . . . matches sequences no longer than length $d$. If we set $d$ to be 4, then the supposedly less specific pattern will not match *Ahab remained*, and the supposedly more specific pattern will match this occurrence. This suggests that the Apostolico-Satta approach of letting $d$ be the distance from the beginning of the left piece to the beginning of the right piece may be preferable. On the other hand, their approach allows the left and right parts to overlap.

### 3.4. No Overlap

The Apoostolico-Satta algorithm is designed to find *tandem* occurrences of two strings, which they explain as follows:

> By the two strings occurring in tandem, we mean that there is no intermediate occurrence of either one in between.

To illustrate the problem of intermediate occurrences, consider the following truncated version of Moby Dick (tokenized by character):

the␣boat.␣the␣white␣whale

The sequence *the␣* occurs twice, so this is a candidate left part. The sequence *wh* occurs twice, both times with *the␣* to the left (supposing $d = 6$, for example). So without taking care, one might extract the nonsense pattern $the\ [|\ white]\ wh$.

The Apostolico-Satta algorithm is designed from the beginning to rule out such overlaps. But the basic algorithm presented in section 4. has a problem with these. An extra step would be required just to filter out such overlaps.

### 3.5. Boundaries

A common feature in the study of (gapped) phrases is that they are allowed to cross many, but not all kinds, of boundaries. For example, in the "lexical bundles" studied by Biber (2009) is that they, more often than not, cross the category boundaries of traditional linguistics. Typical examples are: *as a result of* and *it is possible to*. With tokenizing by letter, one often finds partial words (example from Moby Dick): $contrast\ [between|in|of|to]\ th$. Here the partial word *th* seems to play an important role in English.

Still there are some boundaries that should not be crossed. Dickinson and Meurers (2005), for example, note that the patterns that they were looking for should not cross sentence boundaries. There is therefore a temptation to put such boundary constraints into the phrase extraction program. We believe, however, that this is a mistake. The phrase extraction program is already complicated enough without having to deal with such special cases.

In this case there seems to be a fairly simple-minded alternative. Simply use a tokenizer that replaces each boundary punctuation character (period, question mark, etc) with a unique integer identifier. This requires a bit of bookkeeping to remember which integers have been used to represent

[10] On a personal note, it is examples like this that inspired us to write this paper. We had started off by implementing an algorithm similar to that of Apostolico and Satta (2009), and after encountering problematic cases like this, decided to put the algorithm aside for a while, and to concentrate on writing a specification of desirable features for any gapped phrase extraction program.

which punctuation characters, but it is still much easier than modifying the suffix arrays or trees. A similar approach is described in section 4. to avoid extraction of "phrases" which start near the end of one text in the corpus, and conclude near the beginning of the next text.

### 3.6. Interesting Phrases

To be useful, a phrase extraction program must be equipped with a notion of what kinds of phrases are interesting. Citing Apostolico (2009):

> Irrespective of the particular model or representation chosen, the tenet of pattern discovery equates overrepresentation with surprise, and hence with interest.

In linguistics, there are other ways of defining *interest*. For example, a phrase may be considered interesting if it exhibits some degree of non-compositional semantics, or if it exhibits some particular syntactic pattern. For an overview, see Evert (2009).

Another way of measuring interest is more goal directed. One might say, for example, that a phrase is interesting if it is useful for distinguishing positive camera reviews from negative ones (Tchalakova, 2010). Or alternatively, a phrase could be considered interesting if it is helpful for distinguishing high quality online posts from low quality ones (Burek and Gerdemann, 2009).

A central insight of (Yamamoto and Church, 2001) is that measures of interest are most commonly based upon basic measures of term frequency and document frequency, and that these measures need only be calculated for the saturated phrases.[11][12] So, for example, the term frequency and document frequency for *mumbo* is exactly the same as for *mumbo jumbo*, so this information can be stored just once at the appropriate node in a suffix tree or for an lcp-interval in a suffix array. The problem is, of course, that *jumbo* really ought to be included in this class as well, and neither suffix trees nor suffix arrays provide a natural way of representing such equivalence classes.

A key question to answer is how the interest measure should be incorporated into the gapped phrase extraction algorithm. The simplest approach would be to extract phrases initially without regard to interest, and then use the interest measure as a filter to remove uninteresting cases. Another approach would be to incorporate the interest measure into the algorithm, perhaps by restricting candidate left parts to just the interesting cases before looking for matching right contexts. We leave this as an open question.

## 4. Algorithmic Specifications

In this section, we sketch a rather basic algorithm which may serve as the basis for something more useful.[13] The idea is quite simple. Given a phrase extraction algorithm for non-gapped phrases, candidate left parts can be extracted. To reduce the search space, these candidate left parts may be required to be maximally extended or "interesting" in various ways. For a given phrase $p$, find all occurrences of $p$ in the corpus, and denote each such occurrence as $(i, j)$, where $i$ and $j$ are the indices of the first and last tokens of the occurrence in the corpus. For each such occurrence, specify the right context as $(j + 1, j + d + 1)$, where $d$ is the maximal length allowed for the gap. Clearly, these right contexts can be found efficiently using either suffix trees or suffix arrays. Now form a new corpus by treating each of these right contexts as a single text in this subcorpus. Following the idea of Yamamoto and Church (2001), the texts in this subcorpus should be concatenated, using sentinels to separate one text from the next, and also with one sentinel at the end. Assuming that the text is represented by integer id's, then the smallest otherwise unused integers can be used for the sentinels.

Assuming that a subcorpus is built up in this way, then finding right parts corresponding to each left part is mostly just a matter of running the phrase extraction program again for each subcorpus. There are, however a couple of issues to watch out for. First,pp it is important that a different integer is used for each sentinel. Otherwise the sentinels themselves, including possibly context around the sentinels, will be seen as repeated phrases.

Second, there is a problem with limiting the right context to be of length $d + 1$. If the gap is of length $d$, then the right context is just long enough to include one token from the right part. Consider, for example the following subcorpus for the left part *from one* with $d = 4$: *end of the Earth to $ extreme to the other foo $ shore to the other bar $*.[14] From this subcorpus, one would find the patterns: $from\ one\ [end\ of\ the\ Earth \mid extreme \mid shore]\ to$ and $from\ one\ [extreme|shore]\ to\ the\ other$. It is clear that the first of these patterns has been artificially truncated. This problem is solvable, but it takes a bit of bookkeeping. The idea here is that when a subcorpus is formed, for each token in the subcorpus, a record is kept of where that token was located in the original (parent) corpus.[15] With this record, the end locations of each occurrence of $from\ one\ [end\ of\ the\ Earth \mid extreme \mid shore]\ to$ can be found in the parent corpus. The longest common prefix can then be found for the set of sequences starting at these end locations, and this can be used to extend the truncated right part. There is still a problem, however, since if $from\ one\ [end\ of\ the\ Earth|extreme|shore]\ to$ is extended to $from one [extreme|shore] to the other$, then two instances of this latter pattern will be found. So an efficient way of avoiding such duplications must be found.

[11] This was at least the basic intuition. In fact, the Yamamoto-Church algorithm did not maximally extend phrases to the left since they did not use the Burrows and Wheeler transformation table as advocated by Abouelhoda et al. (2004).

[12] Aires et al. (2008) presents a rather more complicated formula, in which the interest of a phrase is a function of both the term frequency of its subphrases and the superphrases containing the phrase as a subphrase. This is algorithmically more complex, but may be an improvement.

[13] An alternative is presented in Gerdemann (2010).

[14] The tokens *foo* and *bar* are arbitrary. All sentinels are printed as $ even though different integers are used.

[15] Such record keeping is required in any case if document frequencies are required for the phrases.

Another problem also involves maximal extension. Suppose that the saturated pattern $\alpha$ is chosen as the left part. Since it is saturated, it cannot be extended to $a\alpha$ or $\alpha b$ without losing some of its occurrences. Now suppose that $\beta$ is chosen as a corresponding right part, so that the gapped pattern is $\alpha \ldots \beta$. Now it may be that $\alpha$ by itself is saturated, but nevertheless in this context extensions could be made to $a\alpha \ldots \beta$ or $\alpha b \ldots \beta$ without losing any occurrences. Extending the pattern to $\alpha b \ldots \beta$, since it encroaches upon the length of the gap (represented by ...). So rather than extending the left part, it is preferable to filter out cases such as $\alpha \ldots \beta$ where the left part is extendable. Suppose that $\alpha$ can be extended to $\alpha\prime$, where $\alpha$ and $\alpha\prime$ are both saturated. Then both $\alpha$ and $\alpha\prime$ will be considered as candidate left parts. So more specific instances of $\alpha \ldots \beta$ may be found in any case when this pattern is not saturated. The efficiency of the algorithm is, however, an issue, since the filtering turns it partially into a generate-and-test algorithm.[16]

## 5. Conclusion

Gapped phrase extraction clearly has a lot of utility, as witnessed by the number of language researchers who have investigated such phrases, using very imperfect tools. The proper tool for this purpose is an open question which has not been resolved in this paper. The hope is that, as specified in the title, this paper will serve as a challenge, both to someone interested in algorithm design and implementation or to someone who is interested in further specifying what features a gapped phrase extraction program ought to have.

The benefits to eLearning will be that learner texts will be better characterized in terms of the phrases that that the learner uses, instead of simply in terms of a bag-of-words model. Learners should get feedback indicating which phrases are effective, high-quality, appropriate for a particular domain, etc. Such feedback will result in improved writing, in turn leading to better communication. And ultimately, in terms of social theories of learning, better communication will result in improved learning.

## 6. References

Mohamed Ibrahim Abouelhoda, Stefan Kurtz, and Enno Ohlebusch. 2004. Replacing suffix trees with enhanced suffix arrays. *J. of Discrete Algorithms*, 2(1):53–86.

José Aires, Gabriel Lopes, and Joaquim Silva. 2008. Efficient multi-word expressions extractor using suffix arrays and related structures. In *Proceeding of the 2nd ACM workshop on Improving non-English web searching*, pages 1–8, Napa Valley, California.

Alberto Apostolico and Giorgio Satta. 2009. Discovering subword associations in strings in time linear in the output size. *J. of Discrete Algorithms*, 7(2):227–238.

Alberto Apostolico. 2009. Monotony and Surprise: Pattern Discovery under Saturation Constraints. In Anne Condon, David Harel, Joost N. Kok, Arto Salomaa, and Erik Winfree, editors, *Algorithmic Bioprocesses*, pages 15–29. Springer.

Douglas Biber. 2009. A corpus-driven approach to formulaic language in english: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3):275–311.

Gaston Burek and Dale Gerdemann. 2009. Maximal phrases based analysis for prototyping online discussion forums postings. In *Proceedings of the RANLP workshop on Adaptation of Language Resources and Technology to New Domains (AdaptLRTtoND)*, Borovets, Bulgaria.

Markus Dickinson and W. Detmar Meurers. 2005. Detecting errors in discontinuous structural annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, Ann Arbor, MI, USA.

Stefan Evert. 2009. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics: An International Handbook of the Science of Language and Society*, volume 2, chapter 58, pages 1212–1248. Mouton de Gruyter, Berlin/New York.

Dale Gerdemann. 2010. Suffix and prefix arrays for gappy phrase discovery. Presented at: First TübingenWorkshop on Machine Learning; Slides at: http://www.sfs.uni-tuebingen.de/ dg/ks.pdf.

Georgi Jetchev. 1997. *Ghost Vowels and Syllabification: Evidence from Bulgarian and French.* Ph.D. thesis, Scuole Normale Superiore di Pisa.

Dong Kyue Kim, Minhwan Kim, and Heejin Park. 2008. Linearized suffix tree: an efficient index data structure with the capabilities of suffix trees and suffix arrays. *Algorithmica*, 52(3):350–377.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Robert Sedgewick and Kevin Wayne. 2010 (forthcoming). *Algorithms*. Addison-Wesley, 4th edition. Web page: www.cs.princeton.edu/algs4/home (see in particular: www.cs.princeton.edu/algs4/51radix and www.cs.princeton.edu/courses/archive/spring10/cos226/lectures/16-51RadixSorts-2x2.pdf).

Maria Tchalakova. 2010. Automatic sentiment classification of product reviews. Master's thesis, Universität Tübingen, Germany.

Mikio Yamamoto and Kenneth W. Church. 2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Comput. Linguist.*, 27(1):1–30.

D. Zhang and W.S. Lee. 2006. Extracting key-substring-group features for text classification. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 474–483. ACM.

[16] Even as a partly generate-and-test algorithm, initial tests suggest that this approach may be efficient enough for practical purposes. One helpful strategy would be to recognize special cases where the tests can be avoided. For example, if the candidate left part is already supermaximal (Abouelhoda et al., 2004) by itself, then it will not be necessary to check for extensions of this left part when it combines with a right part.

# Towards Resolving Morphological Ambiguity in Arabic Intelligent Language Tutoring Framework

**Khaled Shaalan[1], Marwa Magdy[2], Doaa Samy[3]**

[1] The British University in Dubai, PO Box 345015 Dubai, UAE

[2] Faculty of Computers & Information, Cairo University, 5 Ahmed Zewel St., Giza 12613 Egypt

[3] Cairo University

khaled.shaalan@buid.ac.ae, m.magdy@fci-cu.edu.eg, doaasamy@cu.edu.eg

## Abstract

Ambiguity is a major issue in any NLP application that occurs when multiple interpretations of the same language phenomenon are produced. Given the complexity of the Arabic morphological system, it is difficult to determine what the intended meaning of the writer is. Moreover, Intelligent Language Tutoring Systems which need to analyze erroneous learner answers, generally, introduce techniques, such as constraints relaxation, that would produce more interpretations than systems designed for processing well-formed input. This paper addresses issues related to the morphological disambiguation of corrected interpretations of erroneous Arabic verbs that were written by beginner to intermediate Second Language Learners. The morphological disambiguation has been developed and effectively evaluated using real test data. It achieved satisfactory results in terms of the recall rate.

## 1. Introduction

An Intelligent Language Tutoring System (ILTS) is a computer-based educational system that allows simulation of a human tutor. An ILTS is a valuable tool used in language e-learning programs. Besides, it is highly demanded as an application within the Natural Language Processing field since it helps people in the language learning process either for native or for foreign languages. These NLP tools used in language learning can be used in several ways such as *parsing* of the learner input and *diagnosis* of morphological and syntactic errors (Nerbonne, 2003). However, ILTS for error diagnosis to analyze learners' input and provide intelligent and real-time feedback is highly needed for the following reasons:

- ILTS provide individualized tutoring to learners who are often left to themselves and cannot rely upon teachers and tutors to help them.
- Reliable error diagnosis systems would allow users/authors to overcome the limitations of multiple choice questions and fill-in-the-blanks types of exercises. Besides, ILT systems can provide a suitable platform for introducing more communicative and interactive tasks to learners (L'haire and Faltin, 2003).

Unfortunately, almost all NLP tools such as parsers, morphological analyzer, etc, are designed to handle well-formed input. So, to handle ill-formed input in ILTS, techniques such as constraint relaxation are employed (Faltin, 2003). In any language model, the partial structures can combine only if some constraints or conditions are met. When these constraints are relaxed, an attachment is allowed even if the constraint is not satisfied. The relaxed constraint must be marked on the structure such that the type and position of the detected error can be indicated (confirmed) later on. In ILTS, relaxing the constraints of the language to analyze learner's answer inevitably produce ambiguous solutions, i.e., more corrected interpretations, than systems designed for only well-formed input (Attia, 2006). Consider, for example, the learner input Arabic word عيشت. This would have two interpretations: 1) the learner might mean عشت /Ei$tu/[1] (lived-I) which is related to problems with vowel letters that makes the short vowel الكسرة /i/ long one ياء /y/, or 2) s/he might mean عيشت /Eay~a$tu/ (sustained-I).

This paper addresses issues related to the morphological disambiguation of corrected interpretations of erroneous Arabic verbs written by beginner to intermediate Second Language Learners (SLLs). The proposed system follows the approach a language teacher uses in disambiguating and selecting a preferred analysis. It considers the likelihood of an error which takes into account the level of instruction and the frequency and/or difficulty of Arabic concepts. The concern here is to avoid misleading or incorrect feedback. The result of disambiguation and selecting appropriate analysis is used within ILTS framework to detect the exact source of error and provide the error specific feedback.

Ahmed (2000) addressed the problem of Arabic morphological disambiguation to select the most likely morphological analysis for each well-formed word in the text. He used a powerful dynamic n-gram statistical disambiguation technique. The statistical knowledge of the system may be altered or adjusted anytime to consider any desired text corpus. But, to the best of our knowledge no research has addressed the problem of disambiguating *corrected* interpretations of ill-formed Arabic verbs.

The rest of this paper is structured as follows. Section 2 presents a brief discussion of Arabic morphological ambiguity problem. Section 3 describes the proposed system. Section 4 discusses the results from the conducted experiment. Finally, in Section 5, we give some concluding remarks.

[1] Buckwalter transliteration is used here to Romanize Arabic examples (Buckwalter 2002).

## 2. Arabic Morphological Ambiguity Problem

Arabic language is one of the Semitic languages that is defined as a *diacritized* language where the pronunciation of its words cannot be fully determined by their spelling characters only. Diacritics are special marks put above or below the spelling characters to determine the correct vocalization and, thus, the correct pronunciation.

Unfortunately, diacritics are rarely used in current Arabic writing conventions. The correct pronunciation and interpretation of none or partially diacritized text depends on the native language competence and the context. Due to the optional diacritization, two or more words in Arabic are homographic: they have the same orthographic form, though the pronunciation and meaning is totally different (Ahmed, 2000; Attia, 2006; Habash, 2004). Table 1 listed some homographic examples.

| Word | Lemma | Different Interpretations |
|---|---|---|
| يعد /yEd/ | أعاد />aEAd/ | يعِد /yuEid/ (bring back) |
| | عاد /EAd/ | يعُد /yaEud/ (return) |
| | وعد /waEid/ | يعِد /yaEid/ (promise) |
| | عد /Ead~/ | يَعُدّ /yaEud~/ (count) |
| | أعد />aEd~/ | يُعِدّ /yuEid~/ (prepare) |

**Table 1:** An Arabic word that is homographic

However, other factors contribute to the problem of morphological ambiguity in Arabic. Among these factors (Attia, 2006):

1. Orthographic alteration operations (such as deletion) frequently produce inflected forms that can belong to two or more different lemmas as shown in Table 1. These alteration operations are due to the phonological constraints of certain root consonants. The important irregularity issues are related to Arabic weak verbs that include one or more weak letter. Weak letters can be deleted or substituted by other letters because of Arabic phonological constraints (El-Sadany and Hashish 1989). For example, the deletion of the letter (و) in taking the present (imperfect) tense of the trilateral root و-ع-د /w-E-d/, using regular rules would generate يوعد* /ya-wEid/ but as it is a assimilated (first weak) verb it should be generated according to special weak rules and thus it appears in written texts as يعد /ya-Eid/ (promise).
2. Some Arabic patterns are different only in that one of them has a doubled sound which is not explicit in writing of their corresponding forms such as فعل /faEala/ and فعّل /faE~ala/.
3. Many inflectional operations underlie a slight change in pronunciation without any explicit orthographical effect due to lack of short vowels (diacritics). An example of this is the ambiguity of active vs. passive vs. imperative verb forms.
4. Some prefixes and suffixes can be homographic with each other. For example, the perfect verb suffix ت /Teh/ can indicate either: 1) first person singular, 2) second person singular masculine, 3) second person singular feminine, or 4) third person singular feminine.
5. Prefixes and suffixes can accidentally produce a form that is homographic with another full form word. For example, the word أسد can be interpreted as أسد />asad/ (lion) or أسُدّ />a-sud~/ (I-Block).

Difficulties in the process of Arabic morphological disambiguation are the main reason behind addressing the challenges of developing a morphological disambiguation module/tool/ etc that can handle ill-formed Arabic verbs.

## 3. The Proposed Disambiguation System

The proposed system is an integral part of an Arabic ILTS for SLLs. The system is cable of analyzing both well- and ill-formed learner answers. The ILTS analyzes each input word and produces all of its possible analyses (Shaalan, Magdy and Fahmy, 2010). Afterwards, the ILTS sends these analyses to the disambiguation system to select the appropriate analysis. The selected analysis is then used to detect the exact source of error introduced by the learner and, consequently, the ILTS generates a full diagnosis of the learner input. This is clarified by the following figure.

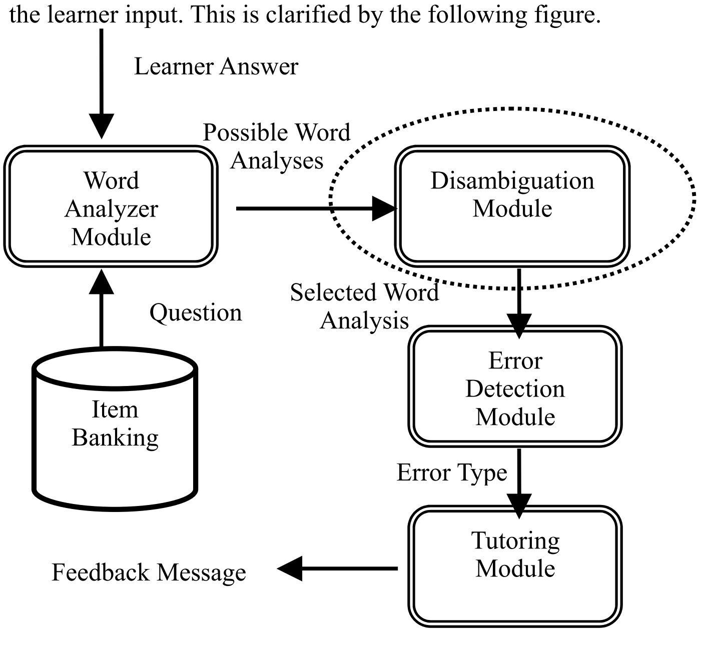


Figure 1: Arabic ILTS Framework

The following example clarifies how the system works. Consider the following question that is presented to the learner:

*Example 1*:
*Complete the following sentence with the correct conjugation of the given root in imperfect tense active voice.*

...... (ب-ي-ع) جدتي الارز
/…. (b-y-E) jad~atiy Al>aruz~/ [my grandmother .... (sell) the rice]

In the above example, the root ب-ي-ع /b-y-E/ contains middle weak letter ي /y/ so it needs special rules to conjugate it in different forms. For example to conjugate it into imperfect passive voice, the middle weak letter should be substituted by ا /A/ so it become تُباع /tu-baAE/ (was sold)

Assume the following two answers; where (a) includes a wrong conjugation of a *Hollow* (middle weak) verb, and (b) is the correct answer.

a. تباع جدتي الارز /ta-biAE jad~atiy Al>aruz~/ (My-grandmother sells the-rice).

b. تبيع جدتي الارز /ta-biyE jad~atiy Al>aruz~/ (My-grandmother sells the-rice).

The ILTS produces two possible analyses for the erroneous word تباع:

- *Third person singular feminine imperfect verb in the active voice with converted middle letter ي /y/ to ا /A/.*
- *Third person singular feminine imperfect verb in the passive voice.*

Then the disambiguation system selects the most appropriate analysis according to: the learner level and difficulty of Arabic concepts[2]. For example in Arabic, the passive voice is a rare construction and it is doubtful that a beginner learner of Arabic would write a passive voice of a verb instead of its active voice. Therefore, the system adopts some *prioritized conditions* to select the most preferred word analysis. Hence, in this case, the system will select the *first analysis*. This analysis is later on used by ILTS to detect the error made by the (incorrect conjugation of verb in imperfect tense active voice)

In the proposed system, we investigated our disambiguation approach on the following three types of ambiguous analysis of erroneous learner input:

1. The orthographic match in non-diacirtized text between Arabic conjugated verb forms in passive voice, and active voice, imperfect or perfect tense, respectively. For example, ( نَقَلَ /*naqala/*) is the perfect tense of the 3rd person singular masculine in active voice, while (نُقِلَ /*nuqil/*) is the perfect tense for the 3rd person singular masculine in passive voice. Same phenomenon is repeated in the imperfect tense (يَنقُل|يُنقَل /*yanqul|yunqal/*)

2. The orthographic match between different affixes in terms of spelling characters. These affixes are used to conjugate different verb forms. For example the prefix (ت) can be used to conjugate the present tense of the 3rd person feminine singular (هي تذهب) and the 2nd person masculine singular (أنت تذهب)

3. The orthographic match between Arabic verb derivation patterns and non-derivative patterns. For example, the verb سعد /saEada/ (to be happy) is a root, non-derivative verb. A possible derivative pattern is أسعد/AsEada/(to make happy). The imperfect conjugation for the first person of the first verb is (أسعد /*AsEada*/), which is identical to the conjugation of the 3rd person singular in the perfect tense of the second verb (هو أسعد /*AsEada*/).

There are some other types of ambiguities[3] that are out of the scope of the current system as the system has no direct knowledge of what the student meant to express. In some systems, where the system has insufficient knowledge to proceed with, a dialogue is established with the learner in order to guide the selection of appropriate expression, e.g. (Hsieh et al., 2002). Figure 2 presents how the system disambiguates multiple analyses and the rest of this section explains in more details.

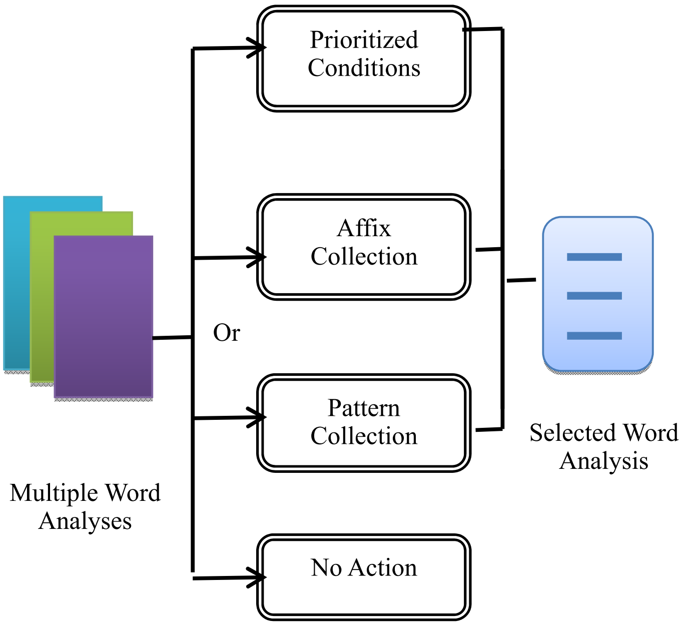


Figure 2: Disambiguation System Structure

In case of the *first ambiguity type*, the system selects the word analysis a student most likely intended. It implements *two* prioritized conditions to selects the most preferred word analysis:

1. If the question goal is to test *passive voice* then the system selects *passive voice* analysis; otherwise, it selects the *active voice* analysis, or

[2] This rule is applied by Arabic language teacher (Heift, 1998).

[3] Example of these types is when the *noun* has the same orthographic form as *verb*

2. If the question goal is to test *imperative tense* then the system selects the *imperative tense* analysis; otherwise, it selects the *perfect or imperfect tense* analysis.

By this way, in Example 1, the system applies the first condition to select the first analysis (*Third person singular feminine imperfect verb in the active voice*). Notice, however, that the question objective is to test conjugation of imperfect active voice verb.

In case of the *second ambiguity type* (i.e. orthographic match between different affixes), the system collects all affixes with the same orthographic form but which differs in their morpho-syntactic features in one entry with a generic feature structure.

For example, consider the following learner input; where (b) is the correct answer:

a. محمد تورطت في جريمة قتل /muHam~ad tawar~aTt fiy jariymap qatol/ (Mohamed was-involved in murder crime).

b. محمد تورط في جريمة قتل /muHam~ad ta-war~aTa fiy jariymap qatol/ (Mohamed was-involved in murder crime).

The learner here has made a subject-verb disagreement between the subject Mohamedمحمد and the verb was-involved تورطت. Four possible analyses of the erroneous verb are produced:

- *First person singular perfect verb in the active voice.*
- *Second person singular masculine perfect verb in the active voice.*
- *Second person singular feminine perfect verb in the active voice.*
- *Third person singular feminine perfect verb in the active voice*

These four possible analyses are combined into the generic analysis:

- *Singular perfect verb in the active voice.*

In case of the *third ambiguity type* (i.e. orthographic match between different patterns), the system collects all these patterns in one entry with a generic feature structure.

For example, consider the following question that is presented to the learner:

*Example 2*:
*Complete the following sentence with the correct conjugation of the given root in perfect tense active voice.*

*جدي وجدتي .... (ن-ق-ل) إلي بيت جديد*

/jad~iy wajad~apiy …. (n-q-l) <ilaY bayot jadiyd/ (my grandfather and my grandmother .... to a new house)

Assume the following learner input; where input (b) is the correct answer:

a. جدي و جدتي نقلوا إلي بيت جديد / jad~iy wajad~apiy naq~aluwA <ilaY bayot jadiyd / (my-grandfather and my-grandmother moved to a new house).

b. جدي وجدتي انتقلا إلي بيت جديد /jad~iy wajad~apiy {inotaqalA <ilaY bayot jadiyd/ (my-grandfather and my-grandmother moved to a new house).

The learner here has made two errors: 1) subject-verb disagreement between the subject "my-grandmother and my-grandfather جدي وجدتي" and the verb "نقلوا", the subject is dual while the verb is conjugated in the masculine plural form and, 2) incorrect use of the root pattern of a perfect verb form; the correct pattern is 'افتعل' while the learner used the pattern 'فعل'. However, the ILTS produced two possible analyses as shown in the following:

- *Third person masculine plural perfect verb in the active voice following the pattern 'فعل'.*
- *Third person masculine plural perfect verb in the active voice following the pattern 'فعّل'.*

These two possible analyses are combined into generic feature structure:

- *Third person masculine plural perfect verb in the active voice.*

# 4. Experiment

We conducted an experiment that measures how successfully the proposed model selects the most appropriate analysis that is used later on to detect the exact source of error the learner has made. The *quantitative* measures are used. These measures rely on collecting different test sets written by real SLLs in a typical teaching/learning environment. It was necessary that these learners have different backgrounds (i.e., differ in their first language) to test if the system is general enough and not aimed to a specific sort of learners. The test set is then fed into the system and the solved ambiguous cases and unsolved are reported. The recall rate is calculated. This measure has been used in evaluating similar research (cf. Wagner et al., 2007; Sjöbergh and Knutsson 2005; Faltin 2003).

The abovementioned methodology is applied on a real test set that consists of 116 real Arabic sentences. The number of words per sentence varies from 3 to 15 words, with an average of 5.1 words per test sentence. The total number of words in all test sentences are 587 words, 118 of them have lexical verb errors. 72 verbs are ambiguous cases. The system successfully solved 46 cases of them while it failed to select the correct analysis for 26 cases. The next section will discuss all failed cases.

## 4.1 Evaluation Problems Classification

In this section, we discuss all problems which the proposed system failed to select the correct analysis. The major problem is it is difficult to determine what the intended meaning of the learner given the complexity of Arabic language.

The 26 failed cases are classified as follows:

- *Orthographic match between un-vocalized forms.* Arabic ILTS handles un-vocalized rather than vocalized written Arabic text. This leads sometimes to more than one possible match between the same and different word categories. The total number of occurrences of this category is 8 cases. They are classified as follows:

○ *Orthographic/homographs match between verb and noun forms*. This case happens when an Arabic verb has the same orthographic form as a noun. For example, consider the word تناول; it can lead to three possible correct words. It is not clear whether the learner meant the word to be: 1) the noun تناول /tanAwul/ (dealing with/ eating), 2) the perfect verb تناول /tanAwala/ (he/it-dealt with/ ate), or 3) the imperfect verb تناول /tu-nAwil/ (hand over/ deliver). The total number of occurrences of this problem is 7 cases.

○ *The special case of the orthographic match between the Arabic third person singular perfect verb following the pattern أفعل />afoEal/ and the first person singular imperfect verb as the word أوقع.* It can lead to two possible interpretations. It is not clear whether the learner meant the word to be: 1) the perfect verb أوقع />awoqaEa/ (he/it-inflicted), or 2) imperfect verb أوقع />u-waq~iE/ (I-sign). The total number of occurrences of this problem is one case.

- *Additional- orthographic matches as a result of relaxing a constraint.* Applying the constraints relaxation technique in order to be able to analyze erroneous learner answers sometimes introduces extra orthographic matches. The total number of occurrences of this category is 18 cases. They are classified as follows:

○ *Orthographic matches produced for Arabic verbs after relaxing the long vowel to the short one.* For instance, consider the erroneous word هجر. It is not clear whether the learner meant the word to be: 1) هاجر /hAjara/ (he/she/it-emigrated) by making the long vowel a short one, 2) هجّر /haj~ara/ (he/it-deported) by using the pattern فعّل /faE~al/, 3)هجر /hajara/ (he/it-left) by using the pattern فعل /faEal/, or 4) هجر /hajor/ (abandoning) by using nouns instead of verbs. The total number of occurrences of this problem is 8 cases

○ *Orthographic matches produced after allowing incorrect conjugation of a verb.* For instance, consider the erroneous word أجوب. It is not clear whether the learner meant the word to be: 1) the imperfect verb أجيب />u-jiyb/ (I-answer), 2) or imperfect verb أجوب />a-juwb/ (I-explore). The total number of occurrences of this problem is 7 cases

○ *Orthographic matches produced for Arabic verbs after relaxing the short vowel to the long one.* For instance, consider the erroneous word عيشت. It is not clear whether the learner meant the word to be: 1) عشت /Ei$-tu/ (I-lived) by making the short vowel a long one or, 2) عيشت /Eay~a$-tu/ (I-sustained) with using the pattern فعّل /faE~al/. The total number of occurrences of this problem is 2 cases.

○ *Orthographic matches produced after allowing incompatible usage of connected pronouns.* For instance, consider the erroneous word أعملت. It is not clear whether the learner meant the word to be: 1) the perfect verb أعملت />aEomal-tu/ (I-employed) or, 2) the perfect verb عملت /Eamiltu/ (I-worked) by using incompatible pronouns أ, ت (Alef, Teh). The total number of occurrences of this problem is one case.

Notice, however, that we asked human linguists about failed cases and he has identified most of theses cases as ambiguous.

## 5. Conclusion

The ambiguity problem is a standard problem in any NLP application. It is the major reason why computers do not yet understand natural language. However, the ambiguity problem presents a challenge to ILTS. That is because selecting the wrong analysis of student input can lead to misleading feedback or an error might be overlooked. Beside that given the complexity of Arabic language, this makes the ambiguity a serious problem and needs to be resolved. The preferred method in ILTS for disambiguating multiple readings of a wrong answer should consider the likelihood of an error and the difficulty of concepts. But with the lack of erroneous corpus, we depend on some linguistic studies that investigate the likelihood of errors. However, the ambiguity problem cannot be resolved totally and there is a need to issue a dialogue with the learner to know what exactly he means. Moreover, if a large tagged erroneous corpus exist then the ambiguity problem can be resolved by considering the likelihood of errors

## 6. References

Ahmed, M. A. 2000. A Large-Scale Computational Processor of the Arabic Morphology, and Applications. Master thesis, Cairo University, Egypt.

Attia, M. A. 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks. In Proceedings of the Challenge of Arabic for NLP/MT Conference, 2006. The British Computer Society, London.

Buckwalter, T. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2002L49, ISBN 1-58563-257-0.

El-Sadany, T. A. and Hashish, M. A. 1989. An Arabic Morphological System. In IBM Systems Journal, 28(4): 600- 612.

Faltin, A. V. 2003. Syntactic Error Diagnosis in the Context of Computer Assisted Language Learning. PhD thesis, University of Geneva, Switzerland.

Habash, N. 2004. Large Scale Lexeme Based Arabic Morphological Generation. In Proceedings of Traitement Automatique du Langage Naturel (TALN-2004). Fez, Morocco.

Heift, T. 1998. Designed Intelligence: A Language Teacher Model. Ph.D. Thesis, Simon Fraser University, Canada.

Hsieh, C.-C., Tsai, T.-H., Wible, D. and Hsu, W.-L. 2002. Exploiting Knowledge Representation in an Intelligent Tutoring System for English Lexical Errors. In Proceedings of the International Conference on Computers in Education ICCE 2002, Auckland, New Zealand, pp: 115-116.

L'haire, S. and Faltin, A. V. 2003. Error Diagnosis in the FreeText Project. In Calico Journal, 20 (3): 481-495.

Nerbonne, J. 2003. Natural Language Processing in Computer-Assisted Language Learning. In Ruslan Mitkov, editors, the Oxford Handbook of Computational Linguistics. Oxford, pp: 670-698.

Shaalan, K., Magdy, M., and Fahmy, A. 2010. Morphological Analysis of Ill-formed Arabic Verbs in Intelligent Language Tutoring Framework. In Proceedings of FLAIRS-23, Daytona Beach, Florida, USA. To appear.

Sj¨obergh, J., and Knutsson, O. 2005. Faking Errors to Avoid Making Errors: Machine Learning for Error Detection in Writing. In Proceedings of RANLP 2005, Borovets, Bulgaria, pp: 506-512.

Wagner, J., Foster, J., and Genabith, J. V. 2007. A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors. In Proceedings of EMNLP-CoNLL 2007, Prague, Czeck Republic, pp: 112-121.

# Language Resources and Visual Communication in a Deaf-Centered Multimodal E-Learning Environment: Issues to be Addressed

**Elena Antinoro Pizzuto[1], Claudia S. Bianchini [2,1], Daniele Capuano [3] Gabriele Gianfreda [4,1] , Paolo Rossini[1]**

[1] Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Via Nomentana, 56 – 00161, Roma. [2] Université Paris 8, Rue de la Liberté, 2 - 93526 St.Denis CEDEX / DIFILILE, Università di Perugia, Piazza Morlacchi, 1 – 06123, Perugia, Italy. [3] Università di Roma "Sapienza", Dipartimento di Informatica, Pictorial Computing Laboratory, Via Salaria 113, 00198 Roma. [4] Dipartimento di Scienze dell'Educazione e della Formazione, Università di Macerata, P.le Bertelli, 1, C.da Vallebona, 62100 Macerata – Italy,

E-mail: elena.pizzuto@istc.cnr.it, chiadu14@tiscali.it, el.einad@gmail.com, hyuga@email.it, paolo.rossini@istc.cnr.it

## Abstract

This paper examines some of the major problems linked to the task of designing appropriate multilingual e-learning environments for deaf learners (DL). Due to their hearing disability most DL experience dramatic difficulties in acquiring appropriate literacy skills. E-learning tools could in principle be very useful for facilitating access to web-based knowledge and promoting literacy development in DL. However, designing appropriate e-learning environments for DL is a complex task especially because of the different linguistic background and experience DL may have, and of the multimodal language resources that need to be provided and integrated (e.g. language produced in the visual-gestural or signed modality, in written texts, closed captioning for vocal language information). The purpose of this paper is twofold: (1) describe and discuss issues we believe need to be addressed, focusing on the limitations that appear to characterize several e-learning platforms that have been proposed for DL; (2) present and discuss ongoing research aimed at overcoming these limitations.

## 1. Introduction

It is widely known that all over the world deaf children and, later, adults, experience dramatic difficulties in achieving appropriate receptive and expressive skills not only in oral or vocal language (VL) but also in written language. The vast majority of deaf learners (DL) achieve literacy levels that are markedly below those proper of their hearing peers (see among others Caselli, Maragna & Volterra, 2006; Garcia & Derycke, 2010; Garcia & Perini, 2010). As a result, in their school years through adulthood, DL experience equally dramatic difficulties in accessing the vast body of knowledge, and the rich learning environments made available by advanced multimedia technologies, most notably e-learning environments. Appropriate written language skills are in fact unquestionably a pre-requisite for exploiting the possibilities arising from such multimedia and multimodal learning environments.

In Italy as all over the world[1] the situation of DL is especially complex due to the very different language background and experience deaf persons may have depending upon the language they use as their primary or preferred means of communication, or L1. It is in fact necessary to distinguish two groups: (1) those who use Italian Sign language (LIS), the visual-gestural, face-to-face language of the Italian deaf community (LIS-L1); (2) those who prefer to use spoken and written Italian (Italian-L1). It is important to stress that, on the whole, *both groups of DL* experience severe difficulties in achieving appropriate literacy levels – though of course 'exceptional learners' who overcome these difficulties can be found within each group.

With respect to signers, the following must be noted. Since the modern study of signed languages (SL) began with Stokoe's (1960) pioneering work on American Sign language (ASL), world-wide research has led to describe, and to recognize as full-fledged human natural languages, a very large number of national SL, including LIS and all the other major European signed languages. The use of SL for instructional purposes has been explicitly recommended by the European Parliament (see Resolution 17-6-1988, art. D).

Bilingual education programs that offer signed and oral/written language instruction to deaf students have been developed in several countries, including Italy where they have been applied for the most to Elementary school children. As reported by Caselli & al (2006), it is unquestionable that the use of a SL, even if limited to its usual, face-to-face- form, can play a very important role in fostering DL's general linguistic competence.

The inclusion of SL within e-learning platforms designed for DL has come as a natural development of the advancement that have been made in our knowledge of SL and of deaf signers. However, as we point out in sections 2 to 4 below, many recent and current attempts to develop appropriate e-learning environments for DL

[1] For reasons linked to the demography of deafness and to the complex sociolinguistic and cultural properties of signed languages the observations we make here with respect to Italy can be easily extended across nations and cultures, with the necessary changes concerning the national signed and vocal/written languages.

exhibit, and/or implicate some major conceptual, methodological and practical limitations.

In section 5 we present and discuss ongoing research aimed at overcoming these limitations.

## 2. Some general problems concerning existing e-learning platforms for DL

For the purposes of this paper, we limit our attentions to e-learning platforms designed for young or adult DL. An overview of several such platforms reveals the following major limitations. First, the guidelines for developing the desired platforms are often just "sketched", and provide fairly general suggestions concerning, for example: - the inclusion of SL videos with SL translations or explanations of the written texts found in a specific e-learning platform; - the development of automatic tools (i.e. avatars) for translating written texts into SL; -the use of cooperation tools such as video conferencing[2]. Second, many existing or planned platforms appear to be designed primarily for DL who know SL, but *seem to neglect the needs of DL who prefer to use VL.*

On the whole, there appears thus to be a general trend towards creating and including SL materials for implementing written text-based environments. The contents encoded in written language are made more accessible to (signing) DL via SL translations and explanations. Other examples are the platform created within the project DEAL [3] for teaching foreign vocal-written languages to DL, or the one designed by Drigas & Kouremenos (2005) for vocational and general educational training.

A fairly large body of work has been dedicated to the development of signing avatars to be added to the users' interface, replacing SL materials presented by real signers (see for example Efthimiou & Fotinea, 2007; 2008; Karouzis, Caridakis, Fotinea & Efthimiou, 2007, or also the recent Italian project "ATLAS" [4]).

Many projects for realizing signing avatars exhibit however, in our view, a rather surprising limitation: they appear to have *a unidirectional, VL-centered perspective.* They start, for the most, from VL written texts and aim at producing avatars that can translate such written texts into individual signs and signed sequences. These project thus ignore or underestimate the problem of *translating from sign to vocal/written texts*. There are only few projects that explicitly aim at realizing signing avatars functioning in both directions, i.e. from sign to speech and/or also written texts, and from speech and written texts to sign. One example is "Signspeak" (see also the project "Dicta sign") [5]

## 3. SL communication and instructional materials: what models of SL to adopt?

Irrespective of whether real signers or signing avatars are used, one additional limitation of many current efforts towards integrating SL materials into e-learning platforms concerns a failure to recognize important differences between SL and vocal/written languages, and the problems posed by the dramatically insufficient reference tools, and overall linguistic descriptions, that are currently available for SL.

It must first be recalled that *all SL are languages without a written tradition.* More importantly from a research standpoint, and even though almost 50 years have passed since the modern study of SL has begun, researchers still have not found an agreement on: (a) what are the constituent elements of SL; (b) what graphic systems can be used for representing SL in written form and, on this basis, develop appropriate reference tools (e.g. dictionaries, grammars, usage-based corpora etc) that are unquestionably necessary for both the communities of signers, and the exploitation of SL for educational and instructional purposes (see Cuxac & Antinoro Pizzuto, 2010; Garcia, 2006; 2010; Garcia & Derycke, 2010).

It is not trivial to stress that, although our knowledge of SL has considerably advanced, we still do not have any monolingual dictionary or grammar, for any of the SL that has been to date investigated - not even for ASL.

In this context, one could expect that well-grounded proposals aimed at exploiting SL for instructional purposes would dedicate particular care in making explicit the models of SL elements and discourse they adopt. This appears especially necessary because, as recalled hereafter, there are at present two major classes of models for describing SL. In agreement with Cuxac & Sallandre (2007) we will refer to these models as "assimiliationist" vs. "non assimiliationist": the first type of models highlight primarily the structural similarities between SL and VL, while the second ones underscore that, in addition to important similarities there are equally relevant differences between SL and VL.

Within the limits of the present context, we illustrate some of the crucial differences between these two types of models in relation to the problem of defining what are the constituent elements of SL.

In substantial agreement with early, very influential descriptions of ASL provided by Stokoe (1960) and subsequently Klima & Bellugi (1979), assimilationist models assume that SL constituents units are *essentially comparable to VL words*, and are *primarily sequentially organized in time.* These models are still largely prevailing in current research on SL and have been for the most acritically adopted in educational applications

[2] See for ex.: *Individuals who are Deaf or Hard of Hearing*, Center for Assistive Technology and Environmental Access (CATEA), http://www.accesselearning.net/mod1/1_02.php; *IMS guidelines for Developing Accessibile Learning Applications*, IMS Global Learning Consortium, http://www.imsglobal.org/accessibility/accessiblevers/index.html; *General guidelines for Inclusive Online Cultural Content,* Canadian Network for Inclusive Cultural Exchange, http://cnice.utoronto.ca/guidelines.php

[3] http://www.deal-leonardo.eu

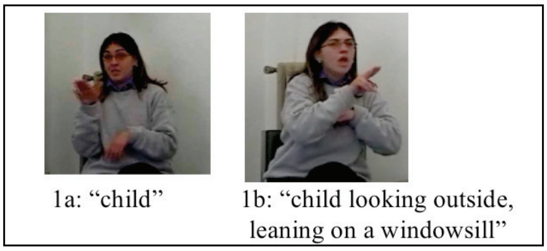
[4] http://www.atlas.polito.it/

[5] http: www.signpeak.eu; http://www.dictasign.eu

of different types, including e-learning platforms.

In contrast, non assimiliationist models, based on extensive analyses of SL discourse, show that SL constituent elements cannot be easily assimilated to VL units. In addition to word-like elements, SL possess complex, highly iconic structures (HIS) that are simultaneously organized in a multilinear fashion that has no parallel in VL. The differences between word-like and non-word-like units are marked by non manual and manual articulators, most notably by modality-specific eye-gaze patterns: when producing word-like units, the signer's gaze is directed towards the interlocutor, whereas when producing HIS the signer's gaze is typically directed away form the interlocutor.

Figure 1 below provides just two illustrative examples of a word-like unit (1a) and a non-word-like HIS (1b) that are commonly found in SL discourse. The examples are taken from LIS discourse but a wealth of similar examples can be found in all SL (for relevant discussions, see especially Cuxac, 2000; Cuxac & Antinoro Pizzuto, 2010; Pizzuto, Pietrandrea & Simone, 2007; Garcia & Derycke; 2010).



1a: "child"

1b: "child looking outside, leaning on a windowsill"

Figure 1: Word-like sign (1a) and HIS (1b)

The point we wish to stress here is the following: HIS are very frequent in SL discourse, ranging from 30% to as much as 90% (depending on discourse genre) of the constituent elements that can be identified and parsed in SL discourse (Boutet, Sallandre & Fusellier-Souza, 2010; Cuxac & Antinoro Pizzuto, 2010; Sallandre, 2003; Di Renzo & al, 2009). It should thus be evident that SL descriptions of any sort, including modelisations via signing avatars, cannot disregard as "marginal" these structures that appear unique of SL (see Cuxac & Dalle, 2007). E-learning materials based on the assumption that SL elements are for the most "just like VL words" thus exhibit severe limitations that need to be recognized, critically discussed and, hopefully, amended.

## 4. Visual attention patterns in DL

An appropriate e-learning environment for DL at large, i.e. for both signers and non-signers, must take in due account a constraint that can be easily observed and yet, to our knowledge, has not been carefully investigated in previous research. When working with a computer, the visual attention patterns proper of DL markedly differ from those observable in hearing learners. This is true especially in situation of cooperative learning where the students must simultaneously attend to visual information concerning written materials of different sorts to be "attended to" and processed, and other information stemming from the interaction with other fellow students and/or with a tutor (e.g. in exchanges taking place in actual classrooms or in videoconferences). Since deaf persons must use their sight, and accordingly orient their visual attention, to process both kinds of information, the two tasks cannot be carried out at the same time: DL cannot simultaneously look at teaching or explanatory materials displayed on the computer screen *and* at linguistic, interaction-based information given on the same materials which they must always decode primarily via vision (e.g. by lip-reading spoken utterances, processing a message in SL, reading subtitles).

This is much unlike what happens, in the same cooperative learning situation, for hearing learners who can simultaneously process communicative messages conveyed through sounds and freely orient their visual attention to other types of information coming from the computer screen. Devising an appropriate e-learning environment for DL thus requires accurate analyses of the ways in which these learners use and distribute their visual attention when performing different learning tasks, and how this can influence the learning process.

## 5. Towards deaf-centered multilingual and multimodal e-learning platforms

Figure 2 schematically illustrates a model for an e-learning platform prototype (ELPP) prototype we are currently developing within the frame of a national project which pursues two major, interrelated objectives: (1) improving multilingual / multimodal e-learning environments for DL (High School and University students); (2) promoting their literacy skills [6].

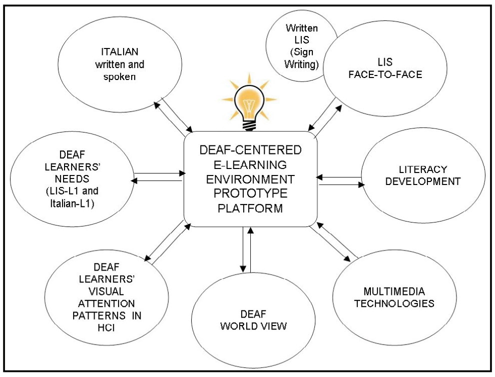

ITALIAN written and spoken
Written LIS (Sign Writing)
LIS FACE-TO-FACE
DEAF LEARNERS' NEEDS (LIS-L1 and Italian-L1)
DEAF-CENTERED E-LEARNING ENVIRONMENT PROTOTYPE PLATFORM
LITERACY DEVELOPMENT
DEAF LEARNERS' VISUAL ATTENTION PATTERNS IN HCI
DEAF WORLD VIEW
MULTIMEDIA TECHNOLOGIES

Figure 2: A model of deaf-centered e-learning platform

The ELLP model illustrated in Figure 2 aims at

[6] The project involves five research teams providing inter- and trans-disciplinary competences across the fields of: -SL linguistics; -special and bilingual education for DL; -multimedia tools for DL and hearing learners; -human-computer interaction and visual learning in e-learning environments; -foreign language teaching methodologies in both traditional and e-learning environments.

overcoming the limitations proper of many e-learning platforms (see discussion above) in different ways. Each of the major "conceptual components" of the model is at the same time *motivated by*, and *necessary for* designing a *deaf-centered e-learning platform*. The platform is grounded upon the idea that research aimed at creating useful products for deaf users needs to be developed, from the very start, *with* deaf persons, not just *for*, or *on* deaf people. Accordingly, and rather differently from what is reported for many past and ongoing projects directed to deaf persons, this idea guides our actual 'project management' practice. The project-leader team includes six deaf colleagues who *participate as protagonists in the planning and articulation of the entire research project*, not only as "end users" or "end evaluators" of the language resources and didactic tools to be produced or implemented. All our deaf colleagues are highly proficient in LIS: three learned to sign in infancy, within deaf signing families, three ad different ages, as it happens to most deaf signers (see Cuxac & Antinoro Pizzuto, 2010); they possess different degrees of knowledge of spoken /written Italian which mirror in part their educational background [7].

We give here just few practical examples of the crucial involvement of our deaf colleagues. The choice of the "contents" we will focus on for developing the ELPP[8], and of the different forms in which such contents will eventually be presented to DL on the ELPP (e.g. spoken and written texts, speech-to-text captions, SL translations and explanations, graphic illustrations), was made following extensive discussions, among the deaf and the hearing members of the project, of different, alternative possibilities. Our deaf colleagues are contributing to the preparation of ad-hoc questionnaires and to a thorough examination and evaluation of language tasks, materials, multimedia technologies we are using and/or are currently developing (including for ex. the ELPP interface). In short, the active involvement of deaf colleagues ensures that the end products of our project be, on one hand, consistent with the "*deaf world view*", (see Figure 2) – i.e. a complex configuration of experiential and conceptual knowledge that is strongly grounded in *vision* (see among others Lane, Hoffmeister & Bahan, 1996), and, on the other hand, effectively respond to *DL needs* (see Figure 2).

One other important element of the deaf-hearing collaboration we are promoting within the project is the following: all the hearing members of the project-leader team possess a good or advanced knowledge of LIS; four of the five (hearing) young researchers of the other research teams involved in the project are currently attending classes to learn LIS. We are also seeking further collaborations with deaf experts who use Italian (rather than LIS) as their preferred language.

As noted above, most e-learning platforms for DL appear to be *designed only for signing DL*. In contrast, as shown in Figure 2, our ELPP aims at *addressing the needs of* ***both*** *signing (LIS-L1) and non signing (Italian-L1) DL*. In fact, as also noted above, both such groups of DL experience dramatic difficulties in *literacy development*. Our research aims at ascertaining the specific communicative-linguistic needs of each group of DL and the extent to which these are, or are not comparable. We expect that the results of our investigations will provide: (a) novel, relevant information on the linguistic-cognitive profile of the two groups of DL, clarifying also whether, and/or how knowledge of LIS as L1 may, or may not, interfere with the acquisition and use of spoken/written Italian; (b) important indications on how we may need to differentiate the multilingual and multimodal materials to be created for promoting literacy development in DL with LIS-L1 as compared to DL with Italian-L1. For example, recalling what noted in section 3, it would be plausible to hypothesize that, for DL with LIS-L1, the simultaneously organized, multilinear linguistic structures that are highly specific of their SL, namely HIS, may negatively interfere with the learning of more sequentially organized linguistic structures that are proper of written language. It would be equally plausible to hypothesize that these potential negative interferences should be absent in DL with Italian-L1. However, these hypotheses can be evaluated only by comparing the linguistic-cognitive profiles of the two groups of DL, as we plan to do in our project.

A substantial novelty of the multilingual / multimodal ELPP e-learning environment we are designing concerns the use, presentation (hence, by the same token, explicit modeling and representation) of the two major types of *language resources* that will be employed for pedagogical purposes, namely: Italian and LIS. What is novel in our model is that, as illustrated in Figure 2, written texts will be provided not only in *written Italian* (the target language in which we aim to promote DL literacy development), but also in *written LIS* – a language resource which, to our knowledge, has never been experimented in e-learning platforms for DL. *Spoken Italian* and *face-to-face LIS* (the latter in the form of digital videos) will also be used (see Figure 2).

For the instructional materials to be provided in *written Italian*, guided easification procedures will be used to facilitate DL's access to textual materials; speech-to-text captioning tools will grant visual accessibility to materials given in *spoken Italian*; linguistic accessibility to the contents and forms of Italian-encoded instructional materials will be enhanced, for DL with LIS-L1, via appropriate videos providing translations and explanations in (face-to-face) LIS. Due to space limits, no further details are given here on these three types of language resources, which will be implemented driving on a consolidated experience in

[7] Spoken/written language proficiency in deaf persons is highly variable and only partially linked to the educational level achieved. Our deaf colleagues include one doctoral student, one college graduate, one University student, three high school graduates.

[8] For space limits we can only mention here the 'general contents' of the ELPP: we will focus on the history, evolution and use of writing, and compare oral/signed vs. graphic/written forms of communication.

bilingual education for DL (Caselli & al, 2006), and more generally in language teaching methodologies, as detailed in our grant proposal. We describe briefly the rationale, empirical grounds, and major aims of our novel experimentation of written LIS.

As noted in section 3, all SL are at present without a written tradition. For DL with LIS-L1, the lack of a written form of their own SL may well be one of the obstacles on the road towards achieving appropriate literacy skills in a language – like Italian - that not only does have a written tradition but is also typologically very different from their own (see especially our remarks above on SL HIS). Recent research shows that Italian signers can profitably use Sign Writing (SW), a graphic system proposed by Sutton (1999) for writing SL, for: -transcribing LIS face-to-face productions; - creating, for the first time in the history of this SL, texts conceived directly in written LIS (SW has been adapted for these purposes to LIS). More importantly for the present discussion, this research shows that, relying on SW-encoded LIS texts, signers can autonomously perform meaningful comparisons between LIS and spoken/written Italian, at all structural levels - lexical, morphological, syntactic, textual, pragmatic.

On this basis, signers can formulate metacognitive and metalinguistic reflections on the structure of LIS as compared to spoken/written Italian, and more generally on the relations between "orality" or face-to-face vs. written communication, in a way that has never been possible, for them, without relying on a written representation of their SL (see among others Di Renzo & al, 2006; 2009; Gianfreda & al, 2009; Pizzuto & al 2006; Antinoro Pizzuto & al, 2008). Taking in due account the crucial role that metacognitive and metalinguistic skills notoriously play in the development of literacy skills, these research findings have motivated us to further experiment written LIS, on our ELPP, as a potentially very powerful pedagogical tool for promoting literacy abilities. SW-encoded, written representations of LIS have also proven to be extremely useful for advancing in the linguistic analysis of the language (Antinoro Pizzuto & al, 2008), paving the way for more appropriate modelisations which may be used for both general descriptive purposes, and for implementing the use of LIS as a linguistic resource on e-learning platforms.

We noted in section 4 that *DL's visual attention patterns in HCI* may significantly differ from those of hearing learners. One other additional novelty of our project concerns the use of eye-tracking equipment for analyzing DL's visual attention patterns, and compare them with those of hearing learners', during learning tasks which demand the simultaneous processing of language resources along with visual information of different sorts. Preliminary results of a pilot study we have conducted indicate that, in processing multimodal / multilanguage materials, the gaze patterns of DL with LIS-L1 markedly differ from those of hearing learners (Capuano, Levialdi & Antinoro Pizzuto, submitted). We trust that the more extensive investigations on this topic we plan to develop within our project will provide us much needed, novel information for a better understanding of how visual information needs to be spatially and temporally structured in e-learning environments for DL, as compared to hearing users. These analyses will also allow to us ascertain whether there are (or not) relevant differences between signing vs. non-signing deaf students, when these DL with different language background access and use visually grounded information, of both linguistic and non-linguistic type.

Finally, recalling the crucial importance of vision in the 'deaf world view', we think that web-based *multimedia technologies and learning tools* for a deaf-centered ELPP may be significantly improved implementing a visually-based graphic interface. Drawing on ongoing research on the topic (Capuano & al, submitted), we aim at designing an interface that DL can access and use easily and 'intuitively' because textual information (which is difficult for them) is significantly reduced, or even entirely replaced by mostly non-textual (iconic) information. This entails the need of creating a new, graphic way for browsing web pages, and interacting with the ELPP.

For the natural, deaf-peculiar visual way of grasping information to be exploited in our platform, we are going to use a new interaction paradigm based on the theories of *embodied cognition* and *storytelling* (Lakoff & Johnson, 1980; Johnson, 1987; Imaz & Benyon, 2007). Within this paradigm, the learning process can be metaphorically represented as a story that includes the user as the main character. Accordingly, the user 'lives' the learning process by physically experiencing it – in the virtual space of the ELPP – as a path with a starting place, a sequence of several learning steps, and a final goal. Such a metaphor seems to be a very intuitive way of representing the learning environment. Moreover, it seems to be an adequate interaction paradigm especially for deaf users, since it exploits the visual channel as the main source of information.

## 6. Acknowledgements

The preparation of this work was partially funded by -the Italian Ministry of Education and Research (MIUR-FIRB), Project *E-learning, Deafness and Written Language / VISEL* –RBNE074T5L (2009-2012), see: http://*www.visel.cnr.it*; -The Association *Progetti Felicità*, Project "Writing LIS and SignWriting" Rome, (2005 -). We thank our colleagues Stefano Levialdi, Marilena De Marsico, Anna Labella and Alessio Di Renzo for helpful comments.

## 7. References

Antinoro Pizzuto, E., Chiari, I. & Rossini, P. (2008). The representation issue and its multifaceted aspects in constructing sign language corpora: questions, answers, further problems. *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages.* LREC 2008, Marrakech (http://www.lrec-conf.org/proceedings/lrec2008/), 150-158.

Boutet, D., Sallandre, M.-A., & Fusellier-Souza, I. (2010). Gestualité humaine et langues des signes: entre continuum et variations. In B. Garcia et M. Derycke (eds.), *Sourds et langue des signes. Norme et variations*, *Langage et Société*, n° 131, mars 2010, 55-74.

Capuano, D., Levialdi, S. & Antinoro Pizzuto, E. (submitted). Embodied visual learning and deafness: a concept paper.

Caselli, M.C., Maragna, S. & Volterra, V. (2006). *Linguaggio e sordità*. Bologna: Il Mulino,

Cuxac, C. (2000). *La Langue des Signes Française; les Voies de l'Iconicité*, Faits de Langues n°15-16, Paris: Ophrys.

Cuxac, C. & Antinoro Pizzuto, E. (2010), Emergence, norme et variation dans les langues des signes : vers une redéfinition notionnelle. In B. Garcia & M. Derycke (eds.), *Sourds et langue des signes. Norme et variations*, *Langage et Société*, n° 131, mars 2010, 37-53.

Cuxac, C., Dalle, P. (2007). Problématique des chercheurs en traitement automatique des langues des signes. In: *Traitement Automatique des Langues, (ATALA) - Traitement automatique des langues des signes*, Vol. 48, N. 3, 15-30 http://www.atala.org/-Modelisation-et-traitement-des-

Cuxac, C. et Sallandre, M-A. (2007). Iconicity and arbitrariness in French Sign Language: highly iconic structures, degenerated iconicity and diagrammatic iconicity, in E. Pizzuto, P. Pietrandrea & R. Simone (eds), *Verbal and Signed Languages, Comparing Structures, Constructs and Methodologies*, Mouton de Gruyter, Berlin-New York, 13-33.

Di Renzo, A., Gianfreda, G., Lamano, L., Lucioli, T., Pennacchi, B., Rossini, P., Bianchini, C.S., Petitta, G., Antinoro Pizzuto, E., (2009). Representation – Analysis - Representation: novel approaches to the study of face-to-face and written narratives in Italian Sign Language (LIS). Paper Presented at the *CILS International Conference on Sign Languages*, Namur, Belgium, November 16-20, 2009.

Di Renzo, A., Lamano, L., Lucioli, T., Pennacchi, B., Ponzo, L., (2006), Italian Sign Language: Can we write it and transcribe it with Sign Writing ? In ELRA (eds.), *LREC 2006, Workshop Proceedings* (*W-15): Second Workshop on the Representation and Processing of Sign Languages* , 11-16.

Drigas, A.S. & Kouremenos, D. (2005). An e-Learning Management System for the Deaf people, *Proceedings of WSEAS Transactions on Advances in Engineering Education*, Vol I, 2, 20-24.

Efthimiou, E. & Fotinea, S. (2007), An Environment for Deaf Accessibility to Educational Content. *Proceedings of the First International Conference on Information and Communication Technology and Accessibility*,

Efthimiou, E. & Fotinea, S. (2008), Tools for Deaf Accessibility to an eGOV Environment, in *Lecture Notes in Computer Science (LNCS*), Vol. 5105, 446-453.

Garcia, B. (2006). The methodological, linguistic and semiological bases for the elaboration of a written form of LSF (French Sign Language). In ELRA (eds.), *LREC 2006 - Workshop Proceedings* (*W-15), Second Workshop on the Representation and Processing of Sign* Languages, 31-36.

Garcia, B. (2010). *Sourds, surdité, langue(s) des signes et épistémologie des sciences du langage. Problématiques de la scripturisation et modélisation des bas niveaux en Langue des Signes Française (LSF).* Mémoire d'Habilitation à Diriger les Recherches, Université Paris 8—Saint-Denis.

Garcia, B. & Derycke, M. (2010). Introduction. In B. Garcia & M. Derycke (eds.), *Sourds et langue des signes. Norme et variations*, *Langage et Société*, n° 131, mars 2010, 5-17.

Garcia, B. et Perini, M. (2010). Normes en jeu et jeu des normes dans les deux langues en présence chez les sourds locuteurs de la Langue des Signes Française (LSF). In B. Garcia et M. Derycke (eds), *Sourds et langue des signes. Norme et variations*, *Langage et Société*, n° 131, mars 2010, 75-94.

Gianfreda, G., Petitta, G., Bianchini, C.S., Di Renzo A., Rossini, P., Lucioli, T., Pennacchi, B., Lamano, L. (2009). Dalla modalità faccia-a-faccia a una lingua scritta emergente: nuove prospettive su trascrizione e scrittura della Lingua dei Segni Italiana (LIS). In C. Consani, C. Furiassi, F. Guazzella & C. Perta (eds.), *Atti del 9° Congresso dell'Associazione Italiana di Linguistica Applicata – Oralità / Scrittura. In memoria di Giorgio Raimondo Cardona*. Perugia: Guerra Edizioni, 413-437.

Imaz, M. & Benyon. D (2007). *Designing with blends.* The MIT Press.

Karpouzis, K., Caridakis, G., Fotinea, S.-E., & Efthimiou, E. (2007), Educational resources and implementation of a Greek sign language synthesis architecture. *Computers & Education*, 49, 54-74.

Klima E.S. et Bellugi, U. (1979). *The Signs of Language.* Cambridge, MA: Harvard University Press.

Johnson, M. (1987). *The body in the mind.* University of Chicago Press.

Lakoff, G. & Johnson, M. (1980). *Metaphor We Live By.* University of Chicago Press.

Lane, H., Hoffmeister, R.J. & Bahan, B. (1996). A journey into the Deaf-World. San Diego, CA: DawnSign Press.

Pizzuto, E., Pietrandrea, P., & Simone, R. Introduction. In E. Pizzuto, P. Pietrandrea & R. Simone (eds.) (2007), *Verbal and Signed Languages - Comparing structures, constructs and methodologies*, Berlin / New York: Mouton De Gruyter, 1-10.

Pizzuto, E. Rossini, P. & Russo, T. (2006). Representing signed languages in written form: questions that need to be posed. In ELRA (eds.), *LREC 2006 – Workshop Proceedings* (*W-15): Second Workshop on the Representation and Processing of Sign* Languages, 1-6.

Sallandre, M.-A. (2003). *Les unités du discourse en Langue des Signes Française (LSF) – Tentative de de categorization dans le cadre d'une grammaire de l'iconicité.* Thèse de Doctorat en Sciences du Langage, Paris, Université Paris 8.

Stokoe, W.C. (1960) Sign Language Structure, in *Studies in Linguistics – Occasional Paper* n. 8, 1960 (rev. ed. Linstok Press, Silver Spring, MD, 1978).

Sutton, V. (1999). *Lessons in SignWriting. Textbook & Workbook*. La Jolla, CA: Deaf Action Committee for Sign Writing (2nd edition, 1st edition 1995).

# Deaf People Education: crossing linguistic borders through e-learning

**Giuseppe Nuccetelli, Maria Tagarelli De Monte**
Istituto per Sordi di Roma
Via Nomentana 56, 00161 Roma, Italy
E-mail: gnuccetelli@uniroma3.it, mariatdemonte@istitutosordiroma.it

### Abstract

The introduction of Web Technologies and the development and spread of portable devices has improved the quality of life of deaf people making distant communication easier. In particular, the development of online systems including video-messaging and the possibility to upload user generated contents, has given deaf people the possibility to rely on other, more direct, means of communication. Similarly, the development of e-learning platforms and their adoption in most Universities worldwide, is shaping the way education is conceived, leading to new and innovative systems merging in-class education with e-learning systems. Our contribution gives a first explanation of how Information and Communication Technology (ICT) can be a strategic resource to give deaf people equal educational opportunities focusing on the development of appropriate language skills, and the strategies through which these opportunities can become effective. Our experience is based on the results and outcomes of DEAL Project (Deaf people in Europe Acquiring Languages through E-Learning), carried out from Istituto Statale per Sordi Roma (ISSR - State Institute for the Deaf in Rome) with co-financing from the European Commission. The objective being that of creating an e-learning model for teaching foreign languages to deaf individuals in professional education, and giving new bases to researches in the field.

## 1. Linguistic competences in Deaf People: an integration problem

Deaf people officially certified in our country (Italy) are about 60,000, but it is estimated that this number does not reflect the true dimension of the problem. About 11 of every 10,000 children born deaf.

Deafness is a deficit, but not a cognitive one. However, School still offers no effective systematic response to the problem of deaf education. The social cost of this situation are enormous: not only deaf people are often excluded from written communication, as well as from the spoken one; in many cases, they cannot perform professional tasks involving minimum competences in written language and cannot access higher levels of education.

Researches done in this field (Caselli et al., 2007; Fabbretti et al., 2006), reveal that deaf people, especially those whose deafness aroused in pre-linguistic age (before 18-30 months), have typical problems in the acquisition of written language and in the development of linguistic skills. These problems are specific for each culture and each language, and they are not always comparable. In Italian, for example, deaf people show lacks in the use of free morphology, clitic pronouns, prepositions, articles and so on. This means they need tools and educational methods aimed at resolving them. This is often a difficult task, due to the differences in deaf people logopedic rehabilitation and educational paths, and, thus, their different writing skills. Any possible solution has to adapt both to the type (genetic, sickness, etc.) and degree of deafness (deep, medium, light, partial), as well as the learners' specific linguistic and communicational competences and abilities.

In this perspective, the evolution of web technologies towards portability and adaptability to users' needs, and the use of educational strategies based on e-learning tools can forecast an enhancement of the effectiveness of the actions directed to this specific target.

On the user point of view, the new forms of digital communication constitute a horizon of authentic interactions in the national written language (or rather, written/spoken) in which deaf people immerge themselves spontaneously and with strong motivation. This means that, inevitably, through these interactions they acquire language skills.

In short, the use of new technologies in deaf people education configures for the first time a domain in which deaf people with medium/low skills in the written language can improve themselves through the involvement in real communication phenomena and not only through learning contexts. They can thus acquire languages, not only learn them.

## 2. Sign Language as a possible tool for promoting deaf people linguistic competences

The condition, however, is that strategies and tools are to be really oriented on the needs and resources of deaf learners. This is the crucial point of the researches and experimentations achieved so far, and can be divided into a number of critical issues that will be considered in the development of our contribution. Most of the findings here described are based on the experience gained working on the DEAL Project (Deaf people in Europe Acquiring Languages through E-Learning)[1].

In the case of deaf people using sign language[2], the role of it in the didactic communication with and within the students is particularly important as part of promoting the development of skills in the target language. In fact, deaf students using sign language find it particularly comfortable as a language to refer to, putting them in the correct emotional condition to become a learner.

Within the process of building these skills, we have

[1] Please refer to the acknowledgement chapter for further information on the project.

[2] All researches and developments of the project here depicted has considered the micro-culture of deaf people using sign language, to which we will refer, from now on, as "deaf people" or simply "the deaf".

considered sign language as the perfect candidate to be one of the cornerstone resources in the design of all activities concerning the didactic communication: research, problem setting and problem solving, meta-linguistic reflection, metacognitive analysis.

Building the e-learning platform, we have chosen to use sign language in both the interactions among peers and with teachers, integrating the online educational path with videos and explanations in sign language, and the possibility for the students to obtain further information through the video-chat system.

The effective implementation of this strategy has brought up the importance of creating tools specially designed not only to allow sign language interactions regulated according to their purposes, but also to support building of feedback structured on a mosaic of codes. This means not only stimulating the use of sign language, but also creating a feedback system among teachers and learners, as well as between the learners themselves, allowing didactic activities to be really effective. Following what learners are doing, teacher will have the opportunity to intervene with different feedback degrees, tailored on the learners needs.

## 3. Deaf People in Europe Acquiring Language through e-learning: the construction of a specific educational path

The actions forecasted in the DEAL project were meant to significantly operate in this framework, through the introduction of educational tools based on an e-learning strategy, targeting the needs and the specific capacities of deaf adults.

In DEAL e-learning based approach, we enhanced the methodological strategies and educational techniques that allowed the action upon those critical features in lexical and grammatical production indicated by the researches carried out in the field: we worked both on a lexicon level and on the linguistic structures for the development of the language skills of deaf learners through the integration of Sign Language in an educational perspective.

The system is based on the use of an open source e-learning platform (Moodle) and a videoconferencing system based on Openmeetings/Red5. The choice of Moodle has followed that of many European Universities, adopting this platform for their online courses. Opportune adaptations were studied and applied to meet the needs of the target group (teenager students of technical schools for enterprise secretaries).

The applications that have been added are:

- Explanation and introductive videos in the local sign language
- Animated segments with subtitles upon which educational activities has been developed.
- Interactive teaching activities where the tutors can work with the students starting from their questions and their doubts in the educational system. Explanations are thus given from the active interaction with the students and not "from above".
- Videoconference possibility
- Forum

While following the teaching activities, at various set points along the course, deaf students uses special supports in their own sign languages. There are two kind of support:

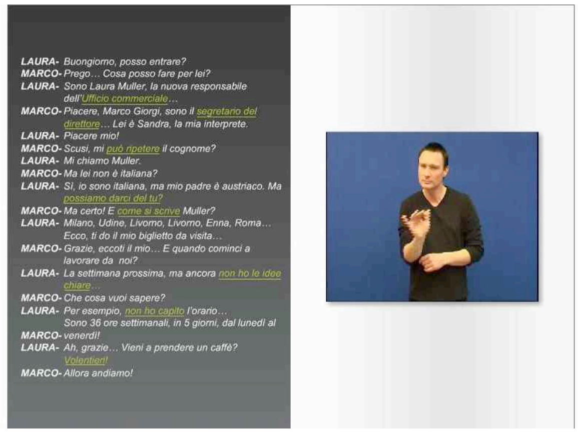
One way:

- Presentation of the teaching unit
- Lexical micro windows on the dialogue
- Grammatical, syntactic and pragmatic support on the key concepts of the unit
- Full translation of the dialogue

Bidirectional:

- Videoconference among peers
- Videoconference with the teaching team

The project has produced three courses: German, Italian and Spanish as second languages for the deaf students of the partner countries. For example: Italian deaf students had a Spanish and a German course available. This means that each course has two sign language to support it: for example, the Italian course has both supporting windows in Catalan Sign Language and in Austrian Sign Language.



LAURA- Buongiorno, posso entrare?
MARCO- Prego… Cosa posso fare per lei?
LAURA- Sono Laura Muller, la nuova responsabile dell'Ufficio commerciale…
MARCO- Piacere, Marco Giorgi, sono il segretario del direttore… Lei è Sandra, la mia interprete.
LAURA- Piacere mio!
MARCO- Scusi, mi può ripetere il cognome?
LAURA- Mi chiamo Muller.
MARCO- Ma lei non è italiana?
LAURA- Sì, io sono italiana, ma mio padre è austriaco. Ma possiamo darci del tu?
MARCO- Ma certo! E come si scrive Muller?
LAURA- Milano, Udine, Livorno, Livorno, Enna, Roma… Ecco, ti do il mio biglietto da visita…
MARCO- Grazie, eccoti il mio… E quando cominci a lavorare da noi?
LAURA- La settimana prossima, ma ancora non ho le idee chiare…
MARCO- Che cosa vuoi sapere?
LAURA- Per esempio, non ho capito l'orario…
Sono 36 ore settimanali, in 5 giorni, dal lunedì al
MARCO- venerdì!
LAURA- Ah, grazie… Vieni a prendere un caffè?
Volentieri!
MARCO- Allora andiamo!

Figure1: example of an Italian comprehension exercise with micro-window explanation in Austrian Sign Language.

An interesting issue in working in such a multilingual environment has been, on several point of view, the lack of human resources having the skills and capacities required from the project: i.e. a tutor capable to sign in Catalan Sign Language to give information about German or Italian language course. This could be an issue to discuss in an international environment, also for the construction of possible professional figures.

## 4. Evaluating the DEAL platform, issues and future develooments

The DEAL project has begun in September 2006 and the main prototype test has been carried out in May 2008 in Italy for the Spanish course. The experimentation took place in the Istituto Statale Superiore Magarotto (ISISS - State High School "Magarotto"). Eight deaf teenagers has participated, all students of a high school for commercial secretaries, of which six have accepted to reply to the final interview. They were all familiar with computers and have never studied Spanish.

The platform has been tested in a blended modality, having a technical support in the classroom as well as a teacher they could ask questions to. The experimentation has also tested both the asynchronous and synchronous interaction modality. During the test, while following the course indications, the students could share their questions both in a Forum (asynchronous modality) or a Videoconference environment (synchronous modality) where the teacher would reply to questions through the help of an interpreter.

The materials used to collect the information coming from the experiments has been: anamnesic questionnaires for teachers, observation checklist filled by the researchers, and a final interview to participant students.

Anamnesic questionnaires for teachers has collected personal data of the participants, information concerning the type of deafness, her familiar situation, and her linguistic competences in Italian and foreign languages, if any, both in vocal or sign language modality.

Observation checklist were filled by 2 researchers per participant, in three sessions of 20 minutes each situated in the beginning, in the middle and in the end of the experimentation. The information collected in this phase being the interaction of the students in the classroom and with the teacher, the chosen linguistic form, and other free observations.

At the end of the test, participants were asked to express their opinion upon the degree and type of knowledge achieved during the course, a comparison with traditional in-class courses, feelings about the interaction with the system as a whole and possible suggestions on how to improve it.

The results have confirmed the validity of the chosen educational methodology, as the participants have confirmed learning something new about Spanish in a more stimulating and fascinating way. Participants liked using the videoconference system as well as the sign language explanatory windows, which has been considered a funny and clear way to achieve knowledge. However, the overall data collected in this phase has revealed the need to improve the overall navigation in the system, making the whole online experience more "friendly".

We believe that a solid evaluation of the platform will come with its use within the deaf community to which the system has been made available on the project website. However, the experimentation has given important information not only for that concerning the methodology to use on an e-learning platform, but also for that concerning the management of language codes and system interfacing.

Not only the educational path needs to be adapted to the e-learning model, but also the quantity and quality of information to give in each step must be managed according to the user's special needs and visual skills, as sight is the only sense in which all the information are conveyed during the interaction with the platform.

## 5. The management of time and space on an e-learning platform for the deaf: the importance of data transmission efficiency

Developing an e-learning platform for the deaf also requires a special attention to the management of time and screen space (Keatin & Miru, 2003).

This has emerged clearly during the experimentation phase of the DEAL project when, for example, giving signed explanations of words or grammatical segments. In cases like the one described here, giving students enough time to pass from the sentence under analysis (written text) to the video/chat is fundamental for both educational and motivational reasons. Teacher, computer screen, (eventual) interpreter, and other students play the role of "educational objects" taking their turn in the construction of sense for the student on both a spatial and linear line. On the spatial line, all "educational objects" must be positioned in order to allow students to return to the selected resource when needed, well localized in space and not undergoing changes. The linear line will be that of "taking turn" in the dialogical relationship among the "educational objects", and the amount of information given.

In a multilingual educational environment, in blended learning, where in-class sessions are completed by sessions with online tutors, this becomes particularly important. The role of the tutor is that of providing further adaptability to the course contents, cut upon the single learners' specific needs. To have the tutor online while developing educational tasks means that every single learner will have the possibility to ask questions about the course content, in a dialogical relationship with the tutor and the other students. Similarly, this feature allows the tutor to monitor the class development in relation to the course contents and to manage the students' community discussion in order to enhance learning in particular fields.

A possible scenario for this case is that of the student being home while the tutor follows her and other students in a separate ambient. Students are given the possibility to follow tutor explanation both on video or written chat.

Deaf students are continuously engaged in following and decoding messages through the only sense of sight. In a context like the one described above, their cognitive resources are thus engaged in processing at least three different codes: text, sign language video and teacher's explanation.

This means that, in the hypothesis of a teacher who is also a sign language speaker, s/he will have to give students enough time to allow sight to complete the video message decoding, eventually integrated with hints given through the written or video chat, think and then reply either in sign language or on written chat, in a distant construction of sense. The depicted situation is furthermore complicated in case of teachers who are non-signers, and the interpreter figure needs to be added.

An incorrect management of these types of interaction could lead to frustration, demotivation and possible abandon of the learning session. This is also the case when working on deaf people writing skills enhancement in the learners' local language (i.e. Italian deaf learner – Italian written language): it is proven that deaf people approach to written language is often affected by the difficulties faced during their linguistic rehabilitation and scholastic path, and the frustration they experience in constructing their writing skills (Fabbretti et al. 2006).

A proper management of screen space and time will impact the emerging relationship between students and teachers and the construction of the learning environment. In fact, while in the case of hearing students speech and sight works contemporarily in the construction of sense and on two different levels (student can watch the screen contents while listening to the teacher's explanation), in the case of deaf students there is only one level to work on, sight, which is engaged in receiving multiple inputs contemporarily. Visual elements in the screen should be managed in order to be highly visible, easy to decode, and giving good navigational cues also for the enhancement of the ongoing interactions in the system.

This great use of video and visual communication tools, makes data transmission quality one of the main issues of e-learning platforms for the deaf. Real-time online video communication such as video-chat for sign language or lip movement are strongly affected by the efficiency of data transmission, as this should be as close as possible to real people movements. Many are, in fact, the cases in which multiple video chats makes communication between deaf people (either bimodal or oralists) nearly impossible, due to the scarce quality of video transmission. This constitutes a strong limit in the development of online educational solutions for deaf people.

As it's possible to understand, a lack of efficiency in video transmission, a poor website visual objects management and a incorrect management of time could end up to a loss in deaf students comprehension of the main topics and their motivation in following the course.

## 6. Conclusions

Being one of the first experiences in Europe trying to teach a foreign language to deaf students through the support of e-learning, DEAL project has focused mainly on the structure of the didactic content, and the use of sign language and short "explanation" windows in a complementary and innovative way, in order to support several type of deaf learners needs. This has challenged other aspects of the educational path, such as the selection of the best technology to use, the design of a correct interface for deaf learners, the combination of multiple communicational channels and the "rhythm" of the ongoing interactions in the system.

One of the points that the DEAL project has aroused is the importance of creating a collaborative network among students and tutors, through the use of an effective and reliable technological support.

In this framework, thus, we need to search the best structure for educational communication with deaf learners and the role given to sign language in the variety of possible codes. This point is strictly related to the interaction regulation (learner/learner, learner/teacher, etc.) and time balancing (synchronous, asynchronous) to grant the maximum efficiency in the learning environment.

One of the results of our researches has been that the educational interaction in video conferences requires a definite number of participant. Basing on the DEAL experience, our hyphotesis is that an optimal number for a smooth interaction could be that of 4 people: i.e. one tutor and 3 students.

However, the problem of a system like this is the regulation of speech turn and the different communicational channels balancing: i.e. video-chat vs. textual chat vs. working area where the student is involved in her educational activity. There is a problem in optimizing sign language as a mean of educational communication in an environment in which the target language remains written and, in multilingual environment, is a foreign language.

The problems we have developed so far are surely strategic with regards to the target group, but they also have a relevance that seems to go beyond this specific scenario. In a "regular" educational environment, there are issues that are normally underrated due to the redundancy of communicational possibilities between hearing people who are able to pick up the information they need from the ongoing communicational process. Working on a multilingual platform for deaf people education has thus opened reflection not only on the specific problems that this type of user could meet but have also given a base for reflection on the nature of educational communication in foreign language learning. In fact, these problems shows that the educational communication in e-learning environments shows inefficiency margins, amplified but not generated by deafness. Working towards the solution of these issues can thus have important theoretical implications also in the frame of second language education in digital learning environments.

## 7. Acknowledgements

The DEAL project has been financed by the European Union within Leonardo's programme, and immediately received the European Label 2008 for Innovative Projects in Language Teaching and Learning [www.deal-leonardo.eu]. Partners of the project has been: Istituto Statale per Sordi di Roma - ISSR, Istituto di Scienze e Tecnologie del CNR - ISTC, Istituto Superiore di Istruzione Specializzata per Sordi Magarotto – ISISS [Rome, Italy], Universitat de Barcelona, Fundaciò del Centre d'Estudios de Llengua de Signes Catalana [Barcelona, Spain], Klagenfurt Universitat [Austria].

The project has been re-funded in the frame of Leonardo's Transfer of Innovation programme 2009-2012 that sees the University College London as a new partner of the

project, in place of Klagenfurt Universitat.

The ISSR has recently begun working on a project for the improvement of deaf people Italian writing skills through e-learning (VISEL).

Both authors are in complete agreement for that concerning the paper's contents. Main contributor for chapters 2,3 and 4 has been Dott. Giuseppe Nuccetelli while Dott. Maria Tagarelli De Monte is to consider the main contributor for chapter 1, 5,6 and 7.

## 8. References

Keatin, E.G., Miru, G.S. (2003). American sign language in virtual space: Interactions between deaf users of computer-mediated video communication and the impact of technology on language practices. *Language in Society*, 32(05):693-714.

Elsendoorn, B.A.G., Coninx, F. (1993), Interactive Learning Technology for the Deaf, *Proceedings of the NATO Advanced Research Workshop on Interactive Learning Technology for the Deaf.* The Netherlands, NATO ASI Series, Computer and Systems Sciences, 13(113): p. 285.

Fabbretti, D., Volterra, V., Pontecorvo, C. (1998). Written language abilities in deaf italians. *Journal of Deaf Studies and Deaf Education*, 3(3):231--244.

Pizzuto, E., Caselli, M. C., Volterra, V. (2000). Language, cognition, and deafness. *Seminars in Hearing*, 21(04):343--358.

Rinaldi, P., Caselli, C. (2009). Lexical and grammatical abilities in deaf italian preschoolers: The role of duration of formal language experience. *Journal of Deaf Studies and Deaf Education*, 14(1):63--75.

Maragna, S., Nuccetelli, G (2008). An e-learning model for deaf people's linguistic training. *Proceedings of the DEAL project final meeting*. Publicacions I Edicions de la Universitat de Barcelona.

Fabbretti, D., Tomasuolo E. (2006). Scrittura e sordità. Roma: Carocci Editore S.p.A.

# BONy: a knowledge centric collaborative learning platform

**Alfio Massimiliano Gliozzo, Concetto Elvio Bonafede, Aldo Gangemi**

*STLab-ISTC-CNR
Via Nomentana 56, 00161, Rome, Italy
alfio.gliozzo@istc.cnr.it, ingc.bonafede@gmail.com, aldo.gangemi@cnr.it

## Abstract

In this paper we describe BONy, a technology enhanced platform for collaborative learning. Semantic technology, and in particular an RDF/OWL ontology, is used to integrate different modules of the system, allowing strong interoperability between linguistic data and structured knowledge. This allows us to develop intelligent advanced functionalities, including expert finding, mentoring and semantic search. Those functionalities largely exceed the capabilities of existing state of the art e-Learning platforms, for example allowing multilingual search. BONy is an unique showcase for the next generation semantic systems for e-Learning. The BONy platform is currently working as a free on-line service.

## 1. Introduction

Electronic learning (e-learning) is a type of education where the medium of instruction is computer technology. It is a planned teaching/learning experience using a wide spectrum of technologies, mainly internet based, to reach learners at a distance. The base units of e-learning systems are called learning objects. They are resources, usually digital and web-based such as HTML pages or animations, that can be used and re-used to support learning. They represent an atomic piece of knowledge and are composed into courses. At their core there will be instructional content, practice, and assessment. The way in which the units can be stored, retrieved and managed has been the focal point of most Learning Content Management Systems (LCMS).

The actual mechanisms to manage the learning objects, mainly based on web standards such as XML, is not able to face the new requirements of collaborative learning, where teachers and users are no longer two different players in the network. In fact, in a web2.0 perspective, students are asked to supervise other students and are supposed to actively contribute to the development of learning objects, playing the role of professors with respect to the areas of expertise where theirs skills are higher. In addition, in a collaborative learning scenario, the student is typically exposed to a very highly unstructured information (e.g. wikis developed by other students, forums, chats), requiring the intervention of a professor or an expert in the field to recommend a personalized learning path and to ensure the selection of high quality content.

On the other hand, non-semantic technology, such as web 2.0 platforms, do not allow us to implement a fully automatic system satisfying the new needs of collaborative learning, and in particular to represent the user profile and assess his skill. To this aim, semantic technology such as ontologies can play a big role, for example to represent the user profile with respect to different subjects and to represent the content of learning objects. To this purpose, within the BONy project, we looked forward to semantic technologies, anticipating the next generation WEB 3.0 solutions for eLearning while providing a showcase of the new generation capabilities.

BONy is a knowledge centric LCMS where a core ontology is used for two main purposes:

1. enhance interoperability and system integration
2. integrating linguistic information from learning objects with structured information from databases
3. allow intelligent services such as expert finding, mentoring and semantic search

The core component of the system is a ”RDF/OWL” ontology, developed according to the best practices and by applying Ontology Design Patterns (Gangemi, 2005; Presutti et al., 2008; Reich, 1999; Svatek, 2004). As far as interoperability and system integration are concerned, the ontology is used to enhance the integration of three existing open source platforms: a LCMS (DOKEOS) (Grandmontagne., 2008; **?**), a framework for social networking (SPREE) (Bauckhage et al., 2007; Metze et al., 2007) and a collaborative authoring tool (Semantic Media Wiki) . The ontology is automatically populated by re-engineering data from the different databases exploited by the three platforms integrated so far.

A mayor role of the ontology is linking the textual data to the knowledge structures. This is done by extracting keywords from the text embedded in the Learning Objects, and associating different keywords to a set of topics of interest for the domain of the course. This allows us to map different courses in different languages to the same topic structure, and to improve search and multilingual retrieval. This allows us to implement a set of intelligent functionalities, including an automatic mentoring algorithm designed for the generation of personalized learning paths, multilingual search and expert finding. To this aim, we connect Learning Objects and user profiles with a shared taxonomy of topics describing the content of the e-Learning course, and we used SPARQL queries and a reasoner. This has been done by extracting keywords for each course

The platform is currently working as an free on-line service, available on the web at the address social.bonynetwork.eu . We invite the reader to join the BONy network and feel the different user experience provided by semantic technology in use.

This paper is structured as follows: in section 2. we illustrate the architecture of the platform, section 3. is devoted to describe the ontology used in system, in section 4. we describe the intelligent functionalities of BONy. Section 5. concludes the paper.

## 2. Architecture

The BONy platform is an integration of three existing open source platforms: DOKEOS, SPREE and Semantic Media Wiki.

The architecture of the system is described in Figure 1: an RDF/OWL ontology is used to represent data coming from the different databases adopted by the integrated open source solutions. The ontology describes semantically the three main components of the platform, and in particular the Learning Objects, the European Project Management domain (Topic ontology) and the user profile in the social network (User Ontology) as in Figure 1. Differently from other e-Learning system, data is represented in the ontology in RDF/OWL format.

In addition, when data is represented into the ontology, it is also linked semantically to a topic ontology, describing the content of the course. In particular, user profiles and learning objects are linked together across topics. The richer expressivity of this formalism allows us to develop semantic functionalities such as user profiling, learning path generation and expert finding.

Thanks to the ontology it is possible to enhance the consistency of the inserted data. This is done by using a reasoner to check the consistency of the entire database every time new data is inserted.

The technology adopted to represent and manage the data in the ontology is based on state-of-the-art Java open source solutions: Jena[1], Pellet 2(Sirin et al., 2006), and Protégé [2]. We used protege to build the ontology, Jena to access the ontology and Pellet to reason on the data. The access to the ontology from the various sub-system is implement by adopting a client/server architecture developed in Java.

## 3. The BONy Ontology

The development of the BONy ontology has been inspired by the following principles:

- re-usability: when adapted to a new course, the OWL schema of the BONy ontology is preserved and only RDF data change, it allows us to minimize the adaptation costs to new domains;
- modularity: the ontology is composed by three modules representing the eLearning content, the social networks and the topics of the course, it allows us to change the course while preserving the community;
- best practices: the ontology has been designed by specializing Ontology Design Patters (ODP)

Regarding *re-usability*, we carefully distinguished the OWL part of the ontology (i.e. the metamodel) from the actual data. In this way, the platform can be adapted to new communities, learning Objects and topics without any change in the ontology. To this aim, the topic taxonomy has been reified, so that topics are instances and not classes.

To allow *modularity*, the ontology has been subdivided into three main components (see Figure 1):

- **Topic Ontology:** it describes the subjects covered by the eLearning course and their conceptual dependencies. Topics are instances of the class TOPIC, and they are related between each other by the object properties isSubTopicOf and nearTopicTo connecting different instances of the same class.
- **eLearning Ontology:** it is about the learning objects and describes different features, e.g. the type of electronic support adopted, dependencies between learning objects and the time required for learning. This part of ontology is composed by different classes such as: LearningActivity, SCO and CourseRole. The instances of those classes and their relations have been mostly derived from the corresponding SCORM descriptions by a reengineering process.
- **User Ontology:** it is about the social network players, representing students and teachers' profiles, their relationships and their skills. All the users in the network are represented by instances of the class AGENT. Specific subclasses are STUDENT, TEACHER and EXAMINER

The topic ontology operates as a link between the eLearning ontology and the user ontology. For example, users and SCOs can share a relation with a common topic, allowing the development of recommending services and the automatic assessment of the user profile.

Users are linked to topics by the knowsTopic relation, reflecting their skills into 5 specific degrees: knowsMediocre, knowsBasic, knowsFair, knowsGood and knowsPerfect. In a similar way, Learning Objects are linked to Topics by the relation hasTopic. This is derived by the keywords annotation performed on the learning objects and represented in the ontology as well. In fact, keywords are linked to topics, allowing to infer the has topic relation between topics and learning objects.

Our ontology is developed according to the Ontology Design Pattern (Gangemi, 2005) (ODP) paradigm, i.e. utilizing and specializing some already existing reusable ontology to describe particular piece of domain knowledge. An ODP is usually a small ontology that solves complex modelling issues to enhance semantic interoperability of different knowledge components. The notion of ODP was introduced in 1999 for a particular problem domain in biology (Reich, 1999). Afterwards, ODP appeared under different names such as semantic patterns, knowledge patterns and the designing patterns for Semantic Web ontologies that are now called ODPs. A large repository of ODP is available on line[3].

### 3.1. Populating the ontology

The ontology is populated by re-engineering data coming from different databases belonging to different applications, and in particular: a) from the e-Learning course (described by the Manifest file in the SCORM syntax) b)

[1] jena.sourceforge.net

[2] protege.stanford.edu

[3] `http://www.ontologydesignpatterns.org`

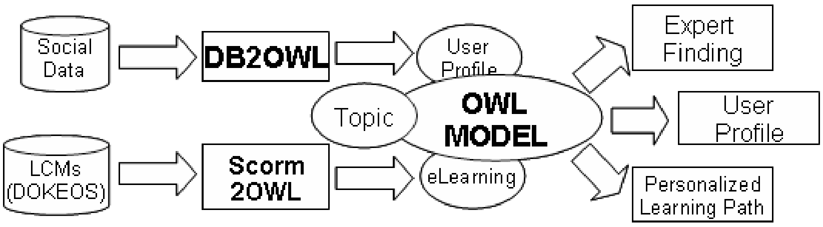


Figure 1: Services and data are linked by the BONy ontology

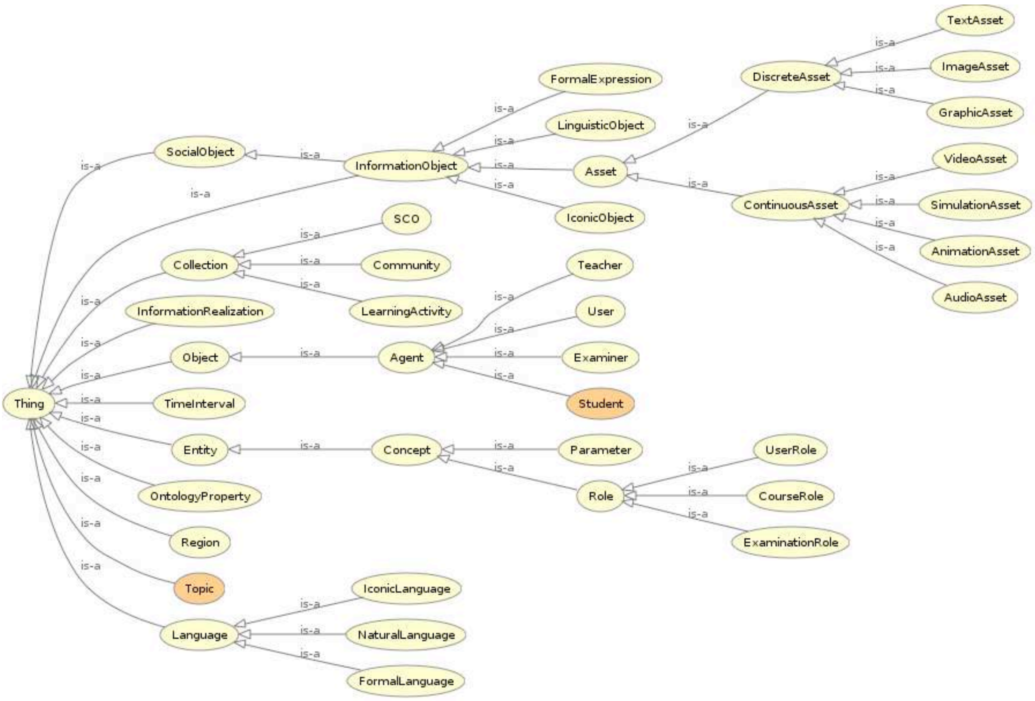


Figure 2: The Bony Ontology: concept hierarchy

from the user profiles in the LCMS and in the Social Network (see Figure 1). In addition, the Topic ontology has been manually populated by topics of interest for the domain of the course and their relationships.

**Populating the Topic Ontology** The ontology class Topic is one of the most important classes. It allows us to link users and learning objects. In our Ontology we use a semiotic notion of Topic as a (usually potential) collection of SocialObject(s). For example, *Project management* is a topic constituted by the set of social objects that are associated with project-management related entities, such as *tasks* and *deliverables*.

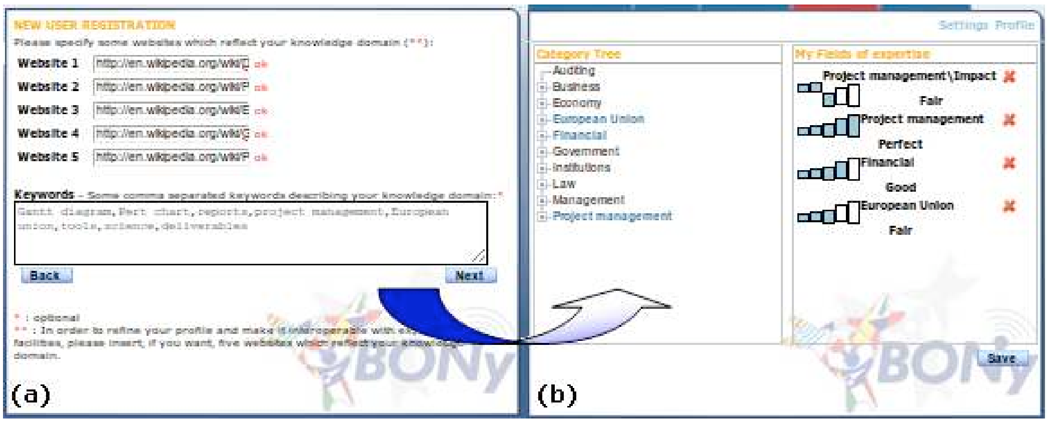
Topics are related each other by Narrower and Broader relations. The procedure adopted to build the topic ontology was entirely manual, but at the same time inspired by quantitative principles aimed at preserving a pretty uniform distribution of learning object for each topic.

To achieve this, we first selected a set of keywords describing each learning object, then we look for their corresponding pages in wikipedia, in order to find their corresponding category. We select those categories as topic after manual revision, and we browse the narrower/broader relationships among them to figure out a meaningful taxonomy a meaningful taxonomy describing the project management domain.

**Mentoring**

**Populating the eLearning Ontology** To populate the eLearnign ontology we re-engineered data from SCORM to RDF following the metamodel developed for the e-Learning ontology, which basically reflected the SCORM distinctions. To this aim, we represent some of the relevant distinctions in the SCORM definition into properties of the ontology. This process is totally automatically and is performed once the course is loaded into the platform. This is done partially by re-engineering the XML based metadata in the SCORM manifest file. In order to connect the Leaning objects to the topic ontology we exploited the keyword annotation developed to build the topic taxonomy and we inferred the relations between topics and learning objects if one or more keyword is associated to both. This is done by a CONSTRUCT query in the SPARQL language.

**Populating the User Ontology** The user data are derived from a variety of different systems integrated in the BONy platform. The ontology allows us to integrate different frameworks such as DOKEOS (where personal data are



NEW USER REGISTRATION

Please specify some websites which reflect your knowledge domain (**):

Website 1 http://en.wikipedia.org/wiki/[ ok

Website 2 http://en.wikipedia.org/wiki/P ok

Website 3 http://en.wikipedia.org/wiki/E ok

Website 4 http://en.wikipedia.org/wiki/G ok

Website 5 http://en.wikipedia.org/wiki/P ok

Keywords - Some comma separated keywords describing your knowledge domain:*

Gantt diagram,Pert chart,reports,project management,European union,tools,science,deliverables

* : optional

** : In order to refine your profile and make it interoperable with ex facilities, please insert, if you want, five websites which reflect your knowledge domain.

Category Tree

- Auditing
- Business
- Economy
- European Union
- Financial
- Government
- Institutions
- Law
- Management
- Project management

My Fields of expertise

| | |
|---|---|
| Project management\Impact | Fair |
| Project management | Perfect |
| Financial | Good |
| European Union | Fair |

Figure 3: Interface for user profiling in the BONy platform

collected in a database) and SPREE (a open source knowledge exchange network) where the data about the knowhow of the single user are registered. Relations between Users and Topics are first established at registration time by the user profiling module, and then refined by the user at any time. Synchronization between the user profile in the Social Network and the ontology is guaranteed by updating the ontology at every change.

## 4. Intelligent Functionalities of the BONy platform

The BONy platform provides three main semantic services which are far behind the capabilities of current eLearning technology: mentoring, (i.e. the generation of personalized learning path within the course on the basis of the user profile), semantic expert finding (i.e. looking for experts within the network which are able to answer to specific questions) and multilingual search (i.e. the capability of retrieving Learning Objects in any of the 11 different languages of the BONy course). Mentoring and expert finding are based on the user profile, automatically inferred by the platform and represented in the ontology.

Even though some of them have been already proposed in the literature, BONy is the first working platform implementing all them at the same time in a integrated environment, thanks to the massive use of semantic technology.

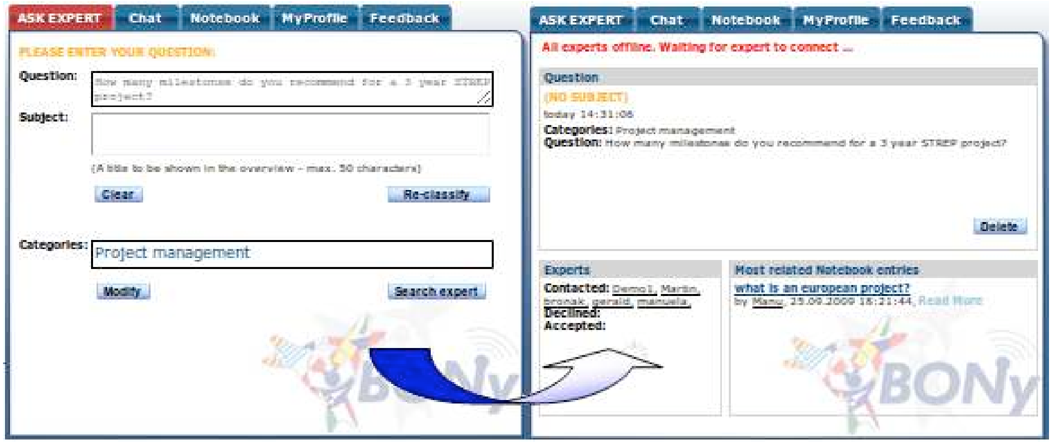
**User profiling** The user profile is represented in the ontology and consists of biographic data, such as email, name, address, as well the assessment of the user skills. The user skills are represented by relations between them and topics in the ontology, as described in section 3.. The BONy platform is able to assess the competence of each users in a semi-automatic way, by looking at web pages and other content indicated by the user as a reference material for his competence. This process is easy, quick and effective, and works as follows.

Every time a new user is enrolled in the system, she/he is asked to enter a set of web pages describing her/his skills (e.g. her/his home page, the home page of her/his university or organization). In addition, she/he is asked to enter a set of keywords describing her/his skills. This is illustrated in the the left part of Figure 3.

Then, the BONy platform uses Information Retrieval and Natural Language Processing techniques to match the content described so far with the topic ontology, in order to establish new relations between her/his profile and topic ontology. To this aim, we exploit one of the core capabilities provided by the SPREE framework, which is able to crawl specified sites, extracting bag-of-words, and therefore representing each page in a vector space model, and then measuring the similarity among vectors associated to users and those associated to topics in the ontology by cosine similarity. To this aim, the SPREE platform generate bag of word vectors for each topic in the ontology when it is installed by adopting a very similar approach to what described for the user (Bauckhage et al., 2007; Wetzker et al., 2007). The result of this process is a preliminary assessment of the user skills that can be further refined by the user itself by adding new topics or modifying the degree of relevance of each category, ranging from basic to perfect. This is illustrated by right part of Figure 3. The user model obtained so far is then stored in the ontology while checking the logical consistence.

The aim of this service is to recommend a minimal sequence of learning objects to a new user on the basis of his profile. This set will be generated automatically by an algorithm whose goal is to select a sequence of learning objects so that the user is not studying subjects he is already aware of, while concentrating on filling the gap between her/his initial user skills (i.e. those inferred by the user profiling module described in the previous section) and the full range of topics covered by the course. The goal of this process is to minimize the time required to study the full topics of the course while avoiding subjects already well known, while taking into account dependencies between learning objects.

The output of this process is illustrated in Figure 5. Clicking on the ”yes, I would like to try”, the automatic mentoring process starts and after a few seconds returns the the sequence of learning objects where a subset has been marked by a green sign (see righter part of Figure 5), meaning that the student does not need to go trough them since he is already skilled in the subject. The effect of this process is that the system generates a minimal set of learning objects, avoiding the student to go thought the full course, which



PLEASE ENTER YOUR QUESTION:

Question: How many milestones do you recommend for a 3 year STREP project?

Subject:

(A title to be shown in the overview - max. 50 characters)

Categories: Project management

All experts offline. Waiting for expert to connect ...

Question

(NO SUBJECT)

today 14:31:06

Categories: Project management

Question: How many milestones do you recommend for a 3 year STREP project?

Experts

Contacted: Demo1, Martin, bronak, gerald, manuela,

Declined:

Accepted:

Most related Notebook entries

what is an european project?

by Manu, 25.09.2009 18:21:44, Read More

Figure 4: Expert finding process. When a question is submitted, the system categorizes the question (categories box on the left figure) and searches the experts (Experts box on the right).

will take around 5 hours in the European Project Management case study. Rather he is supposed to study less, saving time (about 1 hour in the example in Figure 5)

To implement this service, a typical approach in Artificial Intelligence is to use a planner. Given the reduced number of constraints and the relatively small scale domain, it was possible to implement the same set of capabilities in a much simpler way by defining ad-hoc SPARQL queries and using a reasoner. This generates a planner that is different from those using a rigorous logical formalism and a clear definition of goals. Instead using SPARQL we can make an approximation because the objective is not formalized. In fact, each learning object is linked to one or more Topics. This allows us to link the user profile (degree of knowledge in the different Topics) with the learning objects regarding topics he knows better. A simple SPARQL query allows us to select all those Learning Objects about topics that are not in the user profile, generating the mentoring service we are interested in. This service is implemented by adopting the Jena API to perform the SPARQL queries and Pellet 2 to reason on the data.

**Expert finding** The role of the expert finding service is to look for other students in the network which are able to answer a specific question. Every user is regarded as a possible expert on the Topics where is user profile has stronger association. BONy is able to look for suitable experts by simply classifying questions with respect to the topics in the ontology, which is the same adopted to represent the user skills.

To this aim, a bag of words for each topic in the ontology is retrieved from on-line or off-line resources. The same process is done to describe the user profile. Then each expert is mapped to one or more topics by using similarity metrics (Bauckhage et al., 2007; Wetzker et al., 2007).

Every time a new question is submitted, the system classifies it with respect to the topic ontology. Classification is used together with a similarity measure between the query and expert profiles in order to select the first five top scored experts.

The question is then automatically sent to them by email. The answers collected so far are then stored in a public forum and can be ranked by using a feedback mechanism, in order to assess the reputation of users in the network and to promote new experts for forthcoming questions.

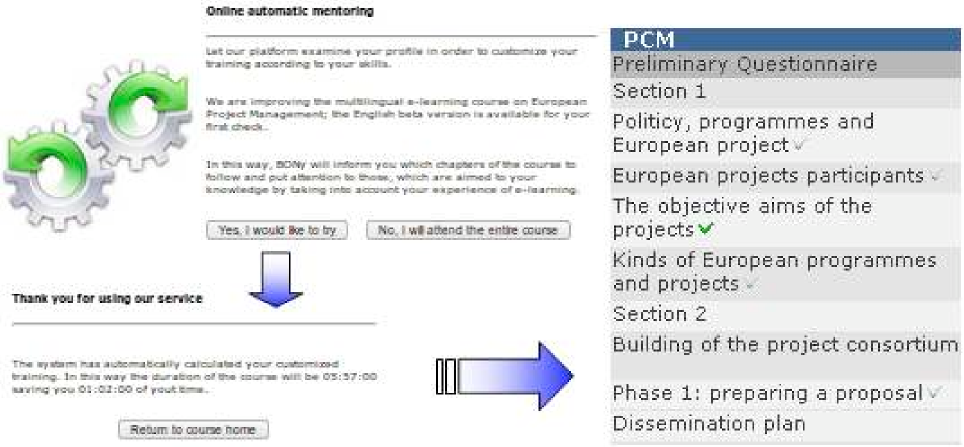
Figure 4 presents an example of the expert finding process, showing the categories of the question and the retrieved users.

**Multilingual Search** All learning objects and their textual content have been indexed by a search engine (i.e. Lucene). The index is done by using the text within the slides and the keywords associated to each of them. As far as the learning objects are aligned among languages by a common representation in the ontology, it is possible to write queries in any language, and to retrieve pages in different languages. Expanding text in learning objects by the keywords in the ontology is also a way to implement semantic search. Figure 6 describes a screen-shot of the search engine and his multilingual capabilities.

## 5. Conclusion and future work

In this paper we presented BONy, a technology enhanced platform for collaborative learning using semantic technology to enhance interoperability between systems and to allow advanced functionalities such as including expert finding, mentoring and multilingual search. Those functionalities largely exceed the capabilities of existing state of the art e-Learning platforms. BONy is an unique showcase for the next generation semantic systems for e-Learning and can be used on line at the address social.bonynetwork.eu .

The main focus of our work has been showing the new capabilities allowed by connecting linguistic data with knowledge bases, how to represent this information into a proper knowledge base and how to make it interoperable with linked data in the semantic web. Therefore we did not concentrated in boosting the performances of the single components, for example by using richer ontologies or more advanced Natural Language Processing techniques. In the future, we are going to develop the 3.0 version of the BONy platform, where semantic web data will play a big role to shift from an information to a knowledge centric system. In particular, we are going to implement a knowledge centric authoring tool for learning objects, where semantic web



Online automatic mentoring

Let our platform examine your profile in order to customize your training according to your skills.

We are improving the multilingual e-learning course on European Project Management; the English beta version is available for your first check.

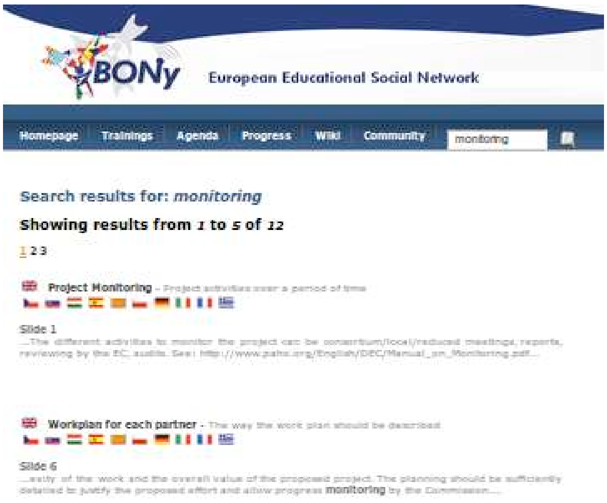
In this way, BONy will inform you which chapters of the course to follow and put attention to those, which are aimed to your knowledge by taking into account your experience of e-learning.

Thank you for using our service

The system has automatically calculated your customized training. In this way the duration of the course will be 05:57:00 saving you 01:02:00 of yout time.

PCM
Preliminary Questionnaire
Section 1
Politicy, programmes and European project ✓
European projects participants ✓
The objective aims of the projects ✓
Kinds of European programmes and projects ✓
Section 2
Building of the project consortium
Phase 1: preparing a proposal ✓
Dissemination plan

Figure 5: Output of the mentoring process



European Educational Social Network

Search results for: *monitoring*

Showing results from 1 to 5 of 12

1 2 3

Project Monitoring - Project activities over a period of time

Slide 1
...The different activities to monitor the project can be consortium/local/reduced meetings, reports, reviewing by the EC, audits. See: http://www.paho.org/English/DEC/Manual_on_Monitoring.pdf...

Workplan for each partner - The way the work plan should be described

Slide 6
...ality of the work and the overall value of the proposed project. The planning should be sufficiently detailed to justify the proposed effort and allow progress monitoring by the Commission...

Figure 6: Fulll text multilingual search inside the eLearning content

data are composed by Ontology Design Patterns specialized on the subject of interest for the course, we are going to exploit agent based technologies for advanced tutoring and mentoring, we are going to replace the retrieval engine with a more powerful recommending engine, looking for semantic web data as well as for internal repositories of learning objects. Last but not least, we are going to explore the potentiality of applying advanced NLP tools for information extraction from text and to link the extracted information to dictionaries like wordnet and other linguistic resources available from the collaborative work, such as wikitionaries and DBpedia.

## Acknowledgments

This work has been supported by the BONy project, financed by the Education and culture DG of the EU, grant agreement N 135263-2007-IT-KA3-KA3MP, under the Lifelong Learning Programme 2007 managed by EACEA.

## 6. References

C. Bauckhage, T. Alpcan, S. Agarwal, F. Metze, R. Wetzker, M. Ilic, and S. Albayrak. 2007. An intelligent knowledge sharing system for web communities. In *In IEEE Int. Conf. on Systems, Man, and Cybernetics*, Montreal, Canada.

A. Gangemi. 2005. Ontology design patterns for semantic web content. In *Proceedings of the ISWC 2005*, volume 1729 of *Lecture Notes in Computer Science (LCNS)*.

Y. Grandmontagne. 2008. Technical report, DOKEOS. Available via http://www.dokeos.com/en/press.

F. Metze, C. Bauckhage, T. Alpcan, K. Dobbrott, and C. Clemens. 2007. A community based expert finding system. In *Proceedings of IEEE Int. Conf. on Semantic Computing*., Irvine, CA.

V. Presutti, A. Gangemi, S. David, G. A. de Cea, M. C. S., E. Montiel-Ponsoda, and M. Poveda. 2008. Deliverable 2.5.1: A library of ontology design patterns: reusable solutions for collaborative design of networked ontologies. Deliverable Project Number IST-2005-027595, NeOn: Lifecycle Support for Networked Ontologies.

J. R. Reich. 1999. Ontological design patterns for the integration of molecular biological information. In *Proceedings of the GCB'99*.

E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, and Y. Katz. 2006. Pellet: A practical owl-dl reasoner. Technical report, Maryland information and network dynamics lab semantic web agent project.

V. Svatek. 2004. Design patterns for semantic web ontologies: Motivation and discussion. In *Proceedings of the 7th Conf. on Business Information Systems*., Poland.

R. Wetzker, T. Alpcan, C. Bauckhage, W. Umbrath, and S. Albayrak. 2007. An unsupervised hierarchical approach to document categorization.. In *IEEE Intl. Conf. on Web Intelligence (WI'07)*, Silicon Valley, USA.

# Social E-SPACES: socio-collaborative spaces within Virtual Worlds

Vanessa Camilleri Matthew Montebello
University of Malta
Malta
E-mail: vanessa.camilleri@um.edu.mt, matthew.montebello@um.edu.mt

## Abstract

This paper presents research based on a current study validating the effectiveness of the teaching and learning process in the context of virtual spaces. A report about teens and social Media (Lenhart, Madden, Rankin Macgill, & Smith, 2007) reveals that 93% of the teens who were interviewed use the Internet as a social meeting place.  This, coupled with recent internet usage statistics, establishes 'digital natives' as active participants in the design of new media as social collaborative tools. Would these social tools be effective in the e-learning context or will they form part of a wider knowledge management framework? The purpose of this study is to outline the design of the measurement of interaction processes in the virtual spaces used for e-learning.

## 1. Introduction

Tiropanis, et al., (2009) discuss how the level of adoption and use of tools and services within the higher education sector in the UK associated with teaching and learning, are various. In addition to these tools and services in the form of Web2.0 applications, or Learning Management Systems (LMS), a number of educational institutions, make use of Virtual Worlds (VWs) for various learning activities (NMC, 2010). These learning activities take the form of seminars or tutorials or simulations as well as other problem solving exercises, engaging learners in their knowledge building process  (Wrzesien & Alcaniz Raya, 2010). More learning activities are described in Petrakou (2009) who illustrates in detail the scope of having a specific virtual environment for the facilitation of transmission of the online content whilst Kumar et al. (2008) portray the VW as being a social environment which holds computer-based simulations which users can make use of without any pre-defined objectives but which yet assimilates groups of people together through an expression of interest. Carey (2007) argues that VWs are intended to be immersive social experiences which not only offer alternatives to face to face interactions but which can also provide new forms of human experiences, built upon a vast array of communication tools which can offer the same emotional satisfaction as gathered from the social exchanges happening on the daily basis. This of course is discussed within the context of the online environment which is discussed extensively in (Dillenbourg, Lehtinen et al., and Slavin in Petrakou, (2009)). These describe how the transition towards the migration of learning content to the online environment is further assisted by a number of interaction processes within collaborative learning. The latter, being one of the pillars of the design for e-learning systems contributes to the construction of new concepts, collectively brought together through communities, most often, established by dialogical interaction (Etelapelta & Lahti, 2008).  The premise of this study is built around learning theories which adopt the socio-constructivist approach (Vygotsky, 1978) describing knowledge construction through interaction processes, rather than knowledge acquisition. The socio constructivist approach in this scenario describes learning as a collaborative meaning-making experience where learners participate in a number of interaction processes which facilitate the learning process. The interactions between learning communities, as well as individuals within the learning communities, as has been argued by Alier (2006), in essence would enforce the reason for existence of Virtual Worlds (VWs) transforming gaming into *serious gaming,* breeding *social communities of practice* (CoP) which eventually develop into *learning communities*.

The scope of this study is to create the framework for the measurement criteria assessing the validity of VWs for the teaching-learning process using human-behaviour parameters.  The rest of the paper is structured as follows: Section 2 gives a brief overview of the insights into e-learning perspectives, whereas Section 3 highlights the pedagogical value of collaborative spaces. Section 4 has a look at established pedagogies which can be implemented in VWs whereas Section 5 and Section 6 propose a design and framework parameters for E-SPACES. Section 7 looks at future developments of the framework for the measurement and validity of the effectiveness of social collaborative process in the VWs.

## 2. E-learning Perspectives

Over the years the use of ICT in education has shifted from mere *Computer Based Learning* (CBL), making use of software as a means of knowledge transmission, to *Computer Enhanced Learning* (CEL), which aims to improve the environment for creative knowledge practice. Studies have in fact shown that merely pushing content online is not returning the results expected (Spalter & Simpson, 2000).

Solimeno et al (2008) show, how up until the late 1980's the first models of computer supported education, put the learner as a solo-user creating an isolated niche where the promotion of learning at one's own individual pace was highlighted. However more recently due to the social networking boom, researchers have been looking at a more advanced form of computer supported collaborative learning, as an additional enhancement to the online teaching model. Such a derivation of the CEL is based upon constructivist learning theories which focus primarily on the social interdependence as affecting the learning process. This in fact has given rise to a new evolution to the use of learning management systems, which in addition to providing content, are also providing some means of online interactivity, paving the way for social interactions as a means of constructing knowledge concepts. Today's e-learning paradigm has shown an evolution from Learning Management Systems (LMS) where the scope is that of utilising the web as a pipe, there only to deliver content, to a meeting point, a place where to hang out with others in specific CoP.

Brophy (in Paechter, Maier, & Macher, 2010) propose five fields of instruction as being core components of e-learning design. These include the course design and the electronic environment, the interaction between students and instructors, the interactions among peers, the individual learning process and the course outcomes. The interactions and processes will be discussed in more detail in Section 4. In addition to interactions, Granic, Mifsud, & Cukusic (2009) further propose that clear pedagogical objectives based on sound pedagogic principles need to be incorporated within the e-learning design for more effectiveness within the teaching learning process to be achieved. The pedagogic approach chosen by the authors for their study is built around the concept of "active learning" with core components which include aspects of constructivism, blended learning and collaborative learning. Engaging and further motivating the learner for a more active involvement using Kolb's experiential learning theories (Kolb, 1984) is one of the basic pedagogic principles which will be adopted in this study. This then leads to the development of a socio-constructivist model for e-learning which will be used to enhance deeper conceptual thinking.

## 3. Collaborative Spaces

Having established that research trends in pedagogies applied to the online learning environment point towards the setting up of communities for collaborative constructivist models, this research proposes to determine the parameters around which such communities are built. Miller & Brunner, (2008) make use of the Social Impact Theory (SIT) (Latane in Miller & Brunner, 2008) to understand how learners' interpersonal characteristics affect peers during collaborative learning experiences. This theory is described as changes in an individual's behaviour, resultant from communication exchanges with 'perceived' and real indviduals. The concept of the perceived peers can be made use of within the virtual world ecosystem, an environment designed and built as a collaborative space. The online environment in itself has been indicated as being more of a support and a supplement to face to face interaction. Research (Tomai, Rosa, Mebane, A, Benedetti, & Francescato, 2010) has shown that the development of online communities and social networks contributes to a possible increase in the social capital for each individual within the group. The social capital is defined as a pool of resources which an individual can accumulate as a result of developed interrelationships. The parameters within which learning communities are assessed include:

- A measure of learners' satisfaction during learning;
- Characterisation of interpersonal relationships during collaborative practice;
- Peer support, indicating connectivity throughout the experience;
- Change in behaviour owing to the social capital constructed.

These parameters will be taken into account when designing the framework for measuring the effectiveness of virtual worlds for knowledge building activities.

## 4. Virtual Pedagogies?

Camilleri & Montebello, (2008) propose a virtual assistant within the social context of the VWs which not only aim to assist and aid in cooperative knowledge building, but which can learn and sustain a mentally stimulating interactive conversation – a two-way communication which finds its roots in every social networking application. Online users' requirements are considered quite distinct however those resident in VWs have unique needs. These include:

- Maintaining student engagement;
- Developing a community;
- Providing immediate feedback;
- Similar learning opportunities;
- Hands-on interactive activities;
- Student ⇔ content interaction;
- Faculty ⇔ student interaction;
- Student ⇔ student interaction.

Furthermore the authors identify a number of strategies which when implemented within the virtual world space contribute to the creation of a 'Learnscape' – a learning environment within the VWs which is built upon:

Flow in balancing inactivity and challenge.

Repetition allowing learners to repeat experiments until they are satisfied with the outcomes.

Experimentation in encouraging learners to try and learn in the process.

Experience which is more engaging than other digitally mediated technologies.

Doing through practice.

Observing through an essential communication platform.

Motivation stimulated by the people's own active part.

Virtual pedagogies are also designed around different approaches and perspectives. Bonanno (2008) in his discussion of learning through collaborative gaming: a process-oriented pedagogy, comes up with a new model which derives its inspiration from "connectionist" and "constructivist" perspectives and which serves the purpose of analysing different "categories of interactions and the major factors that influence them during collaborative gaming".

Monahan & Bertolotto, (2008) describe the transition to the Virtual Reality (VR) environment as one in which the shift is from the 'conventional text-based' to the immersive and intuitive one, where the computer simulates the natural environment thus making it easier for the learner to identify with. The project Virtual European Schools (VES) – (Bouras in Monahan, G, & Bertolotto, 2008) simulates a collaborative learning environment within virtual classrooms themed around specific school subjects. This project has achieved a high level of user satisfaction highlighting the social presence as a 'major advantage'. The authors describe CLEV-R, a 3D learning environment which proposes the social interaction between learners as one of the most important elements which is exploited from its marked absence in other conventional, text based learning systems. E-SPACES will attempt to build upon this research, by making use of pedagogies and principles which have already be ascertained to bring about a change in learner behaviour, within VWs. These will be used to create a measurable standard for the effectiveness of VWs on learning, using distinct parameters for integrating human complex behaviour in community-based learning.

## 5. Proposed Design

One very important component of this study is the implementation of a virtual space, included in which are the key elements which would enable users to experiment with their own learning and interact with each other in a collaborative environment, in a persistent space, facilitating meetings, collaboration, and socialisation for the construction of new concepts. It is also important for the study to establish the validity of the theories proposed and the implementation of the technologies applied in terms of the teaching/learning experience.

Through the use of virtual reality learners can thus become more visually aware of their companions, through their avatars, stimulating the 'perception' of the mind that they are no longer *isolated* in their online learning sphere. The design of E-SPACES proposes that avatar presence is persistent within the VWs. In Camilleri & Montebello (2008) the concept of persistence and scope is emphasised in that VWs without a collective scope or interest remain void and fulfil nothing more than a static representation of content transmission. The pedagogical approaches proposed use concepts of 'active learning' involving learners in their own knowledge building, using the constructivist and collaborative models as well as process-oriented models (Bonanno, 2008). Categories of interactions, influenced by a number of parameters and interpersonal factors in the connected VW will also be applied within the design. One fundamental approach to the design is the specification of the learning communities of practice in the VW context, and the content which will serve to connect learners.

The complex human behaviour relationships which will be targeted through E-SPACES will measure the:

- attitude of users towards 3D VWs;
- Perceived behavioural control of users in relation to the VWs;
- Perceived usefulness of VWs for learning;
- connecting relationships established through learning communities.

The setting will be piloted within a specific case scenario, in the higher education context. In this scenario, students following the teacher training course (B.Ed (Hons.) will experience collaborative learning practices, through a hands-on pilot study held inside an immersive environment, such as Second Life (SL,2010) or Olive (Fortrerra Systems, 2010). This pilot study will embark on offering individual learning 'objects' within the virtual world following the 'FREEDOM' model outlined in section 4 of this document.

## 6. The E-SPACES Framework

Based upon the perspectives of e-learning design, the E-SPACEs framework will take into account all the interaction processes for connectivity and build a virtual space using the 'active' model.

The E-SPACES framework will be designed around a simulating environment exploiting the VW through collaborative, constructivist and experiential activities. Content presented for the pilot study will focus around specific tasks and activities which future teachers can design and create for their students. This means that the through the virtual world, these future teachers, will partake into their own active learning processes to design different activities for school children at different levels. Collaboration will take place within this virtual meeting place which also offers sandboxes, to be able to experience in practice, their peers' task designs. The scope of the framework is to clearly define the measurement parameters and establish whether VWs increase the effectiveness of the teaching/learning process comparing the results to a real world control group participating in the same exercise on a face-to-face classroom setting.

E-SPACES proposes that the content bridges the gap between the pedagogic approaches and the interactions between the actors involved. In the VW, E-SPACES proposes three distinct actors all having a number of interactions; the educator as the instructional designer, the virtual agent as an intelligent assistant facilitating the virtual experience, and the learners actively involved in their own learning process. The interactions proposed involve the three actors, interrelating with the content presented within the socio-collaborative environment. The approaches connected with the actors' interactions will build this virtual ecosystem which will be the niche of the learning experience in the social space constructed.

The research questions will be shared amongst the actors in this framework. The methodology proposes both qualitative and quantitative data collection, taking views from educators and learners, and also measuring students' attainment targets at the end of a pilot course in the E-SPACE framework.

The questions proposed in this study are designed for the measurement of effectiveness within this learning framework and will facilitate a clearer understanding of the findings.

Question #1: How does learning occur in the VW?

Question #2: What are the students' perceptions of learning in the online context?

Question #3: What are the students' perceptions of learning in the VW context?

Question #4: Does learning transfer from the VW to real life?

Question #5: What is the perceived usefulness of the VW context for the learners?

Question #6: How are the interactions in the VW established?

Question #7: How useful for their learning do learners find the interactions within the space?

## 7. Future Developments

The E-SPACES framework is interdependent on a number of parameters including the VW platform chosen, the target sector of learners involved in the pilot study, and the content which is chosen to bridge the gap between the pedagogic approaches and the interactions proposed. It is being proposed that the current study undergoes specific analysis to gather data for this framework. It is then proposed that data and content are integrated within the framework and implemented during a short pilot course. The limitations and challenges of this study, will surface if a limited number of students are chosen for this study. This might occur depending on the content chosen and the participants available for the duration of the course. The quantitative measure of the effectiveness of the social spaces, will need to be performed against a control. Such control might be difficult to establish in the context of the learning environment. It is expected that the future development of E-SPACES is to identify limitations and challenges, for the design of the study measuring the effectiveness of social niches established in the context of VWs.

## 8. Conclusion

Whilst the use of 3D-VWs seem to point towards their increased use for the learning contexts of the future, there is limited research validating their effectiveness based upon pedagogic approaches taking into consideration collaborative learning in the socio-constructivist perspective. Virtual spaces have a number of characteristics which can be found commonly throughout all platforms including the presence of avatars, an immersive experience and a series of interactions between player characters, non player characters and other world components. VWs are a combination allowing for simulation and the "real" virtuality. Can what happens in a real classroom, including all the interactions and exchanges, indeed be transferred to the virtual world? How can this challenge be identified and overcome? Can technology be used to increase the effectiveness of this learning medium? This research is needed to understand how experiential collaborative activities may apply to a number of instructional contexts within the VWs.

## 9. References

Ajjan, H., & Hartshorne, R. (2008). Investigating Faculty descisions to adopt Web2.0 technologies: Theory and empirical tests. *Internet and Higher Education* , 71-80.

Alier, M. (2006). A Social Constructionist Approach to Learning Communities: Moodle. In M. D. L, & N. (. Ambjorn, *Open Source for Knowledge and Learning Management: Strategies beyond Tools.* Idea Group, INC.

Barbour, M. K., & Reeves, T. (2009). The reality of virtual schools: A review of literature. *Computers & Education* , 402-416.

Bonanno, P. (2008). *Learning through Collaborative Gaming: A Process-oriented Pedagogy.* Finland: Joensuu.

Camilleri, V., & Montebello, M. (2008). SLAVE – Second Life Assistant in a Virtual Learning Environment. *RELIVE08 – Researching Learning in Virtual Environments.* Milton-Keyes: The Open University.

Carey, J. (2007). Expressive Communication and Social Conventions in Virtual Worlds. *The Data Base for Advances in Information Systems* , 81-85.

Casamayor, A., Amandi, A., & Campo, M. (2009). Intelligent assistance for teachers in collaborative learning environments. *Computers & Education* , 1147-1154.

Chou, S.-W., & Min, H.-T. (2009). The impact of media on collaborative learning in virtual settings: The perspective of social construction. *Computers & Education* , 417-431.

Etelapelta, A., & Lahti, J. (2008). The resources and obstacles of creative collaboration in a long-term learning community. *Thinking Skills and Creativity* , 226-240.

Granic, A., Mifsud, C., & Cukusic, M. (2009). Design, implementation and validation of a Europe-wide pedagogical framework for e-Learning. *Computers & Education* , 1052-1081.

Jarmon, L., Traphagan, T., M, M., & Trivedi, A. (2009). Virtual world teaching, experiential learning, and assessment: An interdisciplinary communication course in Second Life. *Computers & Education* , 169-182.

Kolb, D. A. (1984). *Experiential learning: Experience as a source of learning and development.* Englewood Cliffs, NJ: Prentice-Hall.

Kumar, S., Chhugani, J., Kim, C., Kim, D., Nguyen, A., Dubey, P., et al. (2008). Second Life and the New Generation of Virtual Worlds. *Computer* , 46-53.

Miller, M., & Brunner, C. C. (2008). Social impact in technologically-mediated communication: An examination of online influence. *Computers in Human Behavior* , 2972-2991.

Monahan, T., G, M., & Bertolotto, M. (2008). Virtual Reality for Collaborative e-learning . *Computers & Education* , 1339-1353.

NMC. (2010). *What is Happening in Virtual Worlds?* US: NMC.

Paechter, M., Maier, B., & Macher, D. (2010). Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction. *Computers & Education* , 222-229.

Petrakou, A. (2009). Interacting through avatars: Virtual worlds as a context for online education. *Computers &*

*Education* .

Solimeno, A., M.E., M., Tomai, M., & Francescato, D. (2008). The influence of students and teachers characteristics on the efficacy of face-to-face and computer supported collaborative learning. *Computers & Education* , 109-128.

Spalter, A., & Simpson, R. (2000). Integrating interactive computer-based learning experiences into established curricula: a case study. *Proceedings of the 5th annual SIGCSE/SIGCUE ITiCSEconference on Innovation and technology in computer science education* (pp. 116 - 119 ). Helsinki, Finland: ACM, New York.

Tiropanis, T., Davis, H., Millard, D., Weal, M., White, S., & Wills, G. (2009). *JISC - SemTech Project Report.* Southampton, UK: JISC CETIS.

Tomai, M., Rosa, V., Mebane, M. E., A, D., Benedetti, M., & Francescato, D. (2010). Virtual communities in schools as tools to promote social capital with high school students. *Computers & Education* , 265-274.

Vygotsky, L. (1978). *Mind and society: The development of higher mental processes.* Cambridge, MA: Harvard University Press.

Wrzesien, M., & Alcaniz Raya, M. (2010). Learning in serious virtual worlds: Evaluation of learning effectiveness and appeal to students in the E-Junior project. *Computers & Education* .

# A Semantic Knowledge Base for Personal Learning and Cloud Learning Environments

**Alexander Mikroyannidis, Paul Lefrere, Peter Scott**
Knowledge Media Institute, The Open University
Milton Keynes MK7 6AA, United Kingdom
E-mail: {A.Mikroyannidis, P.Lefrere, Peter.Scott}@open.ac.uk

**Abstract**

Personal Learning Environments (PLEs) and Cloud Learning Environments (CLEs) have recently encountered a rapid growth, as a response to the rising demand of learners for multi-sourced content and environments targeting their needs and preferences. This paper introduces a semantic knowledge base that utilises a multi-layered architecture consisting of learning ontologies customized for certain aspects of PLEs and CLEs. A number of stakeholder clusters, including learners, educators, and domain experts, are identified and are assigned distinct roles for the collaborative management of this knowledge base.

## 1. Introduction

Personal Learning Environments (PLEs) and Cloud Learning Environments (CLEs) are gradually gaining ground over traditional Learning Management Systems (LMS) by facilitating the lone or collaborative study of user-chosen blends of content and courses from heterogeneous sources, including Open Educational Resources (OER).

PLEs follow a learner-centric approach, allowing the use of lightweight services and tools that belong to and are controlled by individual learners. Rather than integrating different services into a centralised system, PLEs provide the learner with a variety of services and hands over control to her to select and use these services the way she deems fit (Chatti *et al.*, 2007).

CLEs extend PLEs by considering the cloud as a large autonomous system not owned by any educational organisation. In this system, the users of cloud-based services are academics or learners, who share the same privileges, including control, choice, and sharing of content on these services. This approach has the potential to enable and facilitate both formal and informal learning for the learner. It also promotes the openness, sharing and reusability of OER on the web (Malik, 2009).

In the context of the European project ROLE (Responsive Open Learning Environments - www.role-project.eu) we are targeting the adaptivity and personalization of learning environments, in terms of content and navigation, as well as the entire learning environment and its functionalities. We propose the use of ontologies to model various aspects of the learning process within such an environment. In particular, we consider a semantic knowledge base as the core of the learning environment, enabling the collaboration between diverse stakeholder clusters.

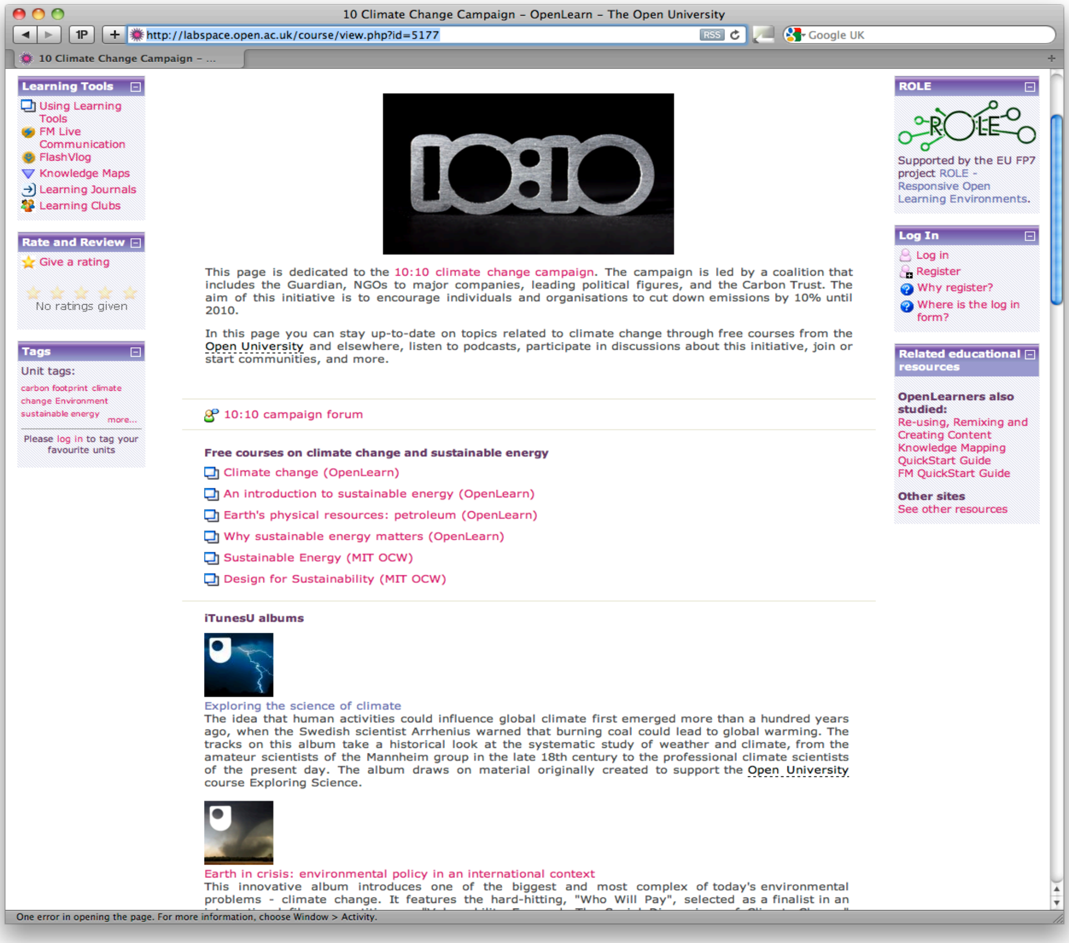
The remainder of this paper is organised as follows. Section 2 describes the OpenLearn case study, consisting of a traditional LMS into transition towards the PLE and CLE paradigms. Section 3 introduces the architecture of the proposed semantic knowledge base and discusses the various learning ontologies that formulate it. Section 4 presents integration mechanisms for the different layers of the knowledge base. Section 5 describes the involved stakeholder clusters and their roles within the management of the knowledge base. Section 6 discusses certain challenges arising from the collaborative nature of the management of the knowledge base. Finally, the paper is concluded and the next steps for progressing this work are provided.

## 2. The OpenLearn case study

The Open University (www.open.ac.uk) provides a wide range of OER through the OpenLearn educational environment (http://openlearn.open.ac.uk). OER can be described as “teaching, learning and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or repurposing by others depends on which Creative Commons license is used” (Atkins *et al.*, 2007). OER are freely available on the web and can be accessed through common web sites or Virtual Learning Environments (VLEs), and more recently through PLEs and CLEs. They can be used, edited and shared by any interested party, such as learners, teachers, institutions, and learning communities.

OpenLearn users have the ability to learn at their own pace, keep a learning journal in order to monitor their progress, complete self assessment exercises, and discuss with other learners in forums. OpenLearn has gathered the interest of a wide audience ranging from governmental and non-governmental entities interested in promoting continuing professional development, public and private higher education institutes, academic teachers, training course designers, graduate and postgraduate students, educational researchers, and generally anyone interested in informal learning (Okada, 2007).

OpenLearn is essentially a traditional LMS, based on the Moodle platform (http://moodle.org), following a course-based paradigm, rather than a learner-based one. It has been built around units of study and not the personal profiles of learners. Currently, OU students are missing a place where they can aggregate the content offered by different OU services, such as OpenLearn and iTunesU, and mix it together with other educational


10 Climate Change Campaign – OpenLearn – The Open University
http://labspace.open.ac.uk/course/view.php?id=5177
RSS
Google UK
10 Climate Change Campaign – ...

**Learning Tools**
- Using Learning Tools
- FM Live Communication
- FlashVlog
- Knowledge Maps
- Learning Journals
- Learning Clubs

**Rate and Review**

Give a rating

☆ ☆ ☆ ☆ ☆
No ratings given

**Tags**

Unit tags:
carbon footprint climate change Environment sustainable energy more...

Please log in to tag your favourite units

This page is dedicated to the 10:10 climate change campaign. The campaign is led by a coalition that includes the Guardian, NGOs to major companies, leading political figures, and the Carbon Trust. The aim of this initiative is to encourage individuals and organisations to cut down emissions by 10% until 2010.

In this page you can stay up-to-date on topics related to climate change through free courses from the Open University and elsewhere, listen to podcasts, participate in discussions about this initiative, join or start communities, and more.

10:10 campaign forum

**Free courses on climate change and sustainable energy**
- Climate change (OpenLearn)
- An introduction to sustainable energy (OpenLearn)
- Earth's physical resources: petroleum (OpenLearn)
- Why sustainable energy matters (OpenLearn)
- Sustainable Energy (MIT OCW)
- Design for Sustainability (MIT OCW)

**iTunesU albums**

Exploring the science of climate
The idea that human activities could influence global climate first emerged more than a hundred years ago, when the Swedish scientist Arrhenius warned that burning coal could lead to global warming. The tracks on this album take a historical look at the systematic study of weather and climate, from the amateur scientists of the Mannheim group in the late 18th century to the professional climate scientists of the present day. The album draws on material originally created to support the Open University course Exploring Science.

Earth in crisis: environmental policy in an international context
This innovative album introduces one of the biggest and most complex of today's environmental problems - climate change. It features the hard-hitting, "Who Will Pay", selected as a finalist in an

**ROLE**

Supported by the EU FP7 project ROLE - Responsive Open Learning Environments.

**Log In**
- Log in
- Register
- Why register?
- Where is the log in form?

**Related educational resources**

**OpenLearners also studied:**
Re-using, Remixing and Creating Content
Knowledge Mapping QuickStart Guide
FM QuickStart Guide

**Other sites**
See other resources

One error in opening the page. For more information, choose Window > Activity.

Figure 1. Climate change OER in OpenLearn (http://tinyurl.com/yene49o)

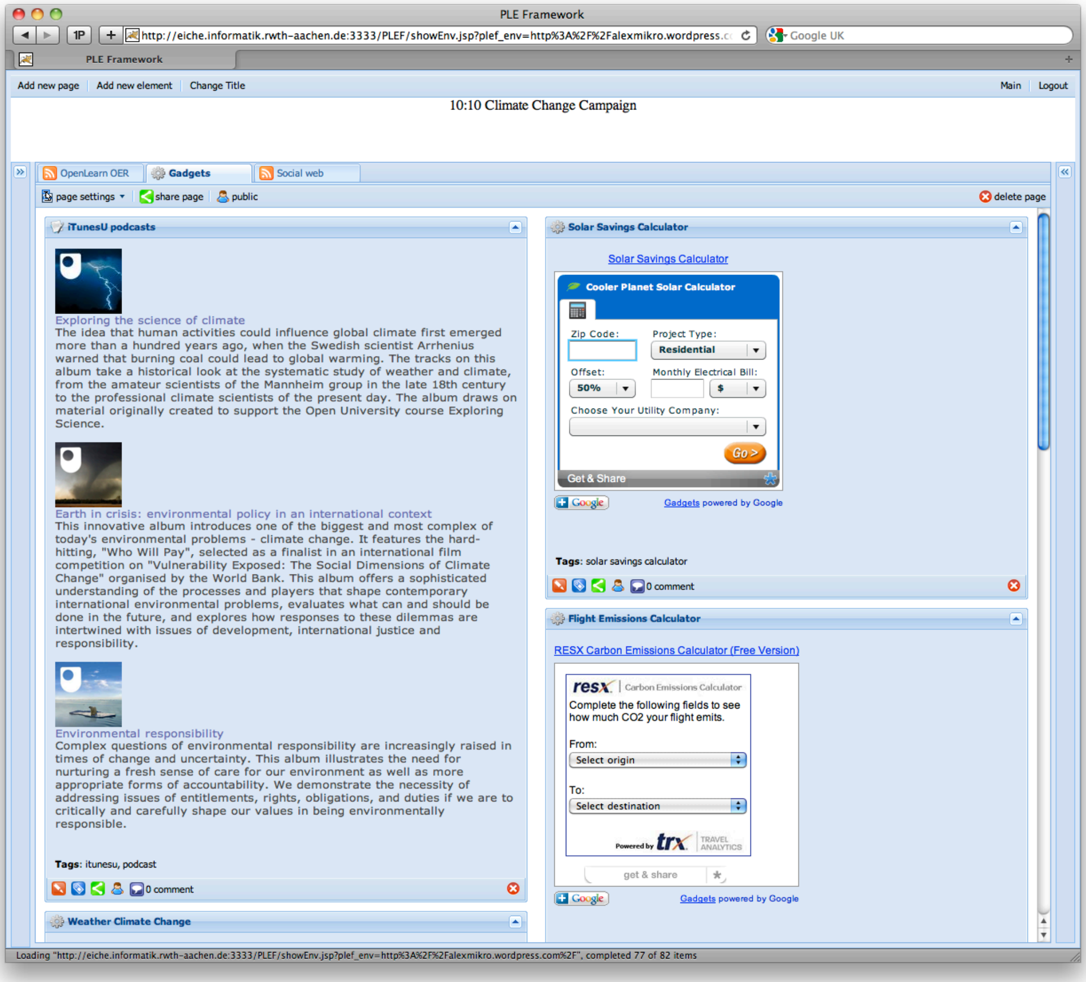
content. Therefore, what we aim to offer OU students in the context of ROLE, is a combined aggregator and e-portfolio, where they can set their learning goals, gather and organise various learning resources, monitor their progress, get recommendations from the system and their peers, and connect with other learners.

In order to explore the present limitations of OpenLearn, we have been comparing its capabilities with those of a PLE, by delivering the same learning resources with both approaches. For this purpose, we have created a collection of OER related to the UK 10:10 climate change campaign (http://www.1010uk.org/). Figure 1 shows this collection delivered by the existing OpenLearn environment, featuring OpenLearn courses and OU albums from iTunesU. In addition, content from external sources, such as YouTube and SlideShare, is included. However, syndication from dynamic Web 2.0 sources, such as the blogosphere, Twitter, and FriendFeed, is not supported.

On the other hand, the PLE of Figure 2 is a showcase of a widget-based environment hosting the same climate change resources as in OpenLearn, in addition to dynamic Web 2.0 sources. Compared to OpenLearn, this approach offers more flexibility in terms of creating new widgets, configuring them, tagging them, and organising them into thematic categories in different tabs.

In the context of the ROLE project, we are working on the transition from the LMS-based approach of OpenLearn towards the PLE and CLE paradigms, by putting emphasis to the needs and preferences of learners. In particular, we aim at providing them with a wider range of OER to choose from, both from OpenLearn as well as from external Web 2.0 sources. However, discovering OER from such a wide range is not an easy task; therefore providing the learners with OER recommendations based on information from their profiles and portfolios is very important.

We propose the use of ontologies to model various aspects of the learning process within the transformed OpenLearn environment. In particular, we consider a semantic knowledge base as the core of this learning environment, enabling the use of metadata and ontologies to annotate learning resources, and model various aspects of the learning process, such as learner profiles. The curation of the proposed semantic knowledge base is supported by the active involvement and collaboration between different stakeholder clusters.


PLE Framework
http://eiche.informatik.rwth-aachen.de:3333/PLEF/showEnv.jsp?plef_env=http%3A%2F%2Falexmikro.wordpress.c
Google UK
PLE Framework
Add new page
Add new element
Change Title
Main
Logout
OpenLearn OER
Gadgets
Social web
page settings
share page
public
delete page
iTunesU podcasts
Solar Savings Calculator
Cooler Planet Solar Calculator
Go >
Get & Share
Gadgets powered by Google
0 comment
Flight Emissions Calculator
RESX Carbon Emissions Calculator (Free Version)
resx Carbon Emissions Calculator
Powered by trx TRAVEL ANALYTICS
get & share
Gadgets powered by Google
Weather Climate Change

10:10 Climate Change Campaign

Exploring the science of climate

The idea that human activities could influence global climate first emerged more than a hundred years ago, when the Swedish scientist Arrhenius warned that burning coal could lead to global warming. The tracks on this album take a historical look at the systematic study of weather and climate, from the amateur scientists of the Mannheim group in the late 18th century to the professional climate scientists of the present day. The album draws on material originally created to support the Open University course Exploring Science.

Earth in crisis: environmental policy in an international context

This innovative album introduces one of the biggest and most complex of today's environmental problems - climate change. It features the hard-hitting, "Who Will Pay", selected as a finalist in an international film competition on "Vulnerability Exposed: The Social Dimensions of Climate Change" organised by the World Bank. This album offers a sophisticated understanding of the processes and players that shape contemporary international environmental problems, evaluates what can and should be done in the future, and explores how responses to these dilemmas are intertwined with issues of development, international justice and responsibility.

Environmental responsibility

Complex questions of environmental responsibility are increasingly raised in times of change and uncertainty. This album illustrates the need for nurturing a fresh sense of care for our environment as well as more appropriate forms of accountability. We demonstrate the necessity of addressing issues of entitlements, rights, obligations, and duties if we are to critically and carefully shape our values in being environmentally responsible.

**Tags**: itunesu, podcast

Zip Code:

Project Type: Residential

Offset: 50%

Monthly Electrical Bill: $

Choose Your Utility Company:

**Tags**: solar savings calculator

Complete the following fields to see how much CO2 your flight emits.

From: Select origin

To: Select destination

Loading "http://eiche.informatik.rwth-aachen.de:3333/PLEF/showEnv.jsp?plef_env=http%3A%2F%2Falexmikro.wordpress.com%2F", completed 77 of 82 items

Figure 2. A widget-based PLE for climate change OER (http://tinyurl.com/m6zrhl)

## 3. Semantic knowledge base architecture

In order to efficiently manage the metadata associated with different aspects of the learning process, we propose their organisation into a number of ontology layers. Figure 3 shows the multi-layered semantic knowledge base adapted from the Heraclitus II framework (Mikroyannidis and Theodoulidis, 2006, Mikroyannidis, 2007, Mikroyannidis and Theodoulidis, 2010).

In this pyramid, the lower layers represent more generic and all-purpose ontologies, while the ontologies of the upper layers are customized for certain uses within a PLE or CLE. When traversing the pyramid from bottom to top, each layer reuses and extends the previous ones. In addition, whenever a layer extends the ones below it (e.g. with the insertion of new concepts), these extensions are propagated to the lower layers. Different stakeholder clusters curate each layer, depending on the expertise that each layer requires. The integration of the ontology pyramid layers is achieved with the use of ontology mappings between ontologies belonging to the same or different layers.

Starting from the top of the pyramid, the *Learner layer* contains ontologies that model the profiles of the learners involved in the learning process. In particular, the ontologies of this layer model the learners' profiles according to their interests, goals, preferences, and skills. Some ontology standards corresponding to this layer are the IEEE Learning Objects Metadata Standard (LOM) (http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf), the IEEE Personal and Private Information for Learner (IEEE PAPI) both developed by the IEEE Learning Technology Standards Committee (LTSC), the IMS Learner Information Package (LIP) (http://www.imsglobal.org/profiles), and the IMS Reusable Definition of Competency and Educational Objective (RDCEO) (http://www.imsglobal.org/competencies).

The *Learning Resource layer* models the learning resources that are employed within a PLE or CLE by learners. These resources are mainly widgets of educational tools and content. For example, the climate change PLE of Figure 2 includes widgets of:

- OpenLearn OER
- iTunesU albums
- External resources, e.g. blog feeds, YouTube videos,

SlideShare presentations, Google gadgets, etc.
- Knowledge maps

The ontologies of the Learning Resource layer are constructed out of annotations of these widgets. These annotations can be user-generated tags, or automatically generated semantic annotations, e.g. with the use of IE (Information Extraction) and NLP (Natural Language Processing) techniques. Apart from the Learner layer, the IEEE Learning Objects Metadata Standard (LOM) also corresponds to this layer, as it defines models for learning objects, including multimedia content, instructional content, as well as instructional software and software tools.

The *Learning Domain layer* models the learning domain of interest. These are more generic ontologies describing a certain domain of interest to the learner, e.g. bioinformatics. The ontologies of the Gene Ontology (GO) project (The Gene Ontology Consortium, 2000) and the Foundational Model of Anatomy (FMA) (Cornelius Rosse, 2003) are some widely used domain ontologies in bioinformatics.

Finally, the *Lexical layer* contains domain-independent ontologies of a purely lexicographical nature. An example of such an ontology is the widely adopted WordNet (Fellbaum, 1998). A lexical ontology is the most generic form of ontology that can be constructed. The ontologies of this layer can be used to model practically any domain. The ontologies of all the other layers are independent of the language used, or other linguistic issues, which concern only this layer.

Although lexical ontologies constitute a strong basis for the construction of any domain-specific ontology, their relations tend quite often to be imprecise and thus not suitable for logical reasoning. This can be addressed with the use of more strictly constructed, general purpose ontologies, such as SUMO (Sevcenko, 2003). Such models can act as structuring mechanisms for lexical ontologies or intermediates between lexical and domain ontologies.

## 4. Knowledge base integration

The integration of the ontology pyramid layers into a single manageable scheme is achieved with the use of ontology mappings. In terms the layers of the ontology pyramid being mapped, ontology mappings are either *intra-layer*, mapping ontologies of the same ontology layer, or *inter-layer*, mapping ontologies belonging to different layers.

From an architectural point of view, ontology mappings can be either *structural*, namely referring to the structure of the mapped ontologies, e.g. via is-a relations, or *semantic* when mapping two ontology objects via a semantic relation, such as an employer-employee relation. OWL Full (Bechhofer *et al.*, 2004) offers a variety of constructs for representing structural ontology mappings, including `owl:subclassOf`, `owl:sameAs`, `owl:inverseOf`, `owl:equivalentClass`, and `owl:equivalentProperty`.

Ontology mappings are particularly useful for the extraction of recommendations to the learner, as they link her profile to learning resources, as well as to profiles of other learners. They can therefore be used to recommend learning resources of potential interest to the learner. They can also be used to recommend a 'study-buddy', with whom the learner shares common abilities and interests.

## 5. Stakeholder clusters

Since each ontology layer represents a different degree of specialization, different stakeholder clusters are required to contribute to the curation of each layer. Starting from the bottom of the pyramid, lexicographers have the knowledge on language structures that is required in this level. Domain experts need to be employed for the next layer. These are professionals on a certain domain, e.g. biologists are responsible for a biology-related ontology.

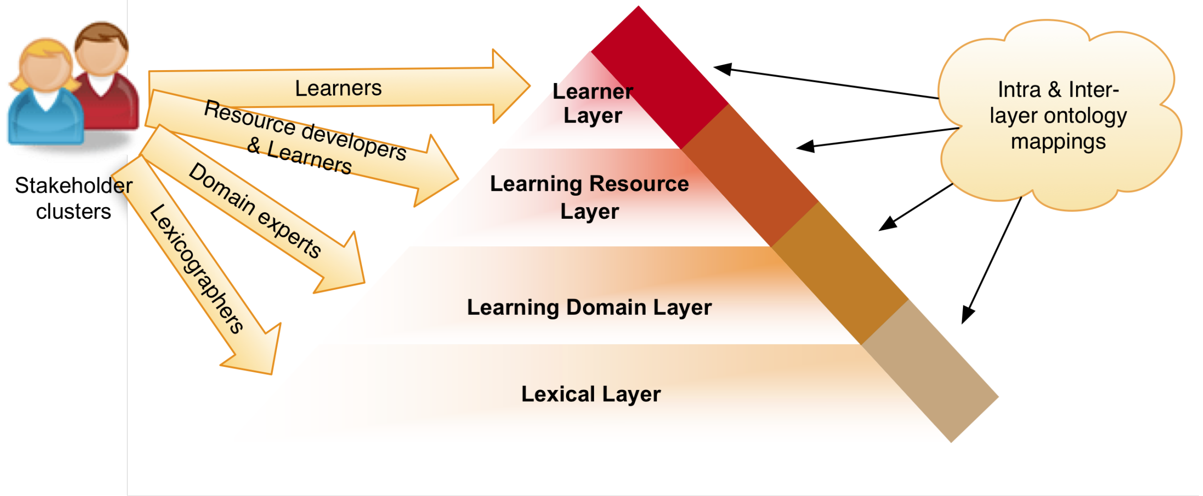


Figure 3. Multi-layered semantic knowledge base

For the Learning Resource layer, a more diverse group is suitable: producers and consumers of learning resources. The producers are those that develop learning resources, either content or tools. They can be lecturers, learning designers, or team leaders who develop new courses, workshops or training sessions and author new learning material. The consumers are learners who use and annotate the offered learning resources.

Finally, the Learner layer is curated by learners, who provide information about themselves in order to receive recommendations about learning resources and create personal networks with users from different learning environments, with whom they may share common learning interests.

Depending on the scope of intra and inter-layer ontology mappings, these are performed by one or more stakeholder clusters. For example, an inter-layer ontology mapping between the lexical and the domain layer will be created jointly by the stakeholder clusters of these two layers, namely lexicographers and domain experts. Intra-layer ontology mappings are performed by the stakeholder cluster of the corresponding layer. The assignment of stakeholder clusters as curators of the ontology pyramid layers is summarized in Table 1.

| **Ontology layer** | **Stakeholder cluster** |
|---|---|
| Lexical layer | Lexicographers |
| Learning domain layer | Domain experts |
| Learning resource layer | Learning resource developers / Learners |
| Learner layer | Learners |
| Inter-layer ontology mappings | Stakeholder clusters of corresponding layers |
| Intra-layer ontology mappings | Stakeholder cluster of corresponding layer |

Table 1. Assignment of stakeholder clusters as curators of the semantic knowledge base

## 6. Challenges in collaborative ontology management

Collaboration between stakeholder clusters in curating the semantic knowledge base is essential; however, it involves several challenges, including concurrency, consistency, and scalability issues. We will be targeting the following set of parameters for collaborative ontology management, as outlined in (Bao et al., 2006):

- **Knowledge integration**: A fundamental task in a collaborative environment is the integration of contributions from multiple participants. The proposed semantic knowledge base consists of a multi-layer architecture that is curated by diverse clusters of stakeholders. Reusability and integration is supported through ontology mappings.

- **Concurrency management**: Different ontology authors need to be able to work on different parts of the knowledge base simultaneously. In case the same part of the knowledge base is concurrently edited by more than one author, this can cause conflicts. Various technologies can be used to address this issue, such as CVS (The Gene Ontology Consortium, 2000), Wiki (Auer et al., 2006, Schaffert, 2006), or peer-to-peer based solutions (Becker et al., 2005, Xexeo et al., 2004).

- **Consistency maintenance**: Parts of the knowledge base curated by different authors may be inconsistent with each other, since an ontology usually reflects the point of view of each author. Mechanisms for structural and semantic consistency preservation as well as change propagation need to be provided to ensure that the knowledge base is free of inconsistencies at all times.

- **Privilege management**: In order to ensure the accuracy of the knowledge base, a collaborative environment needs to assign different levels of privileges to its users, based on their expertise, authority, and responsibility. Our architecture is based on a flat scheme regarding privilege management, by giving each stakeholder cluster equal privileges in their layer of responsibility.

- **History maintenance**: Collaborative environments should provide the means to recover from wrong or unintended changes to the knowledge base. All changes to the knowledge base should be thus recorded in order to be able to track the authorship of a change and to prevent loss of important information. The bitemporal ontology model of Heraclitus II (Mikroyannidis, 2007) retains the necessary information to achieve this goal.

- **Scalability**: Long-term collaboration of diverse parties usually increases the size of knowledge bases; therefore, a collaborative environment has to be scalable to large ontologies. This is particularly important in the abundant environment of CLEs, where a wide variety of cloud-based services is employed.

## 7. Conclusion and next steps

PLEs and CLEs address the crucial demands of today's learner for a personalized and adaptive learning environment. In order to achieve these goals, we propose the use of ontologies for modeling the learning process and assigning distinct curator roles to the involved stakeholder clusters. We perceive a semantically enhanced PLE or CLE as the evolution of the present OpenLearn environment, as well as the evolution of LMS-based approaches in general.

We are currently in the process of refining the specifications of the proposed semantic knowledge base for addressing particular requirements of the OpenLearn case study. This refinement includes reviewing existing ontology standards in terms of their suitability to be reused, repurposed and adapted within an OpenLearn-specific ontology pyramid.

## 8. Acknowledgements

The research work described in this paper is partially funded through the ROLE Integrated Project, part of the Seventh Framework Programme for Research and Technological Development (FP7) of the European Union in Information and Communication Technologies.

## 9. References

Atkins, D. E., Brown, J. S. & Hammond, A. L. (2007) A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities. The William and Flora Hewlett Foundation, http://www.oerderves.org/wp-content/uploads/2007/03/a-review-of-the-open-educational-resources-oer-movement_final.pdf.

Auer, S., Dietzold, S. & Riechert, T. (2006) OntoWiki - A Tool for Social, Semantic Collaboration. *5th International Semantic Web Conference (ISWC 2006).* Athens, GA, USA, Springer LNCS, 736-749.

Bao, J., Hu, Z., Caragea, D., Reecy, J. & Honavar, V. G. (2006) A Tool for Collaborative Construction of Large Biological Ontologies. *17th International Conference on Database and Expert Systems Applications (DEXA'06).* Krakow, Poland, 191-195.

Bechhofer, S., Harmelen, F. V., Hendler, J., Horrocks, I., Mcguinness, D. L., Patel-Schneider, P. F. & Stein, L. A. (2004) OWL Web Ontology Language Reference. IN DEAN, M. & SCHREIBER, G. (Eds.) *W3C Recommendation.*World Wide Web Consortium, http://www.w3.org/TR/owl-ref/.

Becker, P., Eklund, P. & Roberts, N. (2005) Peer-to-peer based ontology editing. *International Conference on Next Generation Web Services Practices (NWeSP 2005).* Seoul, Korea, 259-264.

Chatti, M. A., Jarke, M. & Frosch-Wilke, D. (2007) The future of e-learning: a shift to knowledge networking and social software. *International Journal of Knowledge and Learning,* 3(4/5), 404-420.

Cornelius Rosse, J. L. V. M. J. (2003) A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *Biomedical Informatics 36 (2003)*, 478 - 500.

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*: The MIT Press.

Malik, M. (2009) Cloud Learning Environment - What it is? *EduBlend.* http://edublend.blogspot.com/2009/12/cloud-learning-environment-what-it-is.html.

Mikroyannidis, A. (2007) Heraclitus II: A Framework for Ontology Management and Evolution. *PhD Thesis, Manchester Business School, University of Manchester,* Manchester

Mikroyannidis, A. & Theodoulidis, B. (2006) Heraclitus II: A Framework for Ontology Management and Evolution. *2006 IEEE /WIC/ACM International Conference on Web Intelligence (WI 2006).* Hong Kong, China, IEEE Computer Society, 514-521.

Mikroyannidis, A. & Theodoulidis, B. (2010) Ontology Management and Evolution for Business Intelligence. *International Journal of Information Management,* (forthcoming).

Okada, A. (2007) Knowledge Media Technologies for Open Learning in Online Communities. *International Journal of Technology, Knowledge and Society,* 3(5), 61-74.

Schaffert, S. (2006) IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. *15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE'06).* Manchester, UK, 388-396.

Sevcenko, M. (2003) Online Presentation of an Upper Ontology. *Znalosti 2003.* Ostrava, Czech Republic.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat Genetics,* 25, 25-29.

Xexeo, G., De Souza, J. M., Vivacqua, A., Miranda, B., Braga, B., Almentero, B. K., D' Almeida, J. N., Jr. & Castilho, R. (2004) Peer-to-peer collaborative editing of ontologies. *8th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2004).* Xiamen, China, 186-190

# Semantic Annotation for Semi-Automatic Positioning of the Learner

**Petya Osenova, Kiril Simov**
Linguistic Modelling Laboratory, IPP-BAS
Acad. G.Bonchev 25A, 1113 Sofia, Bulgaria
petya@bultreebank.org, kivs@bultreebank.org

### Abstract

The learner's positioning with respect to a curriculum is of a great importance to the life-long learning (an informal learner needs to achieve a certain level of competency) as well as to the mobility learning (a student spending a semester in another university). In both cases it is necessary to determine learner's prior knowledge. Thus, he might profit in an optimal way from the consequent learning process. The learner's positioning requires grading of pre-course questionnaires by a tutor. This grading is tedious and time-consuming work. In this paper we present the first implementation of a knowledge-rich method for supporting the tutor in the positioning task. Our method exploits the potential of the semantic annotation with regard to the curriculum and the learner's questionnaire answers. The annotation of the curriculum provides the level of the competence to be covered in the course, while the annotation of the questionnaire answers provides evidence for the learner competence per se. The final judgment is assigned to the tutor. The presented method might be well used also for the learner's self-positioning with slight modifications only.

## 1. Introduction

Learner's positioning, on the one hand, has proved to be a very important step in the learning process, and on the other hand, to be a very difficult task. It has been often considered in the context of the self-positioning (Ross 2006) or the context of various groups of practices (Braun and Schmidt 2008). The central role in the positioning task plays the tutor, since there has not been invented yet a completely automatic way, which to be 100 % successful and reliable.

Hence, our aim is to support the tutor in his judgments when positioning the learners. We assume that the tutor is inspecting a set of learner's answers to a questionnaire. The questionnaire would reflect the required knowledge that has to be covered by the learner. The actual positioning is with respect to a curriculum, which presents the following aspects: the knowledge-oriented requirements for a learner, a set of learning materials to support him during the learning process, and links to people who might help him with the learning topics. Thus, the questionnaire is designed on the base of the curriculum. The positioning in these settings is viewed as a set of recommendations from the tutor to the learner which directs the learner within the curriculum, i.e. which materials to study, which people to contact, etc.

Our method relies on the comparison among the curriculum and the related learner's answers, both semantically annotated. This comparison highlights the learner abilities to express the necessary concepts in the answers of the questions from the questionnaire. The tutor can use the results from the comparison to balance his judgments individually for each learner and collectively, as a group. This method has also the advantage that some conceptual or terminological gaps/inconsistencies might also be discovered in the curriculum itself.

The structure of the paper is as follows: Section 2 concentrates on the various aspects of the knowledge rich method. Section 3 overviews the design of the curriculum and the questionnaire answers as well as their interaction. Section 4 describes the semantic annotation of the curricula and answers. Section 5 outlines some preliminary evaluation of the method. Section 6 presents the further extensions over the semantic annotation. Section 7 concludes the paper.

## 2. The Knowledge-Rich Approach

In our work on the positioning of the learner with the help of the knowledge rich approach[1] we rely on the ideas reported in (Kalz et al. 2007). They discuss the notion of the learning network. According to it, learner's competence can be automatically compared to a set of concept evidences of the target competence. Our goal is to achieve an ontology-based positioning where the learner competence is represented by a learner's competence ontology and curriculum competence ontology. However, reliable competence ontologies are still missing. Thus, in our work we rely on domain ontologies, which are supposed to reflect the knowledge part of the learner's competence. The ontological analyses of the learner's portfolio (mainly tests and CVs) and the textual description of the relevant curriculum might be considered an approximation of the learner's (per se) competence against the curriculum (required) competence. Thus we consider the learning network a set of different resources including tutors, experts, learning materials and learners, whose connections are mediated by ontologies. The positioning of a learner within the learning network is identical to the task of creating a learning path for each learner within the established network.

Our method facilitates the tutor in positioning task by analyzing some of the textual elements of the network. Thus, the knowledge rich methods rely on the analysis of the text by using knowledge sources, external to this text, such as ontologies, lexicons, grammars. These sources are used to achieve a semantically rich text analysis which to explicate the conceptual content of the learner's answers

[1] We call the method knowledge rich because it requires an appropriate ontology to represent the conceptual knowledge to be explicated in the curriculum and learner's answers.

to the questionnaire. The main steps in the text analysis that we envisage as necessary in order to support the problem in reliable way, are: (1) grammar-based semantic annotation with concepts from an ontology, (2) discourse segmentation; (3) lexical chains creation to support the disambiguation of concept annotation from (1) and concept distribution within the text; and (4) sentiment analysis for evaluation of the concept usage in the text. The combination of all these analyses should explicate in the best way the conceptual content of the curriculum and the learner texts, which to be used for the positioning.

Our first implementation of the positioning service realizes completely only point (1) and sketches initially a version of the other processing tasks. Therefore, in the rest of this section we will concentrate on analysis (1), namely grammar-based semantic annotation with concepts from an ontology. The reason is that this analysis has already been completely performed and preliminary evaluated in a designed learning-based setting. The other steps are discussed as further extensions to the task design in Section 6.

As mentioned above, the knowledge rich approaches are usually connected with the availability and the usage of knowledge rich databases, such as ontologies and lexicons. The ontologies reflect the conceptualizations in some domain of interest – for example, DAML ontology library, SWOOGLE, LT4eL ontology, etc. These ontologies have to be connected to an upper ontology in order to cover in a better way the general knowledge – for example, DOLCE, SUMO, SIMPLE Core Ontology. The most famous knowledge rich lexicons are the so-called wordnets (WordNet, EuroWordNet, BalkaNet, SIMPLE). Such resources are exploited for the semantic annotation of documents and/or for semantic retrieval.

Within LT4eL project an ontology-to-text relation was defined, which approaches the interaction among conceptualizations and terms (Simov and Osenova 2007; Simov and Osenova 2008). For clarity, this relation is briefly presented here. We assume that the ontology is the repository of the lexical meaning of the language. Thus, we start with a concept in the ontology and we search for lexical items and non-lexical phrases that convey the content of the concept. There are two possible problems: (1) there is no lexical item for some of the concepts in the ontology, and (2) there are lexical items in the language without a concept representing the meaning of the lexical item in the ontology. The first problem is overcome by allowing in the lexicon also non-lexical phrases to be represented. The second problem is solved by extension of the ontology. The lexicon items are then mapped to grammars. We call them concept annotation grammars. These grammars relate the lexicon to the text. Such a mapping is necessary as much as lexical items and phrases from the lexicons allow for multiple realizations in the text. Thus, they require some additional linguistic knowledge in order to disambiguate between different meanings of some lexical item or phrase.

We have been using the relations between the different elements for the task of ontology-based search. The connection from ontology via lexicon to grammars is also relied on for the semantic (concept) annotation of the text. In this way, we established a connection between the ontology and the texts. The relation between the lexicon and the ontology is used for defining user queries with respect to the appropriate segments within the documents. In a more general setting, these relations can be extended to cover the overall learning network. For example, the annotation of a document with concepts connects it to the ontology, which with the help of the lexicon and the grammar connects it to other documents. Similarly, it is possible to annotate other resources, such as images, web sites, people profiles, etc.

## 3. Design of the curriculum and the related questionnaires

As it was mentioned above, we assume that a curriculum consists of set of topics providing the content of a course or a set of courses. Each topic is then associated with a set of learning materials – lectures, tests, descriptions of expected answers, etc.

The learner needs to acquire at least two kinds of knowledge studying the curriculum: the content knowledge and the skills necessary to apply the content knowledge in a community of practise. Here we focus on the content knowledge.

The questionnaire, on the other hand, consists of questions of various types, which check the learner's status with respect to the curriculum topics. They might reflect more surface as well as more profound aspects of the topics.

However, as a first practical approximation we decided to exploit a set, which more or less amalgamates both perspectives – curriculum plus questionnaire, but at the same time is being used in real job seeking situations. As our design setting we used a sample of 10 topic questions, provided by BitMedia within the LTfLL project. The topics are in the IT area. Each question suggested a more surface background when asking about types of things. It also further asked about functions and properties. Some examples are in order: *Explain the meaning of the concept RAM and describe its properties; Name as many PC ports as you can and give some examples.* These topic questions have been equipped with a set of required example answers. Since the set was provided in German, translation was performed into English and Bulgarian. Thus, only the real answers had to be gathered. This part of the setting is described in Section 5.

On the base of this concrete curriculum, we identified the following types of questions: (1) content questions which require answers; (2) skill questions which highlight the learner's abilities to apply the knowledge in practice; (3) personal questions which demonstrate learner's abilities to communicate within a group, etc. Our primary goal is to cover evaluation of questions of the first kind.

## 4. Semantic annotation

In this section the annotation of the curriculum and questionnaire answers will be presented.

The semantic annotation of a curriculum includes two steps. First, all the learning materials related to the curriculum, are annotated automatically with concepts from the domain ontology. Then, the tutor (or the teaching administrator responsible for the curriculum) creates a set of queries to reflect the content knowledge of the curriculum. Each question is also annotated with appropriate concepts to reflect this content knowledge. A comparison is made whether the coverage of the questions meets the requirements within the curriculum.

It is worth mentioning that other questions concerning the skills of the learner might be additionally provided within the question set, but they are not necessarily annotated with concepts from the ontology, and their answers have to be evaluated in a different way.

During the creation of the questions, the tutor has at his disposal the ontology and the semantic annotation of the learning materials. Then the questions are also automatically annotated and the mappings are again presented to the tutor. To sum up, our approach demonstrates the usage of automatic procedures, which alternate with the tutor's intervention, when required.

In our practical setting, the questions, related to the curriculum, were given in advance. Thus, we only provided the automatic annotation of the questions themselves and the example answers. The question annotation was additionally edited by experts in the area of IT.

The following example demonstrates the questions in BitMedia questionnaire with the list of the assigned concepts. The learner's answers to the questions were annotated with concepts automatically by the semantic annotation module, described above. In our setting this step was performed exactly in this way. Here is one example of a query:

> *Name some of the technical specifications of different kinds of monitors.*

The following is a list of concepts, selected as annotations for this question by an expert:

> *CRT monitor, display, contrast, frame rate, graphical elements, image, LCD monitor, monitor, picture, pixel, ratio, refresh rate, rendering, resolution, screen, size, VGA*

This list of concepts demonstrates that the tutor could include not only concepts that are directly answers to the question, but also related concepts which are necessary in order to ensure that the learner uses the concepts related to the answers within the proper context. The above list also demonstrates the case in which concepts and sub-concepts are also included in the list because they define slightly different contexts of usage.

The next example shows the annotation of the learner's answers. The annotation is done within the text of the answer. Then the concepts from this annotation are compared to the concepts from the question annotation and three lists of concepts are created: (1) list of common concepts – the concepts that demonstrate how well the learner competence matches the required competence; (2) list of missing concepts – these concepts determine what is not covered by the learner competence and they can be used to suggest further learning activities; (3) list of additional concepts – these concepts could demonstrate some wrong understanding of the topic by the learner or gaps in the curriculum (topics or semantic annotation).

In the context of the above example a learner responded with the following answer:

> ***Output device**, **monitor**, **display devices** of a **PC**; there are two **types**: **Monitors** with cathode ray tube (**CRT**) - heavy, need more **power**, occupy more **space**; Flat panel **displays** - light, need less **power**, and occupy less **space**.*

The terms in the text recognized as related to concepts in the ontology are highlighted. The three lists are as follows:

Common concepts:

> *CRT monitor, display, monitor*

Missing concepts:

> *contrast, frame rate, graphical elements, image, LCD monitor, picture, pixel, ratio, refresh rate, rendering, resolution, screen, size, VGA*

Additional concepts:

> *types, devices, Output device, PC, power, space*

The concepts in the first two lists are lexicalized on the base of the lexicon, mapped to the ontology. The concepts in the last list are represented with the terms used by the learner. This helps the learner and the tutor to identify the usage context of these concepts. As it was mentioned above, the usage of additional concepts is not always an evidence of wrong knowledge, but could be a good feedback to both - the learners and the tutors. The expression *output device* in the above example might be considered as a good concept to be included in the annotation of the query.

## 5. Evaluation

Having the semantic annotation of the curriculum and the learner's answers, the service calculates several lists of concepts, as it was reported in the previous section. The real evaluation within the learning network of BitMedia is under implementation. Here we report on a small scale evaluation, run by us in order to have some first evidences for the usefulness of the service and to acquire some ideas about the future development of the service.

The concept evidence of the learner's competence can be automatically compared to a set of concept evidences of the target competence (learning network in the terms of (Kalz et al. 2007)). Those are selected, which are not covered by the current learner's competence. For the comparison of the concept evidences we use the standard vector metrics from Information Retrieval community. The automatic evaluation was constructed as a ration of the list of the common concepts with the list of concepts from the annotation of the query.

In order to do evaluation of the automatic method, the 10 questions were given to Bulgarian students in IT area. We gathered more than 10 answers per topic at average. Then, the same answers were given to two tutors in IT area to grade them. First, we compared the concepts, presented in the answers, to those, required in the descriptions. Then, we compared the automatic grading to the tutors' one. The results are as follows: there is a big mismatch between the descriptions and the answers due to short students' productions or avoidance of certain concepts. On the other hand, tutors' grading was different. It accounted certain aspects (such as detailed description of characteristics of the main concepts to be covered by learner's answers), but not others (such as the presence/absence or distribution of concepts). The last point reflects the fact that in a verbose answer it is relatively easy to overestimate the learner's knowledge – especially under time pressure.

Thus, the preliminary evaluation showed that: pure automatic comparison might underestimate learner's knowledge; pure tutor grading also skips some aspects of learner's knowledge while putting more weight to others. The conclusion is that the best way is for the tutor to have at disposal the intersection list of concepts from curriculum and learners' answers in order to present the final judgement with respect to learners' status and future learning materials. The tutor has the concepts from the curriculum, which were mentioned in the answers as well as the list of ones not mentioned.

However, in the long run we aim at a more profiled concept evidence, which would be possible when the extensions to the semantic annotation are added (see the discussion in the next section). In such a case the learner's competence would be set of concept descriptions extracted from the answer. For the moment we envisage to extend the classification of the concepts from three lists to five. We will divide the set of concepts in the following subsets: (1) known concepts; (2) partially known concepts; (3) unknown concepts; (4) concepts with contradictory usages; and (5) additional concepts. The first subset will contain all the concepts which are evaluated as known in the answer. The second subset will contain concepts that are mentioned in the answer, but there is no enough evidence about the level of knowledge of the learner with respect to them. The third subset will contain concepts that are definitely mentioned as unknown by the learner. In the fourth subset we will include the concepts for which there are positive and negative evidences about the knowledge of the learner. The last set is the same as the described above. It can influence the other groups with its relevance or irrelevance. In addition to the extracted concepts, we will extract links to the occurrences of the concepts in the text.

## 6. Extensions to the semantic annotation

For better semantic annotation and its usage in positioning, we consider also additional context-oriented information: co-referential relation annotation, annotation of general lexica and sentiment analysis of the concept usage in the text.

The relation between concept annotation and co-references has been approached from various perspectives. For example, (Lech and de Smedt 2006) and (Nikolov et. al 2009), among others, exploit the semantic features from ontology in order to improve the co-reference chaining; (Kawazoe et al. 2003) designed a software that helps experts in biomedical domain to create ontologies and annotate texts with co-references. In our task, we exploited these papers (together with the work on anaphora and co-reference annotation in general) in the annotation of the corpus. In our future work, we will apply their approaches for the implementation of a new version of our ontology-to-text relation.

One of the reasons for the underestimation of the learner answers by the automatic method is due to the fact that the concept annotation requires very exact answers which sometimes are not present among learners' answers. The learners use freer style of expressing their knowledge. Thus, they rely on similar concepts to the ones in the curriculum annotation – such as, more general or sibling concepts, etc. In order to handle this problem, we envisage extending the annotation from domain concepts via domain terms to general concepts via general lexica.

As it was mentioned in the goal of classifying concepts used by the learner, we would like to evaluate the level of knowledge of the used concept. To do this, we will exploit a version of the sentiment analysis. In our case, the sentiment analysis determines the attitude of the learner to the concepts explicated within the answers. As a starting point for developing of the sentiment analysis, we consider the work reported within (Moilanen and Pulman 2007) and (Liu 2008). It is often underlined that adding knowledge rich features improves the results in sentiment analysis. For example – (Moilanen and Pulman 2007), (Kennedy and Inkpen 2006), (Kim and Hovy 2006). The input for this module will be the results from the previous modules.

In order to construct a concept evidence of the learner's competence, we first need to extract the concepts which are mentioned within the answers text. Then, on the base of the ontological reasoning, the implied concepts will be added. For example, if the answer's holder says that he/she is used to giving injections[2], this automatically means: on more general level, that he/she can intervene in order to improve the situation, and, on more specific level, that he/she can put liquid under the skin by using a syringe. We also need to know in what context each of the concepts in the text was mentioned by the learner. For example, if the learner stated two opposite fact: it is easy to give an intradermal injection, but it is difficult to give an intramuscular one. From this short context a conclusion can be drawn that the learner is not experienced in giving injections as a whole. Thus, comparing conceptual information and discourse relations about the context, each mentioning of a concept will be evaluated by one of the values: 'well known', 'known', and 'unknown'. We will use the methods developed in the areas of sentiment and opinion analysis. As it was already

[2] The examples in this section are from a preliminary work in the medical domain.

mentioned, a pre-defined requirement list of necessary concepts with definitions will be used in order to estimate the degree of competence, delivered by the learner in the portfolio. There will be three types of evaluation: coverage, degree of detailness and relevance. The coverage will be estimated over the number of the mentioned relevant concepts that match the pre-defined list. The degree of detailness will be evaluated over the depth of the conceptual space. And the relevance will be estimated via the ontological relations from a given concept to the other co-occuring concepts within the discourse segment.

## 7. Conclusions

In this paper we presented a knowledge-rich method for supporting the tutor in his positioning task. We presented a preliminary evaluation setting, which showed: the potential of the domain ontologies in the semantic annotation within the life-long learning context; the role of the tutor in the same context; and the natural ways for further extension of the annotation, which aims at a more precise and wider eliciting of learner's knowledge evidences.

The result of the service will be used further to compare the concept evidence of the learner's competence with the learner network. The comparisons will use a vector representation of concept evidence of the learner's competence and concept evidence of the target competence. The vector for the target competence will be fixed within the learner network. The vector for learner's competence will be created by the assessor on the basis of the above sets of concepts. Our goal is not just to calculate these sets of concepts, but also to use them for giving feedback to the learner and thus achieving better results in the learning activities. This kind of feedback will be even more useful when the approach is used for self-positioning of the learner.

Knowledge rich approach requires some initial effort to prepare the necessary resources in order to achieve the goals of positioning of the learner. In our view (also discussed and shared by other colleagues from the LTfLL project – especially Christoph Mauerhofer from Bitmedia), the effort invested at the beginning will pay off during a long and wide exploitation. This could be true in cases of introducing new products of big software companies, where the company itself has the interest to construct appropriate resources (ontologies, lexicons, curriculum, tests, etc). The advantages of the knowledge rich approach are: the exactness of the evidences of the learner competency, the links to the learning materials and the definition of learning paths. Another advantage of the approach is multilinguality – the curriculum and its annotation could be prepared in one language, but it might be reused with little additional effort in many other languages for the learners who do not know the original language of the curriculum.

## 8. Acknowledgements

The work reported within this paper is supported by the European Project LTfLL (http://www.ltfll-project.org/). We would like to thank the colleagues from the project for the discussions on the topics related to the task of positioning of the learner. Also we would like to thank the two reviewers for their valuable comments.

## 9. References

Braun, Simone, Andreas Schmidt. (2008). People Tagging & Ontology Maturing: Towards Collaborative Competence Management. In: 8th International Conference on the Design of Cooperative Systems (COOP '08), Carry-le-Rouet, France, May 20-23, 2008.

Kalz, Marco; Van Bruggen, Jan; Rusman, Ellen; Giesbers, Bas; Koper, Rob. (2007). Positioning of Learners in Learning Networks with Content, Metadata and Ontologies. In Interactive Learning Environments, Volume 15, Issue 2 August 2007 , pages 191 – 200

Kennedy, Alstair and Inkpen, Diana. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence. vol. 22, 2, pp. 110-125.

Kim, Soo-Min and Hovy, Eduard. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text, Sydney, Australia.

Lech, Till Christopher and Koenraad de Smedt. (2006). Enhancing Semantic Annotation through Coreference Chaining: An Ontology-based Approach. In: Siegfried Handschuh, Thierry Declerck, Marja-Riitta Koivunen (eds.), CEUR Workshop Proceedings, Vol. 185, 2006.

Liu, Bing. (2008). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer.

Moilanen, Karo and Pulman, Stephen. (2007). Sentiment Composition. In Proceedings of Recent Advances in Natural Language Processing (RANLP 2007). September 27-29, Borovets, Bulgaria, pp. 378-382.

Nikolov , Andriy, Victoria Uren, Enrico Motta and Anne de Roeck. (2009). Towards instance coreference resolution in a multi-ontology environment. Presented at: Workshop on matching and meaning, Edinburgh, UK, April 2009.

Simov, Kiril and Petya Osenova. (2007). Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects. In Proc. Of RANLP 2007 workshop: Natural Language Processing and Knowledge Representation for eLearning Environments. Borovets, 26. September 2007.

Simov, Kiril and Petya Osenova. (2008). Language Resources and Tools for Ontology-Based Semantic Annotation. In Proc. at the Workshop 'OntoLex 2008' at LREC 2008.

John A. Ross. (2006). The Reliability, Validity, and Utility of Self-Assessment. In: Practical Assessment, Research and Evaluation. Volume 11 Number 10, November 2006.

# Facilitating cross-language retrieval and machine translation by multilingual domain ontologies

**Petr Knoth***, **Trevor Collins***, **Elsa Sklavounou**†, **Zdenek Zdrahal***

* KMI, The Open University
Milton Keynes, United Kingdom
{p.knoth, t.d.collins, z.zdrahal}@open.ac.uk
† SYSTRAN
Paris, France
sklavounou@systran.fr

**Abstract**

This paper presents a method for facilitating cross-language retrieval and machine translation in domain specific collections. The method is based on a semi-automatic adaption of a multilingual domain ontology and it is particularly suitable for the eLearning domain. The presented approach has been integrated into a real-world system supporting cross-language retrieval and machine translation of large amounts of learning resources in nine European languages. The system was built in the context of a European Commission Supported project Eurogene and it is now being used as a European reference portal for teaching human genetics.

## 1. Introduction

A significant amount of research has been carried out in the NLP and Semantic Web technology fields in the last years. A few activities and projects, such as LT4eL (Lemnitzer et al., 2007) or LTfLL (LTfLL, 2008), have been launched with the objective to integrate these technologies with eLearning systems. One of the vital sub-objectives of these projects is to allow seamless access and retrieval of *multilingual* learning materials. In this paper we report on the activities undertaken in the context of *Eurogene (The First Pan-European Learning Service in the Field of Genetics)* project related to the problem of accessing and sharing multilingual learning resources.

More specifically, the article builds on the idea that eLearning systems should not only allow the cross-language retrieval of learning resources, but should be extended with machine translation capabilities to provide a better user experience. The proposed approach synchronizes the adaption of cross-language retrieval and machine translation in such a way that the performance of both systems improves. Although the presented method has been integrated into an eLearning system in the human genetics field, it is applicable in a broader context.

Many of the important players in the information retrieval field (including Google and Yahoo!) offer cross-language information retrieval (CLIR), some of them also provide machine translation (MT). While the performance of these systems is usually sufficient for general queries, CLIR and MT are often inaccurate for domain-specific queries. Large repositories storing domain specific content, such as PubMed which stores vast amounts of scholarly articles, have successfully adopted large thesauri/ontologies of domain terminology to improve the performance of their retrieval system (Lu et al., 2009). While there are efforts targeting cross-language retrieval in eLearning (Lemnitzer et al., 2007; Eichmann et al., 1998; Lu et al., 2008), the combination of the domain-specific retrieval and machine translation is rarely available.

Because of the low frequency of polysemy in domain specific collections, domain-specific MT systems are capable of achieving high performance. However, one of the main obstacles remain in the acquisition of terminology. At the same time, the domain terminology is usually an essential artefact used for query composition. Our method is motivated by this problem and tries to approach it by using a single terminological access point embodied by the multilingual domain ontology for both CLIR and MT. This allows to combine the strengths of ontology-based retrieval and domain-specific machine translation. In Section 2, approaches to domain CLIR with relation to MT are introduced. The theoretical foundation of the method for facilitating domain CLIR and MT is explained in Section 3. The application of the approach in the Eurogene system is then presented in Section 4 and the performance is discussed in Section 5. Finally, the contribution of the paper for the eLearning domain is summarized in Section 6.

## 2. Approaches to domain CLIR

There are two typical approaches to CLIR:

1. MT approach - The user's query is translated from the source language to the target language and submitted to the search system. This approach can be further divided into two cases:
   (a) MT of the query is performed and the query is submitted in all languages of interest.
   (b) A multilingual ontology is developed and used to map the submitted query to different languages.
2. Statistical approaches - The system is trained on a collection of texts (usually parallel). The user's query is then mapped to a language independent document vector using approaches, such as Latent Semantic Indexing (LSI) (Dumais, 1997).

Approach 1(a) requires the search system to be well-adapted for the translation of the terminology of the target domain. Depending on the MT system in hand, domain adaption is rule or statistically based. Rule-based approaches allow specifying rules expressing that a given term $t_{L_1}$ in language $L_1$ corresponds to term $t_{L2}$ in $L_2$. Statistical approaches to machine translation support automatic learning of such pairs from parallel corpora. Approach 1(b) is motivated by the fact that monolingual domain ontologies can be employed to improve the performance of the retrieval system by query expansion leveraging the ability of ontologies to represent synonyms linked to a concept and the hierarchical structure of concepts. Monolingual ontologies can be extended to multilingual ontologies.

Approach 2 is influenced by the size of the available parallel corpora which is critical for the performance of the retrieval system. The approach is, in general, more suitable for bilingual cross-language retrieval as it is usually difficult to find experts to build a domain-specific training set that would contain parallel texts from each language of interest to a common interlingua.

## 3. Synergy of CLIR and MT

Our method is based on the assumption that when we start to build a domain-specific system for sharing language resources, the amount of parallel corpora available is often limited. Our methodology uses a multilingual domain ontology as we argue that ontologies are well-suited for domain CLIR and can also be used for the adaption of the machine translation system. We presume an IR system and a MT system to be available. More specifically, our approach requires a hybrid MT system combining rule-based and statistical-based MT.

The method consists of two phases, which will be discussed in this section in detail: the *initialization phase* and the *bootstrapping phase*. The initialization phase takes as the input a collection of domain texts or an existing monolingual domain ontology and produces as an output a lightweight multilingual ontology of the target domain. While this step is performed just once, the bootstrapping phase is repeated as many times as necessary. The bootstrapping phase takes as the input the multilingual ontology produced in the initialization phase and adapts the MT system by extracting domain specific translation rules from the ontology. As the amount of learning resources stored in the system systematically grows, a statistical module of the MT system can be applied at any time to extract bilingual pairs of domain terms from the available collection of learning resources. These pairs are then used to semi-automatically enrich the multilingual ontology, thus improve the performance of the CLIR and later also the MT system.

The **initialization phase** can be further divided into:

1. Development of a *seed* monolingual ontology.
2. Extension of the ontology to multiple languages.

The **first step** of our approach requires building a small monolingual domain ontology of concepts. For our purposes, we will define the monolingual ontology as a quadruple $O = \langle C, T, E, f \rangle$, where $C$ is a set of concepts (cognitive units of meaning - abstract ideas or mental symbols), $T$ is a set of terms (textual representations of concepts), $E$ is a set of oriented relations (*is-a* relations), such that $\langle C, E \rangle$ is a directed acyclic graph, and $f : T \to C$ is a surjective function from terms to concepts. Note that this implies that polysemy cannot be represented in our ontology. This is for our purposes intentional as we comprehend a domain as an area or part of an area in which the terminology is unambiguous.[1]. Today, lightweight ontologies can be built by reusing existing ontologies or by applying NLP methods for term extraction and ontology learning (Cimiano and Völker, 2005).

In the **second step**, the initial domain ontology is translated using MT and is validated by domain experts. The accuracy of MT is at this moment usually low as the system has not yet been sufficiently trained for the target domain. The resulting multilingual ontology is a 6-tuple $O = \langle C, T, E, f, L, lang \rangle$, where $L$ is the set of languages and $lang : T \to L$ is a mapping from terms to languages. After the validation, the multilingual ontology is integrated with the retrieval system and the available collection of language resources is indexed in terms of the ontology. A set of terms $\{t | lang(t) = \text{language of the resource}\}$ is used for indexing.

The **bootstrapping phase** can be iterated as many times as necessary. The mutual updating procedure is shown in Figure 1. This phase can be further divided into:

1. Adaption of the MT dictionaries
2. Adaption of the multilingual ontology

In the **first step** of the bootstrapping phase, the MT system is adapted to the domain using bilingual substitution rules of form $t_{L_1} \to t_{L_2}$ extracted from the multilingual ontology and satisfying the condition $f(t_{L_1}) = f(t_{L_2})$, where $t_{L_1} \in T_{L_1}, t_{L_2} \in T_{L_2}$ and $T_{L_n}$ is defined as $T_{L_n} = \{t | lang(t) = L_n\}$. For MT systems that translate using an interlingua, the term on the left hand side of a rule is a term in the language of the interlingua and the term on the right hand side is a term in any other supported language. For bilingual MT systems all combinations of terms are exploited and used for the generation of the translation rules. Supplying MT with rules extracted from the ontology can be also useful when a domain is accessed from a general-purpose search engine. IR systems can be equipped with a classification component that can: calculate the most probable domain of a document, select the most suitable domain ontology available, and extract the rules for adaption of the MT system.

For the **second step** of the bootstrapping phase, let us assume that the content stored in our system grows over time. Each time a new learning resource is submitted, it is indexed and put into the document collection. The submitted learning resource may be a translation of an already existing resource stored in the collection. Such parallel texts can be automatically recognized (Resnik and Smith, 2003) and used by the machine translation system for training.[2]

[1] Note that this assumption is not always true.

[2] Most of the statistical MT systems require parallel corpora

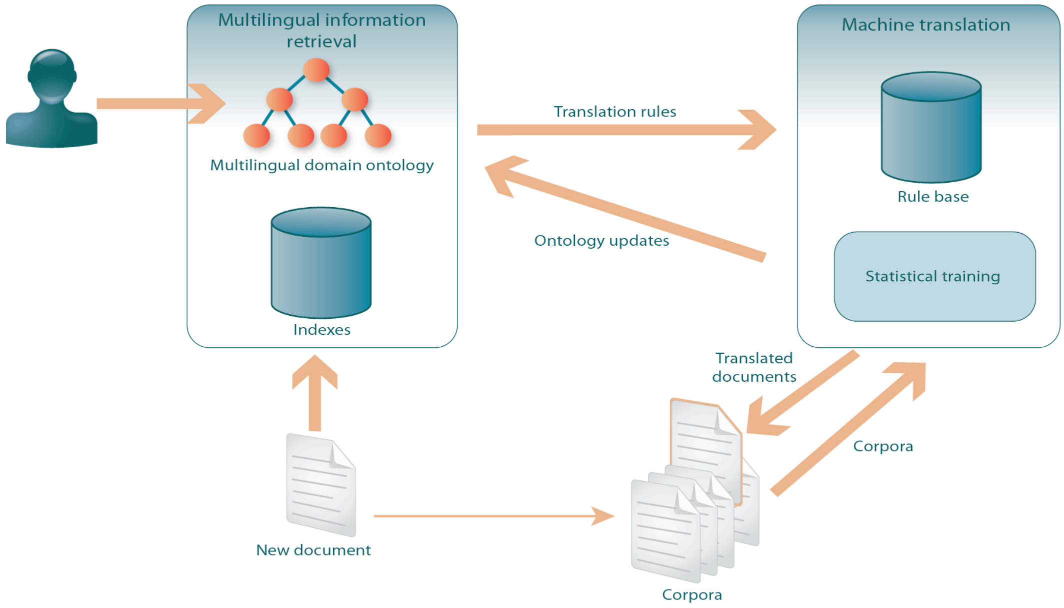


Figure 1: Collaboration of CLIR and MT. Translation rules are extracted from the multilingual ontology and are used to adapt the MT system. New terminology discovered in the statistical training phase is sent to the CLIR system which adapts the multilingual ontology. The updates are validated by a domain expert.

The output of the statistical training is a set of quadruples of the form $(t_{L_1}, t_{L_2}, conf, lang_q)$, where $conf$ is the confidence measure of translating term $t_{L_1}$ to $t_{L_2}$ estimated from text and $lang_q : T \rightarrow L$ is a mapping from terms to languages. The statistical model of the MT system is updated and the quadruples are sent to the CLIR system which uses the following algorithm to update the ontology:

```
Algorithm: Update ontology
Input: Multilingual ontology $O = \langle C, T, E, f, L, lang \rangle$,
a set $Q$ of quadruples of form $(t_{L_1}, t_{L_2}, conf, lang_q)$.
Output: An updated ontology $O' = \langle C, T', E, f', L, lang' \rangle$.
1.  $T' := T, f' := f, lang' := lang, \tau :=$ arbitrary value from $[0, 1]$
2.  for each $(t_{L_1}, t_{L_2}, conf, lang_q) \in Q$ do
3.    if $lang_q(t_{L_1}) \in L \wedge lang_q(t_{L_2}) \in L \wedge conf \geq \tau$ then
4.      if $t_{L_1} \in T_{L_1} \wedge t_{L_2} \notin T_{L_2}$ then
5.        $T' := T' \cup t_{L_2}$
6.        $f' := f' \cup (t_{L_2}, f(t_{L_1}))$
7.        $lang' := lang' \cup (t_{L_2}, lang_q(t_{L_2}))$
8.      end if
9.      if $t_{L_2} \in T_{L_2} \wedge t_{L_1} \notin T_{L_1}$ then
10.       $T' := T' \cup t_{L_1}$
11.       $f' := f' \cup (t_{L_1}, f(t_{L_2}))$
12.       $lang' := lang' \cup (t_{L_1}, lang_q(t_{L_1}))$
13.     end if
14.   end if
15. end for
16. return $O' = \langle C, T', E, f', L, lang' \rangle$
```

The algorithm requires one pass through the set of quadruples $Q$ (line 2). During initialization a sufficiently high value of parameter $\tau$ is set (line 1). Each quadruple is first tested for the compatibility with the ontological language set and for its confidence (line 3). Later, it is checked whether the terms suggested by MT can be mapped to the ontology (lines 4 and 9). The ontology is then updated using the components of the quadruple (lines 5-7 and 10-12). Finally, the algorithm assembles the new ontology (line 16). When the ontology is updated, domain terminology administrators are made aware of the updates by the system and, if necessary, modifications can be performed (for example, new concepts should be added or better translation than the one proposed exists). Performed validation causes new pairs of rules $t_{L_1} \rightarrow t_{L_2}$ to be extracted from the validated part of the ontology and to be submitted back to the rule base of the MT system. As the amount of content grows, the system bootstraps and the performance of both MT and CLIR is improved.

## 4. Application in human genetics

In this section, we describe an application of the method of Section 2 in the context of the Eurogene project, which provides an eLearning system for sharing learning resources in human genetics.[3] The learning resources are submitted to the system typically in the form of slides, books and research articles represented in a variety of formats including Portable Document Format, Word, Power Point and many others. The Eurogene system also supports multimedia resources, such as images and videos in a number

for training, however there have been research studies that investigated learning of multilingual terminology from non-parallel texts, such as in (Fung and Mckeown, 1997).

[3] The system can be freely accessed at http://eurogene.open.ac.uk/

of formats. Resources can be handled in nine European languages[4], which are English, German, French, Spanish, Italian, Greek, Dutch, Czech and Lithuanian. More than 30 universities and other institutions located mainly across Europe, but also in non-European countries are actively contributing to this collection.

In Eurogene, the initial genetic ontology was developed by merging six monolingual ontologies[5] that contained a descriptive, but not too extensive, terminology of the domain. This ontology was translated into the above nine European languages (English is used as an interlingua, i.e. it is used to label the names of concepts) by domain experts and an upper-level ontology has been inferred using Unified Medical Language System (UMLS). A more comprehensive description of the ontology building process can be found in (Zdrahal et al., 2009).

The upper-level ontology helps to organize concepts from a relatively flat structure into a concept hierarchy, which is represented in the Simple Knowledge Organization System (SKOS) format which satisfies our definition of the ontology from the previous section. Figure 2 shows how a genetic concept *linkage analysis* is represented in our ontology.

The multilingual ontology was then integrated with the CLIR system. Since then, available content is being annotated. Textual resources are annotated automatically, multimedia resources are annotated manually, but the annotation procedure is guided by the ontology.

In the first part of the bootstrapping phase, rules were extracted from the multilingual ontology to adapt the MT system as described in the previous section. This typically helps to improve the performance of MT. For example, before the adaption, our system wrongly translated the English collocation *linkage analysis* to French as *analyse de triglerie*, whereas since the rule *Linkage analysis* → *Analyse de liasion* was extracted from the part of the ontology in Figure 2 and it was put into the MT rule base, the system has correctly translated the term as *Analyse de liasion*.

The CLIR system is powered by Lucene extended with a dedicated query parser that allows the user to combine terminological and full-text queries. Queries can be expressed in any of the available languages, and the results can be filtered by a subset of the available languages. Queries are mapped to a language independent representation using the ontology. The CLIR system can also be used during query composition to visualize the concept hierarchy and to interactively control query expansion for broader and/or narrower terms (Figure 3), thus utilizing the benefits of ontology-based retrieval.

A hybrid system developed by SYSTRAN is used for MT tasks, i.e. for the MT of resources and also for the learning of relations from parallel texts (SYSTRAN, 2009). The

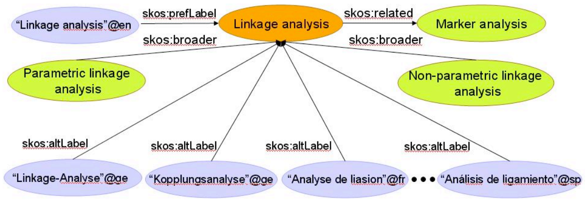
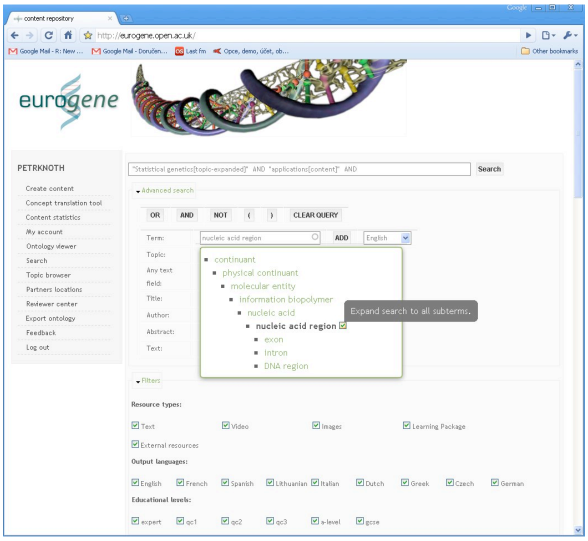


Figure 2: Representation of a concept *linkage analysis* in the multilingual ontology. The preferred label of this concept is the English version *Linkage analysis*. The concept has a two alternative representations in German (*Linkage-Analyse* and *Kopplungsanalyse*).[7] The representation in French is *Analyse de liasion* and in Spanish *Analisis de ligamiento*. The concept Linkage analysis is a broader concept for *Parametric linkage analysis* and *Non-parametric linkage analysis*, and it is related to a concept *Marker analysis*.



"Statistical genetics[topic-expanded]" AND "applications[content]" AND

Term: nucleic acid region English

Resource types:

☑ Text ☑ Video ☑ Images ☑ Learning Package

☑ External resources

Output languages:

☑ English ☑ French ☑ Spanish ☑ Lithuanian ☑ Italian ☑ Dutch ☑ Greek ☑ Czech ☑ German

Educational levels:

☑ expert ☑ qc1 ☑ qc2 ☑ qc3 ☑ a-level ☑ gcse

Figure 3: User interface of the Eurogene CLIR system. The CLIR system allows to control the expansion for broader/narrower terms.

CLIR and MT systems communicate using SOAP messages that allow the sending of extracted translation rules from CLIR to MT, and the sending of newly proposed translations from MT to CLIR. When newly proposed translations are received by CLIR, the ontology is updated using the algorithm in Section 2. Domain experts then perform terminology validation which is supported by the system and results in sending new translation rules to the MT rule base. This synchronization provides a mechanism for continuous semi-automatic adaption of both CLIR and MT systems.

## 5. Performance analysis

The performance of the proposed method and its impact on the resulting CLIR and MT systems can be influenced by a number of factors. These include mainly the suitability of the multilingual ontology for the target domain,

[4] While CLIR allows to pose queries and receive results in any of the mentioned languages, MT is limited to language pairs supported by the Systran system. Please also note that MT is not applied to images and videos.

[5] Published by the University of Washington in Seattle, National Institute of General Medical Sciences in Bethesda, Elsevier, Oracle ThinkQuest, University of Michigan and Centre for Genetics Education in Sydney

the amount of domain corpora available in the statistical phase, the performance of the multilingual keyword extraction system and the validity of the judgements performed by domain experts in the ontology refinement process. Given the number of possible error sources, it seems much more sensible to make sure that the method satisfies certain properties rather than performing a quantitative evaluation that would be biased by too many components.

One of the important properties that the proposed method in Section 3 should have is that the performance of both CLIR and MT should never decrease as a result of any bootstrapping iteration. Let us assume that the initial ontology has been validated by domain experts, so that it does not include any spurious translations. There are now two tasks which could have a negative impact on the performance of the CLIR or MT systems. These tasks correspond to 1) the update of the MT rule base and 2) the update of the multilingual ontology as described in Section 3.

If we assume that our domain is sufficiently small, so that no domain specific term appearing in the multilingual ontology is polysemous in our collection, then updating the dictionary of the MT system may either improve or not change the precision of the MT system. Since it is not possible to extract a spurious translation rule from the multilingual ontology, the resulting MT system cannot perform worse than before the update.

It is essential to expect that the statistical training phase described in Section 3 may produce quadruples describing translations that are in fact invalid and may thus introduce errors to the ontology. However, since all the updates must be validated by domain experts before they can be used by the CLIR system, it is possible to assume that no errors are introduced. This is in reality difficult as humans are in fact vulnerable to introducing errors. Thus, the quality of the ontology used by CLIR can deteriorate only under the condition that an error has been introduced by a domain expert.

To summarize, if all the above mentioned conditions are met, the method is guaranteed to improve or in the worst case not to worsen the performance of the CLIR and MT systems after each iteration.

## 6. Implications for eLearning

This paper showed that current eLearning applications supporting CLIR can also easily adopt MT and tailor it for their domain. In addition, the synergy of CLIR and MT may help to improve the performance of both. The main reason why the method is particularly useful in eLearning is that we should expect that the users of eLearning applications will very often use domain terminology as a part of their submitted queries, thus the added value will become more noticeable than in other contexts.

The paper brought the following contribution:

- Development of a new method for facilitating cross-language retrieval and machine translation by multilingual domain ontologies.

- Development of a real-world eLearning application enhanced by the use of the presented method.

## 7. Conclusion

Multilingual ontologies are particularly suitable for domains where terminology is used for query composition, such as in eLearning. They can be used as a synchronization component for domain adaption of CLIR and MT systems. In addition, the solution is easily readable and adjustable by humans and does not preclude the use of statistical approaches for terminology extraction when a large corpora is available. In the future, publishing of multilingual ontologies on the Web in a standard format may allow an application to decide which domain ontology to use for query expansion and for adaption of the MT system based on the context of the query. This may be helpful when a user accesses a specific domain from a general-purpose search engine.

## 8. References

Philipp Cimiano and Johanna Völker. 2005. Text2onto - a framework for ontology learning and data-driven change discovery.

Susan T. Dumais. 1997. Automatic cross-language retrieval using latent semantic indexing.

David Eichmann, Miguel E. Ruiz, and Padmini Srinivasan. 1998. Cross-language information retrieval with the umls metathesaurus. In *In: Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 72–80.

Pascale Fung and Kathleen Mckeown. 1997. Finding terminology translations from non-parallel corpora.

Lothar Lemnitzer, Cristina Vertan, Alex Killing, Kiril Ivanov Simov, Diane Evans, Dan Cristea, and Paola Monachesi. 2007. Improving the search for learning objects with keywords and ontologies. In Erik Duval, Ralf Klamma, and Martin Wolpers, editors, *EC-TEL*, volume 4753 of *Lecture Notes in Computer Science*, pages 202–216. Springer.

LTfLL. 2008. Language technology for lifelong learning (ltfll).

Wen-Hsiang Lu, Ray S. Lin, Yi-Che Chan, and Kuan-Hsi Chen. 2008. Using web resources to construct multilingual medical thesaurus for cross-language medical information retrieval. *Decis. Support Syst.*, 45(3):585–595.

Zhiyong Lu, Won Kim, and W. John Wilbur. 2009. Evaluation of query expansion using mesh in pubmed. *Inf. Retr.*, 12(1):69–80.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29:349–380.

SYSTRAN. 2009. Systran's machine translation technology url: http://www.systran.co.uk/systran/corporate-profile/translation-technology.

Zdenek Zdrahal, Petr Knoth, Trevor Collins, and Paul Mulholland. 2009. Reasoning across multilingual learning resources in human genetics. In *Proceedings of ICL 2009*.