Using NLP Technology in CALL

Cara GREENE, Katrina KEOGH, Thomas KOLLER,
Joachim WAGNER, Monica WARD, Josef VAN GENABITH
School of Computing, Dublin City University
Dublin 9, Ireland
{cgreene, kkeogh, tkoller, jwagner, mward, josef}@computing.dcu.ie

Abstract

This paper outlines the research and guiding research principles of the (I)CALL group at Dublin City University, Ireland. Our research activities include the development of (I)CALL systems targeted at a variety of user groups including advanced Romance language learners, intermediate to advanced German learners, primary and secondary school students as well as students with L1 learning disabilities requiring a variety of system types which cater to individual user needs and abilities. Suitable CL/NLP technology is incorporated where appropriate for the learner.

1 Introduction

This paper reports on the guiding research and methodological principles and the projects being carried out in the ICALL group at the National Centre for Language Technology based in the School of Computing, Dublin City University. The group aims to leverage currently existing NLP tools in CALL applications for several different learner groups, from ab-initio primary school children to advanced adult learners. Section 2 gives a brief introduction to the composition of the group. Section 3 reviews the group's guiding technological principles. A summary of the four main research strands is presented in Section 4 and Section 5 concludes.

2 Team Composition

The ICALL group is composed of six Computational Linguists and Software Engineers with an interest in CALL. The group's expertise includes Natural Language Processing (NLP), Software Engineering (SE), Computer Science, Pedagogy and CALL.

3 Research and methodological principles

The ICALL group is working on materials for several different learner types. These include advanced speakers of a Romance language who want to learner another Romance language, intermediate to advanced speakers of German (with English L1), Irish primary school children learning Irish and German and students with L1 learning difficulties in secondary schools in Ireland.

The ICALL group has a common set of guiding principles that underpin its research strands. The order in which these principles are presented here do not reflect an inherent importance or ranking. The principles are (i) reuse of existing NLP resources, (ii) reuse of existing CALL research experience, (iii) user-centred design and evaluation and (iv) interdisciplinarity (HCI, SE). We aim to re-use existing CL/NLP technologies where possible and to avoid "re-inventing the wheel". All too often, ICALL projects develop materials from scratch, without building upon existing resources. This may be due to the desire to demonstrate that a certain CL/NLP technology lends itself to CALL, or lack of knowledge of or access to existing CL/NLP resources. "Re-inventing the wheel" is often undesirable for several reasons. The resources (e.g. technical expertise and time) required to develop the materials may not be available. There may not be ample CL/CALL experience in the project group to develop a sophisticated tool set. Sufficiently mature and re-usable CL/NLP resources are now available and can be used as a foundation for developing ICALL resources.

The second principle is the importance of learning from other ICALL projects - successful or otherwise. We try to emulate the successful components of previous projects and to avoid, or at least be aware of, the problems encountered in other projects. For example, Glosser (Dokter & Nerbonne, 1998) was a successful ICALL project that focused on the needs of a particular learner group (i.e. the reading needs of intermediate L1 Dutch learners of French). It used an electronic dictionary (Van Dale Lexicografie homepage), morphological analysis software (XRCE homepage) and corpora to display morphological information, definitions and/or examples of user-selected words from a variety of texts. One reason for the success of Glosser was that it focused on a specific need and used NLP technology to produce a user-friendly ICALL resource for the learner. Another prominent ICALL project was the FreeText (FreeText, 2001) project. This focused on intermediate learners of French and included both

text and audio components. One of the FreeText group's recommendations for future projects (Vandeventer Faltin, 2003) was that more time should have been allocated to user evaluation and that it should have taken place earlier and more frequently during the project.

Our third guiding principle is that the driving force in (I)CALL development should be learner needs and preferences. A common criticism of ICALL projects from CALL practitioners is that they tend to be technology-driven at the expense of pedagogy. Given its mostly technology-oriented skill set, the group is acutely aware of the need to look at ICALL development from a CALL and learner point of view, and not just a technical one.

The group also considers it important to incorporate research from Human-Computer Interaction (HCI) and Software Engineering. All too often, these contributory fields to (I)CALL are overlooked. HCI can substantially improve visualisation of the output of CL/NLP resources. Software Engineering can help in the separation of data and processing by ensuring that the content is independent of the NLP tools.

4 Activities of the ICALL Group

We describe a multilingual ICALL system for Romance Languages in Section 4.1. We outline an artificial co-learner in Section 4.2 while the use of ICALL for Irish primary school students learning Irish and German is summarised in Section 4.3. Section 4.4 describes (I)CALL for students with L1 learning difficulties in secondary schools in Ireland.

4.1 Multilingual ICALL for Romance L2s

The Multilingual ICALL system ESPRIT for Romance languages provides resources for French, Spanish and Italian for advanced speakers of at least one Romance language. The idea behind ESPRIT is to leverage the student's existing Romance language knowledge, rather than forcing them to learn a new language from scratch. Most traditional language learning environments and CALL systems assume an ab-initio learner, which can be frustrating for the target learner group. Learners with previous Romance L2 knowledge can progress quicker but may often make incorrect suppositions which need to be highlighted.

The NLP technologies used include a multilingual parser, animated grammar presentations and the creation and use of small, specialised corpora. The robust chart-based island parser is able to parse ill-formed input and to provide detailed feedback. It reuses an error-sensitive parser for Spanish (Koller, 2003) and supports error recognition on both phrase and sentence level. Constructions which differ between the included languages are recognised and displayed. The feedback generated by the parser feeds into animated grammar presentations. These Flash animations provide a dynamic presentation of grammatical properties and processes and allow learner input as a basis of demonstration. Animation playback can be fully controlled by the learner. Small, specialised corpora provide authentic materials in different topics.

An evaluation platform has been created to enable continuous assessment of different components of ESPRIT (during both development and deployment cycles).

We use a software architecture which combines Flash (Flash, 2004), XML (XML, 2004), Perl and PHP in order to integrate cutting-edge visualisation components, flexible database technologies and NLP tools into a highly flexible and modular language learning environment (Figure 1). This software architecture supports a platform and browser independent representation and a strict separation between language data and processing. In this way, data can easily be reused in different scenarios.

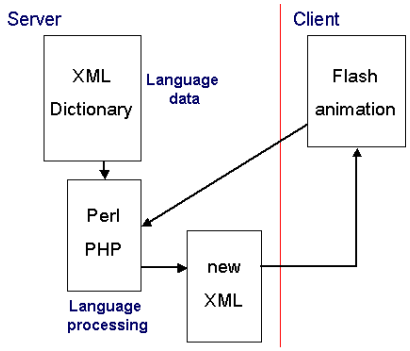

Figure 1: ESPRIT Software Architecture

4.2 Artificial Co-Learner for L2 German

NLP technologies are not perfect and have inherent limitations. Most of these technologies are not designed with language learners in mind, but one can view limitations as an asset, rather than a liability in the learning process. Intermediate and advanced learners can be made aware of the fact that an artificial co-learner based on these technologies is not error-free and that it will sometimes make mistakes. The human learner can then correct the artificial co-learner, improving its knowledge base while at the same time providing a valuable learning environment for the human learner.

We have designed a system for English intermediate L2 learners of German. The system is based on a tool to automatically create "Cognate and False Friends" learning exercises. The NLP technologies used include lemmatisation and corpus processing tools. Figure 2 shows the implemented components of the system (solid lines) and

how we plan to integrate the artificial co-learner (dashed lines). The exercise asks the learner to identify orthographic cognates in a text.

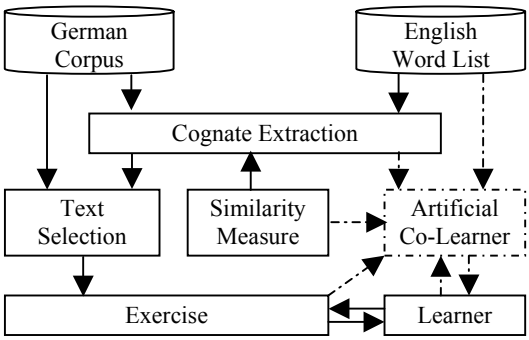

Figure 2: Cognates and Artificial Co-Learner

The co-learner can work with limited resources like word lists, a string-based similarity measure and a list of known cognates. For other types of exercise the co-learner will need access to different linguistic resources: a classification exercise has been implemented in which the learner has to assign words to the classes "true cognate" and "false friend". Here, the artificial co-learner requires knowledge about meanings of words as well as correct translations. A bilingual dictionary can provide the latter information.

4.3 ICALL in the Primary School

Section 4.3.1 discusses an ICALL system for Irish while Section 4.3.2 outlines a system for German. Both are for primary school students in Ireland, ranging from 7 to 13 years of age. Previous ICALL systems have generally concentrated on adult intermediate to advanced learners as they have the linguistic capacity in both their L1 and L2 to interpret and understand the resources presented to them. Younger learners do not have the skills necessary to interpret parse trees nor view the output of morphological analysis.

However, there are several unique features of primary school learners that can be leveraged so that NLP technologies can be deployed successfully in this specific environment. Firstly, the learners' knowledge of their L1 in limited - they do not have the same linguistic range as an adult. Furthermore, the L2 being studied can almost be considered a controlled language in the sense that what the learner knows (or should know) can be quantified if a curriculum and syllabus have been followed. This means that simple Definite Clause Grammars (DCGs) can be written to cater for the anticipated learner language. Learner vocabulary can be controlled and this has positive implications for the deployment of NLP technologies. The CL/NLP engines employed in this context are hidden from the user.

4.3.1 ICALL for Irish in the Primary School

Irish is a compulsory subject in primary schools in Ireland from the age of 4 upwards. Until recently, an audio-lingual method was used to teach the language, but it was very unpopular with both teachers and students. In recent years, a communicative curriculum has been introduced and is proving more popular.

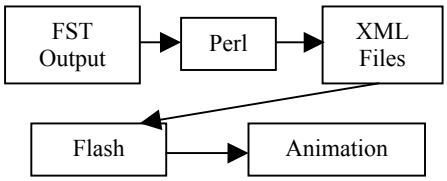

Figure 3: Generating Conjugation Animations

The ICALL system for Irish has two components. One uses the modified output from an existing Finite-State Transducer (FST) morphological engine (Uí Dhonnchadha, 2002) to automatically animate verb conjugations. It uses a combination of FST tools, Perl, XML and Flash to produce the animations (Figure 3).

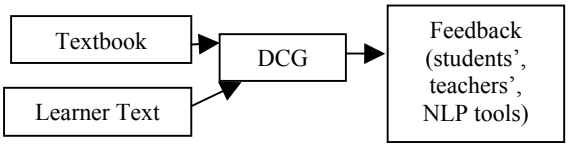

Figure 4: Use of a DCG for Irish

The other component of the system is a tool to analyse learner texts using an incremental DCG. The incremental DCG starts off with a small DCG for Lesson 1 of a textbook and is extended for each subsequent lesson. The DCG can be used to analyse the language used in the textbook as well as student produced text. The findings can then be summarised and used as feedback for the learning process (Figure 4).

4.3.2 ICALL for German in the Primary School

Foreign language teaching has recently been introduced to the senior classes (ages 10-13) of primary schools in Ireland through the Modern Languages in Primary School Initiative (MLPSI).

A fully-fledged curriculum was developed around the National Council for Curriculum and Assessment's (NCCA homepage) draft guidelines and subsequently tagged using Helmut Schmid's TreeTagger (see TreeTagger homepage). The annotated text file was then converted to XML using Perl. Additional information - audio and graphic file references were added at this stage (Figure 5).

The XML annotated corpus can be used as a data source for automatic exercise generation, concordancing and automatic dictionary extraction (along

with morphological analysis). We focus on automatic exercise generation.

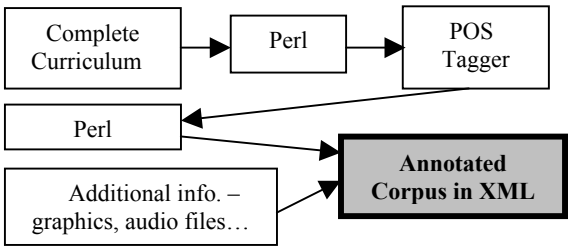

Figure 5: Generating Annotated Corpus in XML

Students in the target context have difficulties with verb inflection as well as nouns being associated with genders. POS tagging identifies verbs and nouns. These feed into 3 types of automatic exercise generation: gap-fill, multiple choice and hangman (guessing an unknown word by guessing letters in the word –with limited number of tries).

Work in progress involves using a FST to generate an audio model for German numerical expressions. A finite number of tokens can be pre-recorded and pieced together for synthesis of (potentially) complex numbers and exercises involving numbers (in definable ranges) can be generated automatically.

Simple CL techniques can be used effectively to meet the learning needs of primary school students and to save time in the generation of learning materials, exercises and feedback.

4.4 (I)CALL and Remedial Learners

Special needs or remedial students are students who suffer from a learning disability. Our research focuses on teenagers in Irish secondary schools who have learning difficulties in their L1 (English). According to the Catholic Communications Office (2003) approximately 10 percent of the Irish population suffer from a learning disability. Students who are in remedial support for their L1 usually have a reading age below their peers. Their main difficulties are with spelling and comprehension of a text. Some students tend to focus on each word separately and not see the word as part of a whole to get the full meaning.

Often, children's remedial primary school CALL software (e.g. Reading for Literacy, 2004) is being used as stopgap remedial software for teenagers in secondary schools. As this software is aimed at children it is inappropriate for young adults and can cause a lack of interest in their support classes.

Our research aims to develop (I)CALL software with appropriate content where CL/NLP technologies are targeted at the specific needs of the remedial student.

5 Conclusion

Our work on integrating CL/NLP technology in CALL applications is guided by four methodological principles: reuse of existing NLP resources, reuse of existing research experience, user-centred ICALL design/evaluation and interdisciplinarity (HCI, SE). The paper illustrates how these principles are manifest in a number of projects.

6 Acknowledgements

This research has been funded by SFI Basic Research Grant SC/02/298, IRCSET Embark Initiative Grant RS/2002/441-2 and DCU.

References

Catholic Communications Office. 2003. *What is a learning disability?* Available at:http://www.cathoiccommunications.ie/ [Accessed: 03 April '04]

D. Dokter & J. Nerbonne. 1998. A Session with Glosser-Rug. In "*Language Teaching and Language Technology"* S. Jager, J. Nerbonne & A. van Essen, ed., pages 88-94, Swets & Zeitlinger, Lisse.

Flash. 2004. Available at: http://www.macromedia.com/software/flash/ [Accessed 10 April '04]

FreeText. 2001. Available at: http://www.latl.unige.ch/freetext/ [Accessed: 10 April '04]

T. Koller. 2003. Knowledge-based intelligent error feedback in a Spanish ICALL system. In *Proceedings of The 14th Irish Conference on Artificial Intelligence & Cognitive Science*. Dublin: Trinity College, 117-121.

NCCA. 2004. NCCA Homepage. Available at: http://www.ncca.ie/j/index2.php?name=currinfo [Accessed: 10 April '04]

Reading for Literacy. 2004. Available at: http://www.englishsoftware.com/ [Accessed: 05 May '04]

TreeTagger Homepage. Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html [Accessed: 20 April '04]

E. Uí Dhonnchadha. 2002. *An Analyser and Generator for Irish Inflectional Morphology Using Finite-State Transducers.* MSc Thesis, Dublin City University, Ireland.

Van Dale Lexicografie Homepage. Available at: http://www.vandale.nl/ [Accessed: 1 May '04]

A. Vandeventer Faltin. 2003. *Grammar Checking for CALL.* Eurocall 2003, Limerick, Ireland.

XRCE Homepage. Available at: http://www.xrce.xerox.com/ [Accessed: 1 May '04]

XML. 2004. Available at: http://www.w3.org/XML [Accessed: 10 April '04]