CL for CALL in the Primary School

Katrina Keogh, Thomas Koller, Monica Ward,
Elaine Uí Dhonnchadha , Josef van Genabith
School of Computing
Dublin City University
Dublin 9, Ireland
{kkeogh, tkoller, mward}@computing.dcu.ie,
Elaine.UiDhonnchadha@dcu.ie, josef@computing.dcu.ie

Abstract

This paper looks at how Computational Linguistics (CL) and Natural Language Processing (NLP) resources can be deployed in Computer-Assisted Language Learning (CALL) materials for primary school learners. We draw a broad distinction between CL and NLP technology and briefly review the use of CL/NLP in e-Learning in general, how it has been deployed in CALL to date and specifically in the primary school context. We outline how CL/NLP resources can be used in a project to teach Irish and German to primary school children in Ireland. This paper focuses on the use of Finite State morphological analysis (FST) resources for Irish and Part of Speech (POS) taggers for German.

1 Introduction

CL/NLP has a lot to offer many disciplines. One particular area of interest is e-Learning for languages or more specifically Computer-Assisted Language Learning (CALL). CALL aims to develop useful learning tools with the focus on the learner. The following sections outline the use of CL/NLP in CALL (also known as Intelligent Computer-Assisted Language Learning - ICALL) for a particular target audience – primary school students in Ireland.

First we review CL/NLP in e-Learning and the case for using CL/NLP in CALL. Next we describe ICALL and the case for its use in primary school. Section 4 goes into detail on the CL/NLP technologies we use for primary school students learning Irish and German.

2 CL/NLP in e-Learning

2.1 CL/NLP – A Broad Distinction

To a first approximation CL/NLP technologies split into two broad categories – A and B. Category A (sometimes referred to as CL proper) typically includes small coverage, proof of concept, often hand-crafted, knowledge- or rule-based systems. They are usually used to test a particular linguistic theory, tend to be of limited coverage and are often quite brittle. Example technologies include DCGs and many (but not all) formal grammar-based parsing and generation systems.

Category B (sometimes referred to as NLP) typically includes broad coverage systems where the lingware is often (but not always – see e.g. FST) automatically induced and processed using statistical approaches. They are usually large scale engineering applications and very robust. Example technologies include speech processing, HMM taggers, probabilistic parsing and FST.

This distinction is, of course, nothing more than a useful over-generalisation with an entire and interesting grey area existing between the two extremes. Khader et al. (2004), for example, show how a wide-coverage, robust rule-based system is used in CALL. In this paper we look at the suitability of type A and B CL/NLP technologies for primary school education, in the context of Ireland in particular.

2.2 e-Learning

CL is generally not to the fore in e-Learning, although it does have a potentially powerful role to play. It can help to enhance the accessibility of online teaching material (particularly when the material is not in the learner's L1), in analysing learner input and the automatic generation of simple feedback. It can also be used with Computer-Mediated Communication (CMC) environments. However, to date, the use of CL/NLP in e-Learning in general has not been a main stream focus of either the Computational Linguistics or the e-Learning community nor has there been much CL/NLP technology transfer into commercially available and deployed systems.

2.3 CALL

Within the domain of e-Learning, the area with the greatest fit and potential deployment of CL/NLP resources is that of Computer-Assisted Language Learning (CALL). This paper focuses on asynchronous e-Learning for natural languages in

the primary school context. CL/NLP resources lend themselves naturally to the domain of language learning, given that the "raw material" in both fields is language. However, attempts to successfully marry the two fields have been limited. Schulze (2003) outlines several reasons for this. Computational Linguists are specifically interested in the use of the computer in analysing, generating and processing language. They are interested in testing out linguistic theories and using the computer to confirm their hypotheses. Researchers in NLP tend to be interested in wide-coverage, robust engineering approaches. For the most part, use of their tools for language learning/teaching applications is not high on their research agenda. A review of COLING papers in the last twenty years reveals that there are very few papers that specifically deal with the use of CL/NLP in language learning. Furthermore, as Schulze (2003) points out, within the unspoken hierarchy that exists in Computer Science departments throughout the world, working with CALL is considered less prestigious than say, working on cryptography. Thus, socio-cultural factors may have played a part in limiting the number of CL/NLP researchers interested in CALL.

From a CALL researcher's or practitioner's point of view, attempts to integrate CL/NLP resources into CALL have not been very successful. Many remain unconvinced about the benefits of using CL/NLP techniques in CALL and whether they can be integrated successfully or not. They sometimes expect an 'all-singing, all-dancing' machine and are disappointed /disillusioned with the results of ICALL research, especially when they incorporate category A CL technologies. CALL practitioners generally come from a language teaching background and are often more interested in pedagogy than technology. Some feel that the technical knowledge required to integrate CL/NLP tools is beyond their scope. They may be wary of claims from CL/NLP developers that a certain CL/NLP resource will be "ideal" for CALL, especially if they have heard such claims before. Even if they are favourably disposed to the use of CL/NLP resources in CALL, it is often very difficult to reuse existing resources, as they demand that a certain (often non-standard) format be used for data (see Sections 4.2 and 5.2 below). Also, the interfaces to the resources may have assumed a techno-savvy or CL/NLP-savvy user, which mitigates against their (re)use.

In summary, apart from notable exceptions (e.g. Glosser (Dokter & Nerbonne, 1998) and FreeText (2001), for various technical and non-technical reasons, CL/NLP resources have not been extensively deployed in main-stream CALL applications.

One of the problems in using CL/NLP resources in CALL materials is that the coverage achieved by the CL/NLP tools has to be broad to be able to handle a general range of learner language. Furthermore, the resources must be robust as learner language will contain input that is not well-formed and this can cause problems for some CL resources. Observations such as these point to type B NLP technologies as being the better type of technologies to employ in the context of language learning. However, below we argue that this is not necessarily the case.

2.4 ICALL in the Primary School

It may be natural to assume that CL/NLP resources customarily lend themselves to intermediate or advanced learners of a language, as they are more likely to have the linguistic competence to understand output generated by CL/NLP resources. Considering the other end of the language-learning spectrum, that of primary school learners, it may be perceived that CL/NLP resources could not be so easily deployed with linguistically less advanced learners - these students will not be interested in viewing concordances, morphological annotations or parse trees.

However, it can be argued that there are certain natural circumstances supporting the use of even type A CL technology in CALL in this environment. Firstly, in comparison to adults, young learners have limited first language (L1) performance (Brown, 1994). The target primary school students are aged between 7 and 13 years (second to sixth class in the Irish primary school system). They tend to produce simpler sentences and have a smaller range of vocabulary than an adult. These L1 features have a number of implications – the students' L1 knowledge further constrains their emerging L2 production. Complex linguistic constructs are less likely to transfer into the target language. Effectively, the target language amounts to a controlled language. Controlled languages are easier suited to type A CL systems and produce better results (Arnold et al., 1994).

Secondly, the students' target language(s) (Irish and German in this context) represent a limited domain or sublanguage. The Irish curriculum is followed in primary schools from the age of 4/5. Students can take German (where it's available) during their senior years of primary school (aged 10-13) and the language domain is limited to a 2 year beginners' curriculum. It is possible to

anticipate students' L2 knowledge, especially since they have been following set curricula. Machine Translation (MT) can be used to highlight an example of the success of sublanguages with CL/NLP. The Météo translation system is used successfully in Canada to translate weather forecasts bi-directionally between French and English (Hutchins and Somers, 1992). The 'weather' sublanguage has a small vocabulary and uses a telegraphic style of writing and omits tense.

Primary school students' L1 and L2 performance characteristics – controlled language and limited domain – imply that some scalability problems that are sometimes encountered in certain type A CL resources can be avoided.

While primary school learners will not be interested in viewing concordances or parse trees – technology can be used but hidden from the learner, to generate exercises and learner feedback and to present students with an animation based on information computed by the underlying CL/NLP engines embedded (but not visible) in the CALL application. In this way the learner will benefit from the technologies but not be confused by linguistic elements that are beyond their capacity as young learners.

3 CL/NLP Resources for CALL

In this paper we look at how CL/NLP resources can be integrated into CALL materials in general, as well as specifically for Primary Schools in Ireland, with a focus on CALL materials for Irish and German. This section will briefly outline how a range of CL/NLP resources can be used in this environment, while later sections will focus on the use of specific CL/NLP resources in more detail.

We return to our dichotomy of A- and B-type CL/NLP systems outlined in Section 2.1. ICALL systems have used a range of technologies, including both type A and type B systems. Examples of type A-like systems include small-scale Lexical Functional Grammar (LFG) –based robust parsers to provide error recognition and feedback (Reuer, 2003) and parsing for viewing sentence structures and error diagnosis (Vandeventer Faltin, 2003). Examples of type B-like systems include a broad-coverage English LFG-based grammar for grammar checking (Khader et al, 2004), the Systran MT system to improve translation skills (La Torre, 1999) and using speech recognition for pronunciation training (Menzel et al, 2001).

It is relatively straightforward to integrate type B (NLP) technology into CALL applications for primary school learners. In Section 4 of this paper we show how broad-coverage FST technology can be used to morphologically analyse word forms or to generate all inflected forms given a root form. Output from a FST morphology engine is fed into an interface engine which sends the information in the appropriate format to an XML/Flash environment for animation (Koller, 2004). The learner input can be collated over time into a learner corpus and later analysed by the teacher to detect common errors amongst students. Part-Of-Speech (POS) taggers can be used to identify the parts of speech in electronic versions of learners' textbooks or a corpus collated around their curriculum (Section 5). The output can then be used for a variety of uses, including the automatic generation of online exercises (e.g. hangman) and together with the FST morphological engine - automatic dictionary extraction.

Mainly due to scalability problems, type A CL technologies can be difficult to deploy in general ICALL systems. However, they can be used in the primary school context quite effectively. As outlined in Section 2.4, the limited linguistic performance knowledge of the learners' L1 and especially their L2 amounts to a 'controlled' language scenario and type A CL technologies can be deployed successfully. Curricula used in primary schools (in Ireland and elsewhere) represent a limited domain in which type A technologies can be highly appropriate. Small coverage DCGs, for example, can be written for the anticipated L2 learner input and can be used to provide immediate feedback to the learner. Problems associated with difficulties in building wider-coverage grammars do not present themselves in this context, as the curriculum is limited.

The are many other potential uses of CL/NLP in this context, but this paper will focus on the FST and POS tagging examples mentioned above.

4 CL/NLP Resources for Irish Primary School CALL

4.1 Background

Irish is a compulsory subject in schools in Ireland. Students generally tend to have a negative attitude towards the language, which hinders learning (Harris & Murtagh, 1999). Until recently, Irish has been taught using the Audio-Lingual method (structural patterns are taught using repetitive drills) and it is only since 1999 that a new communicative curriculum (language teaching is structured around topics in terms of communicative situations) has been developed and integrated. Currently, there are very few CALL resources available for Irish (Hetherington, 2000) and those that do exist may not be as error-free as one would like, are not specifically aimed at

primary school learners and are therefore not tied to the Primary School curriculum which hinders their integration into the classroom.

4.2 A FST-Based Morphological Engine for Irish

Uí Dhonnchadha (2002) has developed an analyser and generator for Irish inflectional morphology using Finite-State Transducers (Beesley and Karttunen, 2003). The FST engine contains approximately 5,000 lexical stems, generates/recognises over 50,000 unique inflected surface forms with a total of almost 400,000 morphological descriptions (due to ambiguous surface forms). The final FST is the result of composing intermediate transducers, each encoding a different morphological process. It is useful to have a record of the morphological processes involved in mapping between lexical (i.e. lemmas and morphological features) and surface forms. By including a marker in the surface form each time a process is applied, a record of the morphological processes involved can be maintained and used in other applications.

The morphological processes covered include: (i) internal mutations such as lenition, ellipsis, stem internal modification and vocal prefixing; (ii) final mutation, such as vowel harmony with suffixes (broadening, slenderising and syncopation); as well as concatenative morphology (prefixing, suffixing).

4.3 Technology - FST, Perl, XML and Flash

Primary school learners are not interested in viewing output generated by a FST Morphology engine. The challenge in CALL applications (particularly in the primary school scenario) is to exploit underlying technology to present information in a manner appropriate to the learner. To this end we developed animation software interfaced with the output generated by the FST engine.

Animation can enhance the learning process and is especially interesting for younger learners. Flash (2004) is a useful software environment to develop animations but it is difficult for non-programmers to use and it is often difficult to use the same animation templates for different inputs. One solution is to use XML (Extensible Markup Language, XML (2004)) files as input into Flash, so that the information displayed is customisable according to the information in the input data file. We outline how animated CALL materials were developed for teaching the conjugation of verbs in the present tense in Irish.

Output from the FST engine is fed to a Perl script which converts the information into a specified XML format. The XML files are then used by Flash to generate the required animation. Figure 1 outlines the software architecture. Figure 2 shows the conjugation of the verb *cuir* (*to put*) in the present tense in Irish. Figure 3 shows modified output from the FST engine to enable automatic animations to be generated (^INF indicates inflectional infix, ^PP indicates inflectional postposition and ^SUF indicates inflectional suffix for Flash).

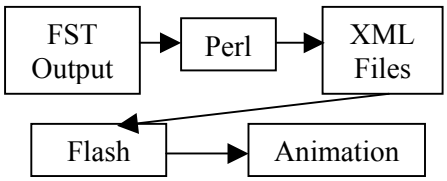

Figure 1: Software architecture

1S	Chuir mé
2S	Chuir tú
3S	Chuir sé/sí
1P	Chuireamar
2P	Chuir sibh
3P	Chuir siad

Figure 2: Conjugation of "c*uir*"

```
PastInd          c^INFuir^PP
PastInd+1P+Pl    c^INFuir^SUF
```

Figure 3: Sample output from FST engine

A section of the XML file that feeds into the Flash program is shown in Figure 4.

```
<verb>cuir</verb>
<stem1>c</stem1>
<stem2>uir</stem2>
<infix>h</infix>
<fir_sg><postpos>mé</postpos></fir_sg>
<sec_sg><postpos>tú</postpos></sec_sg>
<thi_sg><postpos>sé/sí</postpos></thi_sg>
<fir_pl><suffix>eamar</suffix></fir_pl>
<sec_pl><postpos>sibh</postpos></sec_pl>
<thi_pl><postpos>siad</postpos></thi_pl>
```

Figure 4: XML file for Flash program

The animation movie demonstrates that the stem "cuir" is split up into "c" and "uir". Then the infix "h" is inserted between "c" and "uir". Finally the postposition "mé" is added (Figure 5).

Animations can be developed automatically for any verb and morphological process known to the FST engine, as all morphological operations are

coded for Flash. This removes the necessity of hand-coding animations and reduces the risk of human error.

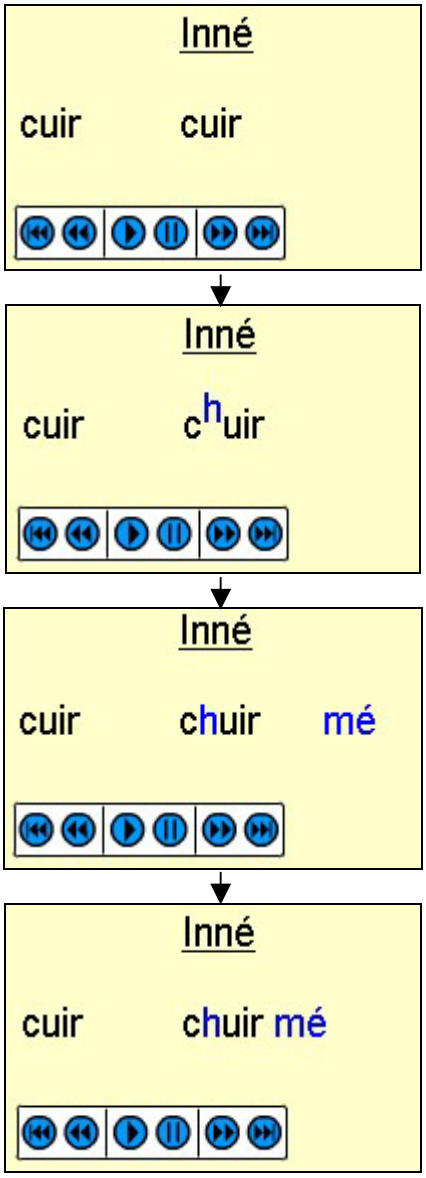

Figure 5: Snapshot sequence from animation movie for past tense 1st person singular for the verb '*cuir*' in Irish (*Inné* means *yesterday*)

The Flash-based interface dynamically displays XML data. It reads in XML data at runtime and generates an animation. Learners have full control over the animation. They can play, stop, rewind and skip through the animation. Further interaction is provided via menus to choose specific conjugations (e.g. number, person and tense.)

The FST-Flash interface is language-independent. The XML files contain detailed information about the different string operations and the corresponding targets. The only operations known to the Flash interface are insert, delete and replace. In this way, the animation of language data is abstracted from linguistic terms like prefixation, suffixation or lenition, thus avoiding the problem of varying definitions of these terms in different languages. The transformation of the (linguistically tagged) output from the morphology engine to the XML data necessary for animated presentation is done by Perl scripts which can be tailored specifically to each combination of language and output of a NLP tool.

5 CL/NLP Resources for German Primary School CALL

5.1 Background

German is gradually being integrated into Irish primary schools through the Modern Languages in Primary School Initiative (MLPSI), which has been running since 1998. At present, over 300 schools in Ireland are involved in the MLPSI.

German is taught during the senior two years of the primary school cycle (children aged 10-13). Irish students do not receive any instruction in Modern Foreign Languages (MFL) up until this point (Irish is not considered a MFL). The communicative curriculum we developed is based on a draft curriculum which was developed by the National Council for Curriculum and Assessment (NCCA) (NCCA, 2004) for teachers participating in the MLPSI.

The integration of type A CL technology into CALL in this environment is ideal. The target language is restricted to a beginner's curriculum. This represents a restricted domain. Sentence constructions are simple with few structures that could present coverage or ambiguity difficulties to CL systems. Given that the target language is German, many CL tools are available for almost every aspect of language processing.

In this section we will focus on the use of type B NLP technology in this environment to *meet the needs* of students learning German. These needs have been researched qualitatively through observation during German language lessons in a primary school in Ireland during the school year 2003/4. Irish students are native English speakers (some are also native Irish speakers) and as such are unfamiliar with nouns being associated with genders as in German. These students also require extra practise with inflecting verbs correctly. Having being asked '*Wie heißt du*?', students will often respond with '*Ich heißt ...*', for example. We present the use of a POS tagger in the development of a tailored corpus which subsequently feeds into the automatic generation of exercises.

5.2 Technology – POS tagging, Perl and XML

CALL courseware generally presents users with exercises to complete after they have studied a particular topic. These are usually static in content and are very time consuming to develop over the full set of language topics. Students are usually presented with a small number of exercises, which they will have completed in their entirety and become familiar with in a limited space of time. Larger sets of exercises prove beneficial in providing variety for the student – they will not be presented with the same set of exercises each time they visit a topic. In addition, some students will complete exercises faster than others. This puts pressure on slower students to keep up and on teachers to provide alternative work to keep faster students occupied. Larger sets of exercises mean that exercise selection can be randomised so that students are presented with new material each time they visit the courseware; slower students will feel less pressure to work at a faster pace when faster students complete additional exercises within the same language topic and teachers will not be required to provide alternative material.

CL can significantly reduce the time needed to generate sets of exercises around language topics.

A complete curriculum was developed around the NCCA guidelines and tagged using Helmut Schmid's TreeTagger (see TreeTagger homepage). The annotated text file was then automatically converted to XML using Perl. The corpus is divided into separate XML files for each language topic. Additional information - audio and graphic file references were added manually to each topic file at this stage.

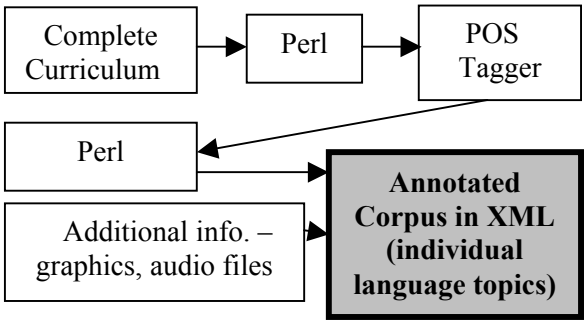

Figure 6: Generating annotated corpus in XML

Once the annotated corpus has been converted to XML it can feed into a number of applications such as lesson generation, automatic dictionary extraction, a concordancer and automatic generation of various exercise types. In focusing on the latter, we are particularly interested in the verbs, articles and nouns that the POS tagging identifies. Inflection and article-noun combinations can be practised when a student chooses the correct verb ending or article from a selection or types in the correct answer. A version of hangman (a game where students try to guess an unknown word by guessing letters in the word - they only get a certain number of chances for incorrect answers after which the game ends) can also be played with article-noun combinations. By simply specifying the topic section in the curriculum and the type of game, exercises are automatically generated. Each particular exercise is randomised so that the user is presented with a new variant of the problem each time they attempt an exercise or game.

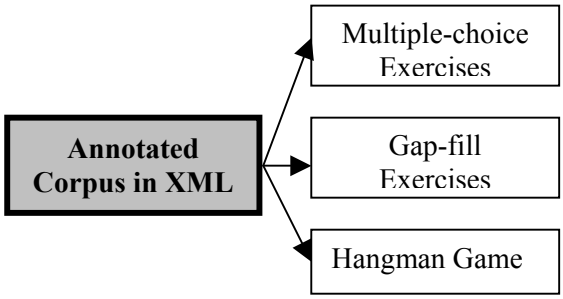

Figure 7: Automatic Exercise Generation

Previous work in automatic exercise generation from corpora highlighted a number of potential pitfalls (Wilson, 1997). Most importantly, the language in the corpus used is best when the linguistic quality of the texts is appropriate for learning a language. Long and complex sentences are best avoided. Our design employs a corpus collated and tailored around the learner's curriculum, thus avoiding this pitfall.

The benefit of using CL resources here is similar to the situation in the Irish context. Exercises can be developed automatically for any verb or noun phrase within the curriculum and provide variety for the user. This removes the necessity of hand-coding each exercise and reduces the risk of human error.

6 Conclusion

It is difficult to integrate CL/NLP resources into CALL, especially as these resources are not generally designed with a CALL audience in mind. However, there are environments where they can be successfully integrated, especially if used in an imaginative and useful way. The technology does not have to be particularly revolutionary or complex - what is important is that it is appropriately deployed.

This paper outlined how two NLP resources can be used in the development of CALL resources for primary schools. It is novel in the ICALL world to employ CL/NLP technologies for young learners, especially when they are beginners in learning a

language. We outlined how the output of a FST engine can feed into the generation of Flash animations for Irish verb conjugations. We showed how a POS tagger can be used to annotate a curriculum to produce a corpus which can in turn be used to automatically generate exercises. Both of these initial modules will be comprehensively deployed and evaluated in the classroom during the coming school year (Sept. 2004-June 2005). Future modules will include type A CL technology like DCGs and will take advantage of the controlled languages and limited domains which exist in the primary school environment. Each module of the overall system is being developed so that concurrent evaluation can be carried out.

This paper highlighted the point that even though neither of these NLP resources was developed with CALL applications in mind, when combined with relatively straightforward programming and interface techniques, they can be used fruitfully in a CALL environment.

7 Acknowledgements

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

References

D. Arnold, L. Balkan, S. Meijer, R. L. Humphreys and L. Sadler. 1994. *Machine Translation - An Introductory Guide* NCC Blackwell Ltd., London, USA.

K. R. Beesley and L. Karttunen. 2003. Finite-State Morphology. Series: (CSLI-SCL) Center for the Study of Language and Information.

H. D. Brown. 1994. *Principles of Language Learning and Teaching.* Prentice-Hall Inc, London, Sydney, Toronto, Mexico, New Delhi.

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

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

FreeText. 2001. FreeText Homepage. Available at: http://www.latl.unige.ch/freetext/ [Accessed 10 April 2004]

D. Hetherington. 2000. *Computer Resources for the Teaching of Irish at Primary and Secondary Levels.* Language Centre NUI Maynooth, Ireland.

J. Harris and L. Murtagh. 1999. *Teaching and Learning Irish in Primary School.* ITE, Dublin.

W. J. Hutchins and H. L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press, London.

R. Khader, T. Holloway King and M. Butt. 2004. *Deep CALL grammars: The LFG-OT experiment*. DGfS 26.Jahrestagung, Mainz, Germany.

T. Koller. 2004. *Creating user-friendly, highly adaptable and flexible language learning environments via Flash, XML, Perl and PHP*. Presentation at the EuroCALL SIG-LP workshop "Innovative Technologies and Their Didactic Application", Vienna, September 2004.

M. D. La Torre. 1999. A web-based resource to improve translation skills. *ReCALL,* 11(3): 41-49.

W. Menzel, D. Herron, R. Morton, D. Pezzotta, P. Bonaventura, and P. Howarth. 2001. Interactive pronunciation training. *ReCALL*, 13(1): 67-78.

NCCA. 2004. *National Council for Curriculum Assessment (NCCA) Homepage*. Available at: http://www.ncca.ie/j/index2.php?name=currinfo [Accessed: 10 April 2004]

V. Reuer. 2003. Error Recognition and Feedback with Lexical Functional Grammar. *CALICO,* 20(3): 497-512

M. Schulze. 2003. *AI in CALL: Artifically Inated or Almost Imminent?* WorldCALL 2003, Banff, Canada.

TreeTagger Homepage. Available at: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html [Accessed: 20 April 2004]

E. Uí Dhonnchadha. 2002. *An Analyser and Generator for Irish Inflectional Morphology Using Finite-State Transducers.* Msc Thesis.

A. Vandeventer Faltin. 2003. Natural language processing tools for computer assisted language learning. *Linguistik Online* 17, 5/03

E. Wilson. 1997. The Automatic Generation of CALL Exercises from General Corpora. In "Teaching and Language Corpora" A. Wichmann, S. Fligelstone, T. McEnery, and G. Knowles, ed., pages 116-130, Addison Wesley Longman, London.

XML. 2004. Extensible Markup Language. Available at: http://www.w3.org/XML [Accessed 10 April 2004]