Developing Integrated Editions of Minority Language Dictionaries: The Irish Example

Julianne Nyhan
Corpus of Electronic Texts, University College Cork, Ireland

Abstract

The *Corpus of Electronic Texts* (CELT) project at University College Cork is an on-line corpus of multilingual texts that are encoded in TEI conformant SGML/XML. As of September 2006, the corpus has 9.3 million words online. Over the last five years, doctoral work carried out at the project has focused on the development of lexicographical resources spanning the years c. AD 700–1700, and on the development of tools to integrate the corpus with these resources. This research has been further complimented by the *Linking Dictionaries and Text* project, a North–South Ireland collaboration between the University of Ulster, Coleraine, and University College Cork. The *Linking Dictionaries and Text* project will reach completion in October 2006. This article focuses on CELT's latest research project, the *Digital Dinneen* project, that aims to create an integrated edition of Patrick S. Dinneen's *Foclóir Gaedhilge agus Béarla* (Irish-English Dictionary). In this article, the newly developed research infrastructure—that is the culmination of the doctoral research carried out at CELT and the *Linking Dictionaries and Text* collaboration—will be described, and ways that the *Digital Dinneen* will be integrated into this infrastructure established. Finally, avenues of future research will be pointed to.

Correspondence:
Julianne Nyhan Knockrea Mews, CELT corpus, 2 Carrigside, University College Cork, Ireland.
E-mail:
julianne.nyhan@ucc.ie

1 Introduction

Patrick S. Dinneen published a variety of Irish language materials during his career including plays, essays, collection of poetry, biographies, and scholarly editions of texts. However, it is his lexicographical work that he is most celebrated for, and his first Irish-English dictionary was published in 1904. When the plates of that dictionary were—luckily for us as one scholar has indicated (Ua Súilleabháin, 2005, p. 71)—destroyed during the 1916 Easter rising, Dinneen began work on a new and enlarged edition that culminated in the publication of *Foclóir Gaedhilge agus Béarla* (Irish-English Dictionary) in 1927.

Dinneen's *Foclóir Gaedhilge agus Béarla* (hereafter *FGB*) is one of the richest works available for modern Irish, and is also useful for some aspects of medieval Irish. Though editors of later dictionaries could call on better resources than those at Dinneen's disposal, *FGB* presents a 'fuller coverage of the modern Irish lexicon' (Ó Murchú, 2005, p. 100) than Ó Dónaill's dictionary, which succeeded Dinneen in 1977. Further, in some instances, Dinneen presents a more complete treatment of words than is to be found in Ó Dónaill (Ó Murchú, 2005).

In addition to its acknowledged scholarly value, *FGB* is held in esteem by many for reasons as diverse as the sense of Dinneen's passion for the

Literary and Linguistic Computing, Vol. 23, No. 1, 2008.
doi:10.1093/llc/fqm038 Advance Access Published on 1 February 2008

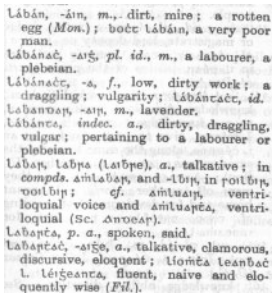
Lábán, -áin, *m.*, dirt, mire; a rotten egg (*Mon.*); bocc lábáin, a very poor man.
Lábánach, -aigh, *pl. id.*, *m.*, a labourer, a plebeian.
Lábánachd, -a, *f.*, low, dirty work; a dragging; vulgarity; lábánchaċċ, *id.*
Labanvan, -ain, *m.*, lavender.
Lábánta, *indec. a.*, dirty, draggling, vulgar; pertaining to a labourer or plebeian.
Labar, labra (laibre), *a.*, talkative; in *compds.* amlabar, and -lbir, in roilbir, roilbir; *cf.* amluain, ventriloquial voice and amluaidhe, ventriloquial (Sc. Anteap).
Labaidhe, *p. a.*, spoken, said.
Labaraċ, -aighe, *a.*, talkative, clamorous, discursive, eloquent; liomhċa leanbaċ i. léigeanta, fluent, naive and eloquently wise (*Fil.*).

Fig. 1 Example of Cló Gaelach font (Example 1)

Irish language that exudes from its pages; the fact that it preserves important citations from the Irish song tradition (Ó Madagáin, 2005); the dictionary's sometimes unintentionally humorous citations or definitions, for example, *caithim gamhain* 'I give birth to a still-born calf' (Ó Murchú, 2005, p. 82); as well as the occasional digressions it contains, described by O'Connell as seducing 'the reader into many fascinating byways' (O'Connell, 1984). In short, though almost a 100-years old, Dinneen's dictionary remains an invaluable resource for scholars, writers, teachers, linguists, historians, and all others who speak and learn the Irish language.

Despite the popularity of Dinneen's dictionary, many non-specialists find that aspects of its presentation and orthography form a barrier to efficient use of it. The font of the dictionary is *Cló Gaelach* (see *Example 1* in Fig. 1), a font that many modern Irish speakers are unfamiliar with, and find difficult to read. The orthography of the dictionary pre-dates the spelling reform of 1946, and because many present day speakers of modern Irish are unfamiliar with such orthography, locating the required lexical entry can be somewhat of a hit and miss process. It was recognized that a digital edition of the dictionary would enable new solutions to these problems to be developed, and that the potential value of a fully searchable edition of *FGB* to the research communities of Irish studies and Humanities Computing is high. Therefore, the Corpus of Electronic Texts (hereafter CELT) applied to the *Irish Research Council for the Humanities and Social Sciences* for funding to design an integrated electronic edition of Dinneen's dictionary and was successful. In order to explain the concept of an integrated edition, and the type of integration that is envisaged for the digital edition of *FGB*, the CELT project, the electronic *Dictionary of the Irish language* project, the electronic *Lexicon* and the *Linking Dictionaries and Text* collaboration will now be described.

2 Role of Dinneen in Irish Language Humanities Computing

2.1 Corpus of electronic texts

The CELT project at the History Department, University College Cork, is a scholarly corpus of multi-lingual texts of Irish literature and history. The texts in the CELT corpus are written in Irish of all periods, Latin, English, French, and both English and English translations of original texts are represented. It is the largest Irish text corpus of its kind available freely to members of the public. CELT is conformant with the TEI guidelines, encoded in SGML and will soon switch to XML as the meta-markup language of choice for its master files. As of September 2006, the corpus has 9.3 million words available online. My doctoral research, carried out at CELT, examined the application of XML to the historical lexicography of Old, Middle, and Early modern Irish (Nyhan, 2006b). Part of that work focused on the creation of an electronic *Lexicon* of Irish covering the period c.700–1700.

2.2 Electronic *Lexicon* of old Irish

The *Lexicon* forms a re-edited subset of the Royal Irish Academy's *Dictionary of the Irish language* (Marstrander *et al.*, Compact edition, 1998) and is supplemented with other sources and the findings of recent scholarship. Six letters of the *Lexicon* have been completed to date. The *Dictionary of the Irish language* (hereafter *DIL*) is a scholarly historical dictionary of Old, Middle, and Early modern Irish, and though it is widely accepted as the most authoritative reference work of its kind available for the Irish language, it exhibits serious limitations in the structure and presentation of its data.[1]

These limitations are compounded by aspects of the Irish language, which experienced complex sound changes at various stages of its development, and the language therefore displays a complicated orthography (Beekes, 1995; Thurneysen, 1998). Words beginning with *air-*, for example, can appear as *aur-*, *er-*, *ir-*, and *ur-*. Thus, the word *airthach*, is also attested as *aurthach*, *urthach*, and *erthach*. A student who encounters the forms *urthach* and *erthach* in a text is unlikely to be aware of such spelling variants—especially in light of the fact that a table explaining such sound changes is not provided in *DIL*—and so is wholly dependent upon cross-references to find the word in question. In the case of the word *airthach*, however, no cross-reference is provided to the forms *erthach*, *aurthach*, or *urthach*. The only form that is cross-referenced is *urrach*. Thus, students and even scholars of medieval Irish frequently spend a large amount of time locating a required word, and many fail to find the object of their search altogether.

In order to resolve these limitations, among others, the electronic *Lexicon* was designed. The *Lexicon* contains the essential data, encountered by the scholar in the first instance, deeply encoded in XML. The many inconsistencies in *DIL* including, *inter alia*, its inconsistent use of abbreviations and citations, proliferation of lexical entries that lack a repeatable structure, and lack of cross references to variant forms have been remedied in the *Lexicon*, which offers advanced search and interrogation possibilities. The excerpt below (*Example 2*) from the possessive pronoun *mo* indicates the treatment of material in it. The *Lexicon* will be published incrementally on http://www.epu.ucc.ie/lexicon during 2007.

```
<form.gr type="simple">
  <form type="simple" person="1" number="sg">M'<ps>Aedóc
</ps></form>
  <form type="simple" person="1" number="sg">M'<ps>Aedocc
</ps></form>
  <form type="simple" person="1" number="sg">M<ps>Aodhocc
</ps></form>
  <form type="simple" person="1" number="sg">Mo<ps>edocc
</ps></form>
  <form type="simple" person="1" number="sg">Mo<ps>eóg
</ps></form>
  <form type="simple" person="1" number="sg">Mo<ps>bí
</ps></form>
  <form type="simple" person="1" number="sg">Mo<ps>chaemóc
</ps></form>
  <form type="simple" person="1" number="sg">M'<ps>Ernóc
</ps></form>
  <form type="simple" person="1" number="sg">Mo<ps>fheca
</ps></form>
  <form type="simple" person="1" number="sg">Mo<ps>nanna
</ps></form>
  <form type="simple" person="1" number="sg">M<ps>unnu
</ps> <note ref="general">abbreviated form of
<form type="simple" person="1" number="sg">Mo<ps>Fhinnu
</ps></form></note></form>
</form.gr>
```

Example 2 Encoding of hypocoristic forms of the possessive pronoun *mo*

2.3 The *eDIL* project and the *LDT*

The Centre for Irish and Celtic Studies at the University of Ulster, with funding from the *Arts and Humanities Research Board*, is currently preparing a digital edition of the entire *DIL* (Fomin and Toner, 2006). The *eDIL* is due for completion in 2007. The CELT project at University College Cork and the *eDIL* team at the University of Ulster have been collaborating since 2003 in the *Higher Education Authority* funded *Linking Dictionaries and Texts* (hereafter LDT) project. Upon commencement of that project, many of the texts cited in both the *Lexicon* and *eDIL* were available in the best research libraries only, and scholars frequently invested considerable time and money travelling to research libraries to access these texts. The LDT project, which also reaches completion in 2007, identified the most commonly cited texts in *DIL*, prepared TEI-encoded electronic editions of them and generated links from citations in *eDIL* to the corresponding electronic text editions in the CELT corpus.

2.4 The *Digital Dinneen* project

When completed in 2009, the *Digital Dinneen* will extend CELT's lexicographical material from the Old Irish period (c.700–900 AD) up to modern literary Irish. The aim of the *Digital Dinneen* project is not only to develop an electronic edition of *FGB*, but to publish it as an integrated edition that contains both fixed links to the CELT project, the *Lexicon*, and *eDIL*, and also arbitrary,

user-generated links from the corpus to the *Digital Dinneen* and vice versa. Creating such an edition can be described as having two broad stages of research: the first is the creation of a stable, interoperable, and extensible electronic edition that meets the highest scholarly standards. The second is the further development and transformation of the markup contained in that edition, as well as its supplementation with further technologies, such as Javascript, to support the development of the kind of integration with centralized and de-centralized text collections that is envisaged. These two stages will now be used as a framework to discuss the development of the edition further.

3 Methodology and Scope of the Dinneen Project

3.1 Desiderata of the digital edition

Important aims of the digital edition include resolving the problems of orthography and font that many present day Irish speakers encounter in *FGB*. It is intended to offer users of the electronic edition the choice of rendering the data with one of two stylesheets: one will present the data in Roman font and the other in *Cló Gaelach* font. Using De Bhaldraithe's *Innéacs Nua-Ghaeilge don Dictionary of the Irish language* (1981), as well as additional information supplied by Dr Seán Ua Súilleabháin, Irish department, University College Cork, pointers will be generated from headwords in the *Digital Dinneen* (to the unique ids on headwords in *eDIL* and the *Lexicon*) to allow a user to follow a modern Irish form in Dinneen's dictionary back to its earlier forms in *eDIL* and the *Lexicon*. Further, the feasibility of incorporating the post-spelling reform orthography of headwords, as meta-text of the headwords in the *Digital Dinneen*, is currently being research to see if such an undertaking will be possible in the three year time span of the project. In regard to the markup of the data contained in the dictionary, all information is being encoded including but not limited to: forms, senses, etymologies, parts of speech, registers, sources, grammatical information, personal names, and places names. Markup of the information contained in the dictionary is drawing to completion, and it has already been possible to use the electronic edition to support a range of applied research. For example, a word list supplied to Dinneen by the poet Riobard Bheldon, that has since disappeared, was reassembled based on information extracted from the dictionary (Nyhan, 2006a).

3.2 Data capture

The data capture was outsourced to Godrej Global Solutions Ltd, who returned the data with basic structural and presentational markup, and an accuracy level in excess of 99.995%.

3.3 XML markup

During work on the *Lexicon*, the data were encoded in middle of the road XML, defined in a transitional DTD, and TEI conformance was also achieved at a later stage through XSLT post-processing. This approach is also being applied to the encoding of the *Digital Dinneen*. The main benefit of such an approach is that when working to understand, interpret, and describe complex data, complete focus can be given to designing markup that accurately describes that data. Considerations about how the data model in question relates to the TEI guidelines can then be dealt with at a later stage in the project, when the understanding of the type and structure of data in the dictionary is better developed.

3.4 XSLT transformations

In marked contrast with the Royal Irish Academy's *Dictionary of the Irish language*, Dinneen's dictionary can be described as being reasonably consistent in its structure, layout, use of abbreviations, bibliographical citations, and presentation of data. Earlier research into using XSLT to automate XML tagging of Irish lexicographical material highlighted that consistency of structure in the base text is an important factor for raising the accuracy rate of transformations. So, with this strong foundation in place, a series of XSLT transformations were developed based on a combination of matching patterns in the text and also upon testing the

```
<entry><head><d>3 A,</d></head> <emph type="italic">interj., aspir.</emph>, precedes voc.
case, <emph type="italic">sing.</emph> and <emph type="italic">pl.</emph>, often slurred over
in pronunciation; silent before vowel or <d>fh</d> and somet. not written, being replaced by an
apostrophe or simply omitted; <emph type="italic">smt.</emph> Eng. O, though not an equivalent,
represents it; <d>a fhir an fháinne,</d> O man of the ring; <d>a Phádraig,</d> O Patrick;
<d>éistidh, a dhaoine,</d> listen, O people; <d>a mhic,</d> my son; the usage, <d>a mh'
anam,</d> my soul! is exceptional.</entry>
```

Example 3 Capture text

content of elements and proceeding in specified ways once a match was found.

As indicated earlier, the captured text was returned with basic XML structural and presentational markup that preserved the formatting of the original hard-copy edition. As illustrated in *Example 3*, this markup included an <entry> element that contains and delineates each lexical entry; a <head> element encoding each headword; an <emph> element to encode text in italics; and a <d> element encoding pieces of Irish language text.

The file was processed using a Sed script, and the linux utilities sort and uniq, to get an alphabetical list of the content of <emph> elements, accompanied by a number indicating the number of times each text string re-occurred.

Interestingly, some of the problems of consistency, from the perspective of XSLT transformations, were introduced into the text at the data capture stage, and were highlighted by this routine. When the dictionary was being captured, and presentational (rather than descriptive) markup was being applied, strings of grammatical information, for example, were rightly encoded (based on the type of markup that was being applied) in one <emph> element that spanned the length of the italicized grammatical information. However, this presentational markup is not retained in the electronic edition, but rather transformed into descriptive markup. To increase the success rate of the XSLT generation of elements, necessary aspects of pre-processing such as this were identified, and, for example, a tokenising template was applied to all <emph> elements to separate its contents based on punctuation.

After the pre-processing stage was completed, a series of XSLT transformation were written, based on the results that had been obtained from the

```
<xsl:template match="text()
[name(preceding-sibling::*[1])='cit']
[name(following-sibling::*[1])='cit']">
<def>
<xsl:value-of select="."/>
</def>
</xsl:template>
```

Example 4 Automating definition elements

analysis of the content of <emph>. If the content of the element matched a specified string in the XSLT file, the element was transformed accordingly, otherwise <emph> was encoded as bibliographical information. This proved a very convenient way to assess the accuracy of the scripts across the entire dictionary, because at the test stage, it was possible to extract the content of all bibliographical elements, thus easily identifying information that was being encoded incorrectly. All the markup of grammatical information, place name information, and bibliographical citations was accurately automated in this routine. The next XSLT transformation was based on observable patterns in the text, for example, if the string preceding <emph> was *Lat.* or *Fr.*, then <emph> was transformed to <frn lang="la"> or whatever.

In the case of <d> elements (marking Irish language text), it was observed that inflected forms are usually listed after the head word and that the root of the word is usually indicated by a dash. Therefore, the contents of <d> elements were tested for an initial dash, if it were present, <d> was transformed to a <form>, otherwise it was transformed to a citation.

Then, building on the results of these transformations, another set of scripts was written

```
<entry><head><d>Annáil,</d></head> <d>-ála,</d> <emph type="italic">pl. id.</emph>,
<d>-álaigh,</d> and <d>-álacha,</d> <emph type="italic">f.</emph>, year, age,
date; <d>míle go leith a. Íosa,</d> fifteen hundred the age of Jesus (the <emph
type="italic">Annus Domini</emph>); <emph type="italic">genly.</emph> in
<emph type="italic">pl.</emph>, annals.</entry>
```

Example 5 Pre-XSLT transformation

```
<entry id="d1e31556"><form><orth>Annáil</orth></form>, <form>-ála,</form>
<gramGrp><number>pl.</number> <lbl>id.</lbl></gramGrp>, <form>-álaigh</form>, and
<form>-álacha</form>, <gramGrp><gen>f.</gen></gramGrp>, <sense><def>year, age,
date</def></sense>; <eg><q>míle go leith a. Íosa</q>, <trans><tr>fifteen hundred the age of
Jesus</tr></trans> <note>(the <frn lang="la">Annus Domini</frn>)</note></eg>; <note>genly. in
<gramGrp><number>pl.</number></gramGrp>, <sense><def>annals</def></sense>.</note>
</entry>
```

Example 6 Post-XSLT transformation

to generate markup of further portions of the dictionary, such as definition text.

The automation of tagging with XSLT proved to be very successful. As already indicated, all the markup of grammatical information, place names, bibliographical citations, foreign languages, citations, translations, and senses was successfully generated. The automation of forms of the headword was not as successful because it was based only on the criterion that the form begins with a dash, but this transformation will be further refined.

3.5 TEI modifications

As has already been stated, conformance with the TEI guidelines will be achieve through XSLT post-processing, but it is nevertheless clear, at this stage of research, that some modifications and extensions of the *TEI Set for Print Dictionaries* (Sperberg-McQueen and Bernard, 2002) will be required.

It is frequently observed by meta-lexicographers when discussing the development of dictionaries from a historical perspective, that many dictionaries cannibalize the ones that went before them. So too, it seems, is the case with markup of dictionaries. The *Digital Dinneen* project will incorporate some TEI modifications proposed by the *eDIL* team, including the element <egForm>, devised to allow the grouping of 'sections of related grammatical information associated with any morphological

```
<form><mut type="l">mo</mut></form>
```

Example 7 Mut element

item' (Fomin and Toner, 2006, p. 87) and the tag <modifier> to markup 'different linguistic items that modify the meaning of the headword, such as prepositions, prefixes, and particles' (Fomin and Toner, 2006, p. 86). Modifications of the TEI made by the electronic *Scottish National Dictionary* project will also be availed of, including the attributes made available on the <cit> element and on the <geo> element to allow identification of oral and print sources (Rennie, 2001, pp. 156–57).

The *Digital Dinneen* project will add a new element to the TEI DTD called <mut>, to allow initial mutations that occur in Irish such as nasalization, gemination, and lenition to be encoded. For example, the possessive pronoun *mo* can lenite some nouns noun that follow it, and the value of the attribute type available on <mut> will be permitted to be one of n/g/l to specify the type of mutation in question (See Example 7).

Further, Dinneen frequently gives detailed information about how a headword combines with other words in a sentence. The merits of creating a new element called <syntax> to encode this information are being researched. While it would be possible to encode this information in a generic note modified

<syntax>*An* is often used in close combination with prep[ositions], especially those ending with a vowel, as *do 'n* or *don*, *i san*, *'san* or *san*, *i sna* or *'s na*, *ó'n* or *ón*, *gus an*, *leis an*, *do na* or *dos na*, in which the old initial s of the art[icle] is preserved </syntax>

Example 8 Syntax elemant

with an attribute, working on the basis of core context and the expectation that researchers will want to manipulate this information further, it seems likely that this element will also be added to the TEI DTD at post-processing stage. A simplified example is given in *Example 8*.

4 Integration and Role of *Digital Dinneen*

The type of integration envisaged for the *Digital Dinneen* includes both links that are hard coded into the edition itself, and also the extension of mechanisms that are being developed to support arbitrary, user generated data queries.

4.1 Extending *LDT* to *Dinneen*

As indicated earlier, the *LDT* project was an Irish cross-border collaboration between University College Cork and the University of Ulster, Coleraine. It was based on the observation that the effectiveness of CELT, the *Lexicon* and *eDIL* as research tools could be maximised through cooperation. It was envisaged that this cooperation would take the form of conventional predetermined (fixed) links, either encoded in a static HTML file or using HTML generated from an SGML or XML source. This would enable users to click on a bibliographical citation (in *eDIL* or the *Lexicon*) that would resolve to the specified text in the CELT website.

The technology that facilitates this is commonplace and well established on the Internet and involves the citational information being appended to the URI in the conventional manner, as illustrated in *Example 9*.

After the citational information has been appended to the URI, a server side application uses a table of the known bibliographical data (based on the document abbreviation and the

<cit><bibl>ZCP</bibl> <biblscope>xiii 169.36 </biblscope></cit>,
the link would be in the format
<xptr to="http://celt.ucc.ie/cgi-bin/lookup? cite=ZCP\%40xiii\%40169.36"/>

Example 9 Fixed linking

format of the bibliographical numeration) to determine whether the required document is on-line or not. If it is, the server will return the document fragment surrounding the reference; if not, it will return an explanation. In this way, as documents are added to the CELT text-base over the coming years, further references from *eDIL* and the *Lexicon* will be satisfied, because the citation data required for recognition can be stored in the TEI header of the CELT documents themselves, and extracted for use by the server on an automated basis.

While the technology that facilitates the type of fixed linking described here is easy to implement, devising a system to automate the XML encoding of bibliographical references proved much more complex. This was achieved with a 90% accuracy level with regular expressions applied in a Sed script.

The mechanism described here will also be extended to the *Digital Dinneen* and will be a boon for scholars of modern Irish. Because many of the texts available at CELT date from the Old and Middle Irish periods, the corpus is currently being supplement with further editions of modern Irish texts.

4.2 Linking corpus to *Lexicon* and the *Digital Dinneen*

During research on the Lexicon, a mechanism to enable arbitrary, user generated links between centralized and de-centralized text corpora and XML encoded lexicographical resources was developed. In order to ensure that the impact of the

technicalities of such a system can be kept to a minimum for the end user, and to enable ease of use, it was decided that it should be possible to highlight any word or phrase in the CELT corpus and retrieve it in the *Lexicon*. The nature of Internet development over the past decade encourages this approach, although there are misgivings in related fields such as Library Science about the levels of software and skills needed for anything more than occasional use of real-time referencing (Janes, 2002).

A robust mechanism for arbitrary reference that could be adapted to enable textual connections between CELT and the *Lexicon*, or indeed vice versa, was kindly brought to my attention by a colleague.[2] The piece of Javascript illustrated in *Example 10* is the work of authors or contributors of blogging systems like Wordpress[3] and others, where it was designed for the capture of URIs.

Suitably adapted, the code must be entered all on one line (it is broken over several lines in *Example 10* for convenience) and installed in browsers that support bookmarklets. The code initiates a DOM function getSelection() that captures the highlighted area to a variable Q, and a window is then opened to a suitable URI, with the escaped value of Q appended as a value to the field text. Once installed, the user can highlight a word in the CELT corpus in the browser, and click on the new bookmarklet. The text gets sent to the URI specified, which in this implementation is the development lookup function for the *Lexicon*. Finally, the XSLT stylesheet that has been developed generates the lookup display.

An example of a user having highlighted a word in the saga *Fingal Ronáin* and clicked on the lookup button is illustrated in Fig. 2 (Example 11).

This mechanism will also be extended to the *Digital Dinneen* to allow users to highlight words in modern Irish texts, click on a toolbar button, and retrieve the required entry from the *Digital Dinneen*. In addition to enabling students, scholars, and lay readers to use existing resources such as CELT, the *Lexicon*, *eDIL*, and *Digital Dinneen* more effectively, this method of arbitrary textual linking enables existing scholarly resources to be examined from different perspectives. For example, words highlighted in the CELT corpus that are not retrieved in the *Lexicon*, *eDIL*, or *Digtial Dinneen* will most likely represent words, that for one reason or another, were not included in these dictionaries, whether as head-words or variants, thus providing information for future supplements to all of the lexicographical materials described here.

5 Future Plans

It is hoped that further funding will be secured to allow research on the *Lexicon* and on the research infrastructure described here to continue. While many research questions have been answered, some still remain. For example, the question of how to refine the arbitrary linking mechanism between Corpus and lexicographical resource so that the correct homograph is returned still remains unanswered. It is hoped that research into text mining may provide some solutions. It is also hoped that funding will be secured to develop electronic teaching and learning applications for use in conjunction with the resources described here, and, indeed, in conjunction with other similar minority language humanities computing projects.

```
javascript: Q=document.selection?document.selection. createRange().text:document.getSelection();
void(window.open(' http://www.ucc.ie: 8080/cocoon/lexicon/find? text=' +escape(Q),
`textselectionbookmarklet' , ` scrollbars=yes, width=480, height=300, left=100, top=150,
status=yes' ));
```

Example 10 Javascript bookmarklet implementation

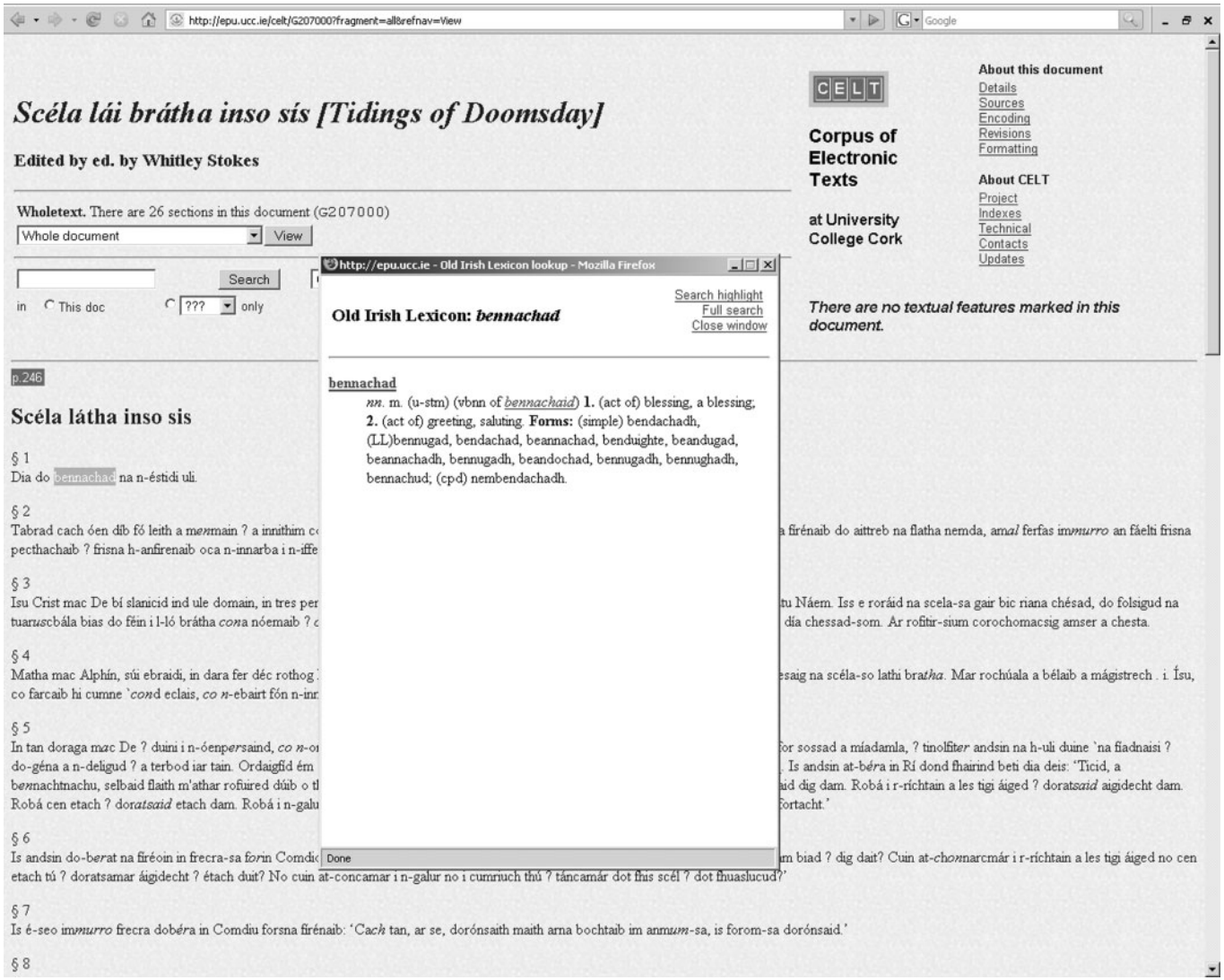

Fig. 2 The lookup bookmarklet in operation (Example 11)

References

Beekes, R. S. P. (1995). Vertalers, Uva and Gabriner, Paul trans. *Comparative Indo-European linguistics*. Amsterdam/Philadelphia: John Benjamins.

De Bhaldraithe, T. (1981). *Innéacs Nua-Ghaeilge don 'Dictionary of the Irish language'*. Dublin: Royal Irish Academy.

Fomin, M. and Toner, G. (2006). Digitizing a dictionary of medieval Irish: the *eDIL* Project. *Literary and Linguistic Computing*, **21**(1): 83–90.

Janes, J. (2002). Live reference: too much, too fast? *Library Journal* October (2002). http://www.library-journal.com/index.asp?layout=article\&articleid\newline=CA251681. (accessed on 2 June 2006).

Marstrander, C. J. S. *et al.* (1998, Compact edition). *Dictionary of the Irish language based mainly on Old and Middle Irish materials*. Dublin: Royal Irish Academy.

Nyhan, J. (2006a). Findfhocla an Chomaraigh: Cnuasach Riobaird Bheldon. In Nic Dhonnchadha and Aoibheann (eds), *An Linn Bhuí Iris Ghaeltacht na nDéise*, **10**: 97–111.

Nyhan, J. (2006b). The Application of XML to the Historical Lexicography of Old, Middle and Early Modern Irish: a Lexicon-based Analysis. PhD thesis. University College Cork, pp. 79–90. http://epu.ucc.ie/lexicon/JNyhan.html.

O'Connell, N. (1984). *Father Dinneen – his Dictionary and the Gaelic Revival*. Dublin: Irish Texts Society.

Ó Dónaill, N. (1977). *Foclóir Gaeilge-Béarla*. Dublin: An Gúm.

Ó Madagáin, B. (2005). Pádraig Ó Duinnín: Key witness to the Irish Song Tradition. In and (eds), *Dinneen and The Dictionary 1904–2004*. Dublin: Irish Texts Society, pp. 51–61.

Ó Murchú, M. (2005). Dinneen and Ó Dónaill. In Riggs, Pádraigín (ed), *Dinneen and The Dictionary 1904–2004*. Dublin: Irish Texts Society, pp. 78–101.

Rennie, S. (2001). The *Electronic Scottish National Dictionary* (eSND): Work in Progress. *Literary and Linguistic Computing*, **16**(1): pp. 153–90.

Sperberg-McQueen, C. M. and Burnard, L. (ed) (2002). *Guidelines for electronic text encoding and interchange (P4)*. TEI Consortium: Oxford.

Thurneysen, R. (1998). Binchy, D. A. and Bergin, Osborn trans. *A grammar of Old Irish*. Dublin: Dublin Institute for Advanced Studies.

Ua Súilleabháin, S. (2005). Dinneen's Dictionaries of 1904 and 1927: Background, use of Historical Dictionaries and of Contemporary Informants. In and (eds), *Dinneen and The Dictionary 1904–2004*. Dublin: Irish Texts Society, pp. 62–77.

Notes

1 For a discussion of these limitations, see Julianne Nyhan (2006b, pp. 79–90).

2 I am especially grateful to Peter Flynn, manager of the Electronic Publishing Unit, UCC, for bringing this mechanism to my attention.

3 See: http://wordpress.org.