Using Irish Language Corpora in the University Classroom

Karin Hansson

1. Introduction

This paper concerns the use of electronic corpora in the Irish language courses taught at the Celtic Section, Uppsala University. Corpora are increasingly being recognised as an important tool in language teaching, in particular at university level. For example, electronic corpora can make up for a lack of suitable textbooks and grammars, a problem facing both teachers and students in Uppsala. I will present my experiences of working with the *PAROLE Irish Distributable Corpus* (*Corpas Náisiúnta na Gaeilge*) to find instances of use that can help explain grammar rules and to create grammar exercises and assignments, as well as to provide material for student term paper projects. In my experience, an electronic corpus is an important tool for the teacher, providing an easily accessible source of authentic examples that reflect actual use and this is of particular benefit in teaching a foreign language. However, there are also problems and caveats involved when working with an electronic corpus regarding the set-up, compilation method, documentation and necessary search and concordance programmes, which I will also address.

Electronic corpora have been used in linguistics since roughly the 1960s but it is only recently that they have started to be used in language teaching.[1] Gradually, teachers have realised that reliable information about various aspects of language use cannot be provided by textbooks or introspection only. Instead, more and more teachers now turn to electronic corpora to retrieve relevant material from a large variety of authentic sources. Electronic corpora are widely available, affordable, and easy to use. Today, many textbooks, grammars and other teaching materials are based on corpus-derived data. However, whereas this development is well underway in the teaching of some languages, especially English, the situation for Irish, being a lesser used language, is vastly different. The emergence of electronic corpora of contemporary Irish is very recent. As a consequence, there are still very few resources available to teachers of Irish who want to use corpus linguistic methods in class and teaching material based on corpus-derived data.

1 For an historic overview, as well as examples of corpus-aided foreign lanugage teaching projects, see Sinclair (2004).

Karin Hansson, karinoh@gmail.com
Celtic Section, Department of English, Uppsala University
Studia Celtica Fennica IV (2007) 58-62, ISSN 1795–097X,

2. Irish language corpora

There are (at least) two major electronic corpora of 20th-century Irish texts. The larger of the two is the *PAROLE Irish distributable Corpus* (also referred to as *Corpus Náisiúnta na Gaeilge* / *The National Corpus of Irish*). This corpus contains over 8 million words from sources published between 1970 and 1990. The source texts are mainly books, both fiction and non-fiction, and newspapers, covering a wide range of topics and genres. *The National Corpus of Irish* was compiled and published as part of the EU-financed *PAROLE* project. The aim of this project was to offer a large-scale harmonised set of corpora for all EU languages.

The other corpus is *Tobar na Gaedhilge* ('The source of Irish'), a 2.5 million word corpus containing 45 Irish language texts (42 of which are written in the Ulster dialect, one of the three main dialects of Modern Irish), mainly fiction (novels and short stories) published between 1907 and 1967.[2]

The National Corpus of Irish is available on CDROM, priced at €50 for non-commercial use. The *Tobar na Gaedhilge* can be downloaded for free.

Table 1. Summary of the PAROLE and Tobar na Gaedhilge corpora.

Corpus	**Corpus na Gaeilge**	**Tobar na Gaedhilge (version 1.2)**
approx. no. of words	> 8,000,000	2,500,000
no. of texts	304 (109 books)	45 books
Genres	mixed	mainly fiction
time span	1970-1990	1907-1967

2.1 Concordance programs

Tobar na Gaedhilge includes a multi-functional retrieval program; the corpus is intended to be used only with this program. *The National Corpus of Irish*, however, consists merely of a set of texts, which means that you also need a search and concordance program to be able to retrieve data from the files. The advantage of this is that the user is free to choose whatever search programme he or she prefers. There are several concordance and search programs available to buy (including *WordSmith Tools*, which is the program that I used) or download for free (for example, *TextSTAT*).[3]

2 This applies to version 1.2, released in 2004. Version1.3 was released in September 2006, containing more texts and improved search and concordance program. For further details, see http://www.smo.uhi.ac.uk/~oduibhin/tobar/index.htm

3 For more information about *TextSTAT*, see http://www.niederlandistik.fu-berlin.de/textstat/software-en.html

3. Corpora in the classroom

In this section I will describe how I have used the *PAROLE* corpus and corpus linguistics in my own work as a teacher. I have used the corpus with my second and third semester students attending courses in Modern Irish at the Celtic Section, Uppsala University. I have not used the corpus with first semester students because at that point the students' vocabulary and grammar are very basic and the material found in textbooks is sufficient and thus there is no real need for additional corpus-derived data.

With second and third semester students I have used the corpus in two ways. First, I have introduced the students to corpus linguistics by presenting the *PAROLE* corpus and *WordSmith Tools* to them. After acquainting them with the basics of the corpus set-up and how to search it I gave them questions to answer by retrieving data from the corpus, for example regarding the use of synonymous expressions (such as *gasúr* and *páiste*, both meaning 'child') or alternating verb forms (such as *tá mé* and *táim*, both meaning 'I am'). For example, the students would be asked to compare the frequencies of the phrases under investigation. This is at first glance a very simple task but for the students it involves taking into account some important points of grammars since they have to consider the mutated and / or declined forms of the nouns in order to make their searches complete: *gasúr, ghasúr, ngasúr, ghasúir, gasúir* etc. When it comes to verb phrases, students need to consider the contrast between independent and dependent verb forms as well as initial mutations caused by verb particles when formulating search strings: *tá mé, níl mé, bhfuil mé* etc. The main aim of exercises of this kind was to encourage students to explore Irish grammar and usage independently, and to find language facts that would not be explained to them in textbooks, dictionaries and grammars.[4]

Second, I have used the corpus as an aid in collecting relevant examples when preparing lessons, handouts, exercises, and assignments. For example, I have prepared exercises where the task is to fill in plural and / or genitive forms of nouns and adjectives in a sentence based on the context and the translation of the sentence. The main purpose of this was to be able to explain and describe better the use of a particular construction in different types of text beyond the limited examples presented in grammars and textbooks.

3.1 Student essay projects

Apart from grammar exercises and independent student work in class, two third semester students in the Celtic Section have used corpus-based data for their essay projects (corresponding to half a semester's work). One of them is currently working on a project about the use and structure of subclauses introduced by *agus*, 'and', in

4 For example, both *tá mé* and *táim* are listed in *New Irish Grammar by the Christian Brothers* (1986, 112) but their use is not commented on.

Modern Irish texts. She used parts of the *PAROLE* corpus, focussing on fiction and newspaper texts, to collect around 200 instances of use to base her study on. The other student studied non-standard forms of the irregular verb *déan*, 'do, make', in his essay, using material from both the *PAROLE* and *Tobar na Gaedhilge* corpora. In total he found around 12,000 verb forms. As a result of the use of corpora, these students had a large set of contemporary empirical data from a wide variety of sources at their disposal that would otherwise have been virtually impossible to compile, especially considering the limited time available to them.

4. Advantages of using electronic corpora

In my experience, there are several advantages to using corpora in language teaching. Firstly they offer an excellent way to remedy the lack of suitable teaching material for university students of Irish, in particular those who have very little previous knowledge of the language. Apart from providing more examples and from a wider range of sources than can be found in text books and grammars, one of the greatest advantages of using corpora-derived data is that it helps the teacher to focus on the most frequently occurring aspects of grammar rather than lesser-used constructions. This is particularly important for me personally as a non-native speaker of Irish since in many cases I cannot rely on introspection alone. For example, the declension of adjectives together with nouns in the genitive is a complex matter. It is discouraging for students to study paradigms and examples like *hata an fhir bhig*, 'the small man's hat', as prescribed by grammarians (*New Irish Grammar by the Christian Brothers* 1986, 61), instead of ones like *hata an f[h]ear beag*, which is more common in genuine Connemara Irish. However, corpus-derived data reveals that adjectives in the genitive are relatively infrequent in actual use. The relevance of practising examples like the one cited above is thus limited and in my experience, it is of great comfort to students to realise this. Also, searching an electronic corpus for relevant data is considerably faster than searching printed sources manually.

4.1 Caveats

Nevertheless, despite all the advantages of using corpora in language teaching there are also caveats involved. The main problem with using the *PAROLE* corpus concerns the selection of material from which it is composed. Several types of texts included in the corpus are unsuitable as sources of examples of language use for teaching purposes because they are extracts from, for example, course books or translated works, or because some texts contain quotes, lyrics, poems, titles, names of products etc. It may be more difficult to detect unsuitable examples in the concordance list of a search program than in printed texts due to the limited context

displayed. Therefore, it is necessary to study examples carefully before using them and also to select files from the corpus manually before conducting a search.

Furthermore, the texts in the *PAROLE* corpus have not been marked up for dialect, which must be kept in mind when looking for examples of expressions or structures where there may be dialectal variation. This is of particular relevance for non-native speakers of Irish and learners who may lack the necessary linguistic knowledge to be able to assess the text sources with regard to dialect. Dialectal variation must be taken into account when formulating a search string as well as in the interpretation of the search results. Obviously, this also applies to *Tobar na Gaedhilge* which contains many dialectal forms.

5. Conclusion

To conclude, then, it is obvious that using corpora in the teaching of Irish as a foreign language has great potential, both as a source of material for exercises and descriptions of Irish usage in various contexts for the teacher and as a tool for, in particular, more advanced students for independent study. However, a lot of work in this field still needs to be done. For example, textbooks, dictionaries and grammars based on corpus-derived data would be a very welcome development in Irish language teaching. Further, spoken corpora would also be a very useful tool that could provide invaluable information for teachers and students alike.

In addition, the little research that has been done in the field of corpus-aided language teaching mostly concerns English and it is thus important to explore further the benefits of using corpora in the teaching of lesser-used languages like Irish. However, most importantly, corpora and corpus-based teaching material can help remedy the lack of exposure to Irish that students face due to the limited use of the Irish language today.

Bibliography

Primary material

PAROLE Irish Distributable Corpus. 2000 ITÉ (Linguistics Institute of Ireland). http://www.elda.org/

Ó Duibhín, C. 2004. *Tobar na Gaedhilge. Gaelic textbase and retrieval system for use under MS Windows*. http://www.smo.uhi.ac.uk/~oduibhin/tobar/index.htm

Secondary material

New Irish Grammar by the Christian Brothers. 1986. Dublin: C. J. Fallon.

Scott, Mike. 1999. *Wordsmith Tools version 4*. Oxford: Oxford University Press. http://www.liv.ac.uk/~ms2928/

Sinclair, J. McH. (ed.). 2004. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.