Williams, B. (forthcoming) The phonetic manifestation of stress in Welsh. In Hulst, H. van der (ed.), *Word Prosodic Systems in the Languages of Europe.* Berlin: Mouton de Gruyter.

Centre for Language and Communication
Cardiff University
P O Box 94
Cardiff CF1 3XB
mueller@cardiff.ac.uk

JOURNAL OF CELTIC LANGUAGE LEARNING, 4, 33-41 (1999)

Preparation Aid for Text Based Irish Course

ANNETTE MCELLIGOTT
Ollscoil Luimnigh

AND GEARÓID Ó NÉILL
Ollscoil Luimnigh

In this paper is described facilities for the preparation of text based exercises for delivery in a computer aided language system. Supplied texts are annotated automatically and then revised by the teacher. An index of word forms is built. There are both computer generated and teacher supplied exercises.

In text based exercises, a teacher specifies a text, typically augmented with notes on words, phrases and grammatical points. Questions are posed on various aspects of the text, such as, vocabulary, grammar and general understanding. In this paper are described facilities for the preparation of text based exercises for delivery in a computer assisted language learning system.

The system has, as its kernel, a knowledge base comprising a dictionary, spell checker and grammatical rules (McElligott and Ó Néill 1996,1995). The dictionary is an extension of *An Foclóir Beag*, (Roinn Oideachais, 1991) giving meanings and inflected forms. There are two classes of user: the person (the teacher) who annotates the text and the person (the student) who uses the annotated text for learning the language.

Annotating the text is achieved through a number of processes. These processes are initiated by the teacher. The results of each process are submitted to the teacher for

correction, modification or for additional information. Texts may be linked to form a series of texts. Such information is also used in the analyses of further texts. The following serves to illustrate some points in the paper.

> Bhí dhá mhoncaí ag siúl ar an mbóthar lá. Chonaic ceann acu cnó ar an mbóthar agus ar sé "chím cnó". Leis sin, léim an dara moncaí agus thóg sé an cnó. "Tabhair dom an cnó sin", arsa an chéad mhoncaí, " is liomsa é mar is mise is túise a chonaic é". "Ní leatsa é", arsa an dara moncaí, "is liomsa é mar is mise a fuair é". Bhíodar ag achrann mar sin ar feadh tamaill nuair chonaiceadar fear an phoist ag teacht chucu. "Cad tá oraibh", ar sé, "nó cad chuige go bhfuil sibh ag troid?". D'inis siad an scéal dó. "Cá bhfuil an cnó?", arsa an fear. Fuair sé an cnó agus bhris sé é. Thug sé plaosc amháin don mhoncaí a chéad-chonaic an cnó agus an plaosc eile don mhoncaí a fuair an cnó. Ansin mar gheall ar a saothar, d'ith sé féin an eithne.
> (MacGiolla Phádraig 1953, p. 76) (with some changes)

First the text is split into tokens. To facilitate tokenisation the teacher may indicate whether the text is general or is in a specialised domain. For example, an e-mail address might contain @ and . as part of a token. Two representations result, the original text - the surface representation - and the tokenised text - the internal representation. The first sentence in the text sample in tokenised form is simply the words with the punctuation mark.

As the text is split into tokens, a check is made to determine whether the word is in the dictionary. If the word is not in the dictionary, the teacher is alerted and may alter the word, replace the word or make an entry for the word,

Source Text:	*Bhí dhá mhoncaí ag siúl ar an mbóthar lá.*
Tokenised Representation:	*tokens(['Bhí','dhá','mhoncaí','ag','siúl','ar','an','mbóthar','lá','.'])*

Figure 1: Source Text and Tokenised Representation

together with any inflected forms and derived words. These entries are added to the teacher dictionary. For example, *chéad-chonaic* is not in the dictionary. When the student comes to use the system, clicking on *céad-chonaic* will display any information which the teacher has supplied.

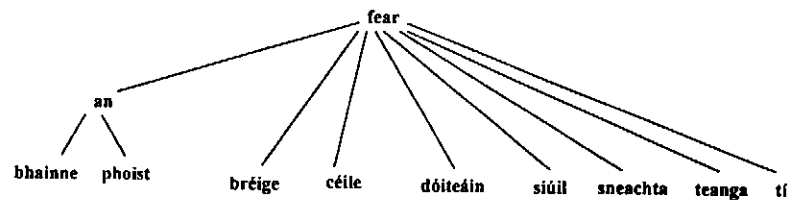
The next step is compound processing which works in two stages. First a search tree is created with the set of compounds from *An Foclóir Beag* forming the initial nodes of the tree. Each compound is inserted into a database as seen in Figure 2. The first attribute in this database is the list of atoms forming the compound, the latter two attributes will be used later for the compound recognition phase and the syntactic tagging phase, respectively.

```
compound_elements_db( [ fear, an, bhainne ], 'fear_an_bhainne', 'FA').
compound_elements_db( [ fear, an, phoist ], 'fear_an_phoist', 'FA').
compound_elements_db( [ fear, bréige ], 'fear_bréige', 'FA').
compound_elements_db( [ fear, céile ], 'fear_céile', 'FA').
compound_elements_db( [ fear, dóiteáin ], 'fear_dóiteáin', 'FA').
compound_elements_db( [ fear, siúil ], 'fear_siúil', 'FA').
compound_elements_db( [ fear, sneachta ], 'fear_sneachta', 'FA').
compound_elements_db( [ fear, teanga ], 'fear_teanga', 'FA').
compound_elements_db( [ fear, tí ], 'fear_tí', 'FA')
```

Figure 2: Compounds associated with the headword *fear* from *An Foclóir Beag*

From the relations in this database is built a compound tree. The tree for the data in Figure 2 can be represented diagramatically and textually as shown in Figure 3.

Graphical Representation:



Textual Representation:

compound_tree(fear, [fear, [an, [bhainne, []], [phoist, []]], [bréige, []], [céile, []], [dóiteáin, []], [siúil, []], [sneachta, []], [teanga, []], [tí, []]]).

Figure 3: Graphical and Textual Representations of a Compound Tree

The creation of the textual representation of the compound tree can be summarised as follows (Sutcliffe et al 1996):

1. Take the head of the first argument in the first clause of the *compound_elements_db*.
2. Find all clauses whose first word is the same as this, and make a list of their first arguments, minus the first word.
3. Expunge the original clauses.
4. Extract all lists from the list which start with the same word.
5. Append to the tree a successor tree for those lists.
6. Repeat steps 4 and 5 until all elements of the list have been used.
7. Assert the finished structure into the *compound_tree* database.

The second step relating to compound processing concerns the recognition of compounds in an input utterance. On receipt of an input utterance the compound recognition software works down the sentence searching for potential compounds at each level using the *compound_tree* database. If a word does not form a compound it is copied to the output string. If a word does form part of a compound the search procedure continues to obtain the longest compound that the current word is the head. Once determined the text forming the compound is replaced with the second argument of the *compound_elements_db* (cf. Figure 2.) and then appended to the output string. This replacement is used for internal representation purposes only. The internal representation is also augmented with the syntactic category associated with the compound as found in the *compound_elements_db*, that is, the third argument in this database (cf. Figure 2). The overall recognition phase is called by the predicate *determine_compounds* that takes as input an atomised string and returns as output the string with compounds differentiated with the underscore symbol, for example,

determine_compounds(['Bhíodar','ag','achrann','mar','sin', 'ar','feadh','tamaill','nuair','chonaiceadar','fear','an', 'phoist','ag','teacht','chucu','.'], 'Bhíodar ag achrann mar sin ar feadh tamaill nuair chonaiceadar fear_an_phoist ag teacht chucu.').

If the teacher identifies a compound phrase in a passage of text it is first added to the *compound_elements_db* database and then to the *compound_tree* database following processing. From this process *fear an phoist* is tagged as a noun phrase (FA)

fear an phoist → fear_an_phoist/FA

Unless the teacher identifies a compound as indivisible, each element of the compound is also tagged. The third

process is the completion of the syntactic tagging phase which is achieved by part-of-speech filtering in conjunction with the grammatical rules that are contained in the lexical database. For each word in an utterance that is not tagged after the compound recognition phase a search is performed to ascertain all potential syntactic categories for a word. Such information is obtained from *An Foclóir Beag* and any additional word information that has been added by the teacher. On completion of this search the grammatical rules, that are stored in the form of rewrite rules, are applied in an attempt to resolve syntactic ambiguities. The internal representation is updated with the results of this process prior to presentation to the teacher. The teacher may be involved in a number of different ways at this stage that can be summarised as:

a) If the system is not able to resolve syntactic ambiguities the internal representation is presented to the teacher in order to do so.
b) On presentation of the internal representation the teacher may choose to make alterations to the output obtained from the system.
c) Addition of a syntactic category for a word to the lexical database.

For example, in the second sentence (of the sample text) ***ar*** occurs twice, once as a preposition (***ar an mbóthar*** - on the road) and once as a verb (***ar sé*** - says/ said he). From the entry in the lexical database for ***ar*** (preposition) we get the following extract (McElligott and Ó Néill 1994),

```
ar
    + article
        singular, masculine, nominative
            +noun
                singular, masculine, nominative
                first letter can be eclipsed
                    action:    eclipse
```


So *ar an mbóthar* is consistent with ***ar*** being a preposition, whilst the other categories of ***ar*** do not give a match. For example, ***ar*** (as a verb says / said) requires a pronoun, which would not match the eclipse.

The teacher can augment the knowledge base (or even modify it), such as, by entering a new compound or by giving additional information for an existing word. A text may be identified as the first text in a series. For the first text in a series, the teacher may split the words of the text into three sets

a) words which the student is expected to know
b) forms of words new to the student
c) words new to the student

The system provides a list of the words in the root form together with the various forms of that word used.

Headword	Forms
ag	ag, acu
ar	ar, oraibh
ar	ar, arsa
bheith	bhí, bhíodar, bhfuil, tá
dá	dhá
moncaí	mhoncaí, moncaí
siúil	Siúl

Figure 4: Word root and forms appearing in the text

A list of words is shown, as in Figure 4, and the teacher clicks on the headword if the word is new or on the form if the form is new. By scrolling through the list all words of the text are accessible. After selection the teacher may review each list. For subsequent texts in a series, the system supplies the lists, which the teacher can then modify.

The final step of the teacher's preparation is the specification of exercises. The system can generate lexical and grammatical related exercises such as case, tense, number, gender, lenition (McElligott and Ó Néill 1995). The teacher can indicate for which words or grammatical points he or she wants exercises, for example, to convert from one tense to another for specified verbs (either from the text or the dictionary). General questions of comprehension, multiple choice questions and cloze tests can also be specified by the teacher. Apart from the cloze tests, the answers to these exercises must be supplied by the teacher.

The cloze tests are generated according to the teacher's requests, for example, he or she can specify that one noun from each sentence be omitted. The system then randomly selects the noun from each sentence. The corrections are then automatic. Synonyms may be supplied by the teacher for a word and should the student use a synonym then that would be accepted as correct input.

Once a text is prepared by the teacher, it is made available to the students. The dictionary is available together with the teacher's notes. The student can then do the exercises relating to the text. The system logs the words looked up by the student and generates a corresponding "vocabulary", which can be modified by the student. The student may do additional lexical or grammatical exercises if he or she chooses (since such exercises can be generated by the system). If the text being studied is part of a series, the student can search for words in the other texts and see their use in context. Such an environment, with an on-line dictionary, facilitates the student's vocabulary acquisition (Ellis 1995).

In conclusion, the system should be of use to both types of users. The system should help reduce the time and effort involved in preparing individual texts for class and be particulary useful if there is a series of linked texts. For the student it provides the benefit of easy access to the dictionary and the teacher's notes while working on the text.

REFERENCES

Ellis, N.C., 1995: "The Psychology of Foreign Language Vocabulary Acquisition: Implications for CALL." *Computer Assisted Language Learning* Vol. 8, No. 2-3.

Mac Giolla Phádraig, B.,1953: *Bun-Chúrsa ar Cheapadóireacht Gaeilge*, Cuid a hAon, Baile Átha Cliath: Brún agus Ó Nualláin.

McElligott, A. and G. Ó Néill, 1995: "CALL with Methodical Explanations," *Journal of Celtic Language Learning* 1, pp. 38-52.

McElligott, A. and G. Ó Néill, 1996: "Morphological Objects," *Journal of Celtic Language Learning* 2, pp. 21-32.

Roinn Oideachais, An, 1991: *An Foclóir Beag*. Dublin: An Gúm.

Sutcliffe, R.F.E., D. O'Sullivan, L. Relihan, L. Sheahan, A. McElligott, 1996: *Sift deliverable D27a: Collate Terminology for Sift*. Luimneach: Ollscoil Luimnigh. (Note: Sift - selecting information from text: LRE European funded project.)

Luimneach
Éire
annette.mcelligott@ul.ie
gearoid.oneill@ul.ie