Efficient corpus development for lexicography: building the New Corpus for Ireland

Adam Kilgarriff, Michael Rundell
Lexicography MasterClass Ltd
UK

Elaine Uí Dhonnchadha
Institiúid Teangeolaíochta Éireann
Ireland

1. Introduction

In this paper we describe the development of the New Corpus for Ireland (NCI) – a substantial lexicographic corpus in two-parts, one being Irish (the Celtic language of Ireland), the other Hiberno-English (the variety of English that is spoken in Ireland). We describe its *design, collection,* and *encoding* and these are the main concerns of the paper.

The NCI was developed as part of the set-up phase of a project for a new English-to-Irish Dictionary (NEID).[1] The NEID is intended to be used by scholars, school and university students, translators, people working in the media, and the general public. It will replace the current main reference work, Tomás de Bhaldraithe's *English–Irish Dictionary* (1959), a highly-regarded dictionary but now almost 50 years old.

The island of Ireland includes both the Republic of Ireland and, in the North, six counties of the province of Ulster, which form part of the United Kingdom. The border was not critical to the project; collaborators and texts alike were sought both North and South of the border, and the language and dialects of Ulster were treated on a par with those of other regions. In this paper, "Ireland" means the whole island.

62,000 speakers use Irish as their main everyday language, and almost 340,000 speakers use Irish on a daily basis[2] It was the main language of Ireland until English displaced it (substantially as a result British imperialist language policies). It remains the chief language in a few parts of the island, collectively known as the Gaeltacht, which are mainly located along the western seaboard. There are three main dialects of Irish – Connacht, Munster, and Ulster – corresponding respectively to the most westerly, southerly, and northerly areas. The language has an important place in Irish culture and identity and is very widely taught in schools[3].

Irish is one of the two official languages of Ireland, the other being English. The Irish language belongs to the Celtic branch of the Indo-European family of languages, and

[1] The project is under the direction of Foras na Gaeilge, the government-funded body responsible for the promotion of the Irish language throughout the island of Ireland, whose statutory functions include the development of new dictionaries (http://www.forasnagaeilge.ie). Full details of the NEID project can be found at http://www.focloir.ie. The main contractor for setting up the project, including corpus preparation, is Lexicography MasterClass Ltd (http://www.lexmasterclass.com/).
[2] Figures from the 2002 Census.
[3] Irish is taught throughout the school system, and about 30,000 students are educated in Irish-medium schools, 'Gaelscoileanna'.

within this branch, it forms part of the Goidelic tradition along with Manx and Scots Gaelic, the other tradition being Brittonic, which comprises Welsh, Cornish, and Breton.

The remainder of the paper describes the *design, collection,* and *encoding* of the NCI in sections 2, 3, and 4. A particular area of innovation was the use of the web as a source of some of the constituent texts, and the issues arising there are covered in some detail, as are the practical issues of data organization and 'cleaning'. As part of the process we developed a morphological analyzer and part-of-speech tagger for Irish, described in section 5. Section 6 describes the project team and resources, with a view to assisting others with comparable projects in mind to assess the resources they require. Section 7 outlines possible further developments, and section 8 concludes.

2. Design

In the first instance, a detailed corpus-design document was prepared, and the target sizes for the two major components were agreed as 30 million words for Irish, and 25 million words for Hiberno-English. The other key requirements were that the corpus should form an adequate data source to support a major programme of lexicographic work, and that it should be collected and encoded within the one-year set-up phase for the new dictionary.

2.1 The English component

For the NEID project, the source language (SL) for the dictionary is English, and, more specifically, the English language as spoken in Ireland, with standard forms of British and American English also accounted for. The methodology proposed for compiling the dictionary (and used in creating over 100 sample entries) is the "translated framework" model (see Atkins 2002: 4-11), which entails three stages:

- developing a source-language framework, in which each SL headword has a fine-grained, example-rich database entry
- inserting target-language translations of key elements in this framework
- deriving final bilingual dictionary entries from the translated framework.

The detailed level of analysis in the first of these stages requires a very large corpus for the source language To this end, the Hiberno-English side of the corpus was supplemented by the 100-million-word British National Corpus (BNC[4]) for British English, and 100 million words taken from the Linguistic Data Consortium's English Gigaword corpus[5], for American English. Thus this larger corpus – "NCI+" – comprises 225 million words of English. The BNC was designed for lexicography and includes a wide range of text types, including 10 million words of transcribed speech. The Gigaword is journalism, taken from four newswire services.

Target proportions were set for different text types. These were based, in the first instance, on the design principles developed for the BNC (see Atkins, Clear and Ostler

[4] See http://natcorp.ox.ac.uk
[5] See http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

1992), but then modified in response to local factors. The factors that led us to adjust the BNC model included:

- the social and cultural salience, in Ireland, of certain genres and domains which had played a less central role in the BNC, for example reminiscences, rural folklore and the Catholic religion
- the fact that, within the category of journalism, high-status national newspapers such as the *Irish Times* were more likely to approximate to the norms of standard British English than papers with a more local remit; a higher proportion of journalistic data was therefore selected from smaller local publications
- the impossibility, given time and budget constraints, of developing new spoken corpus data, in light of which it was decided that the only transcribed speech would be taken from already-existing spoken corpora
- the plan, agreed at the outset, to include the entirely new category (in BNC terms) of data from the web

The Hiberno-English component would cover the period since the foundation of the Irish Free State in 1922, with a focus on current language. Wherever possible, texts would be classified according to whether the author was from the north, south, west or east of Ireland. We were able to record this information in a majority of cases.

2.2. The Irish component

2.2.1 Native speakers

For English in Ireland, the cases where authors are not native speakers are marginal. For Irish, however, it is a critical issue. In the whole population of users of Irish, native speakers form a relatively small percentage. As noted above, however, a majority of Irish children learn some Irish in school, and substantial numbers go on to work with Irish and write in it. Consequently, quite a high proportion of the Irish that is produced, in books, newspapers, and official documents, and on radio, television, and the web, is produced by non-native speakers.[6] It was desirable that a significant proportion of the Irish corpus should be taken from native-speaker sources. For most newspaper, web, and official material, it would not be practical to determine whether the author was a native speaker. But for books, which were to make up 50% of the corpus, it was usually possible to determine the author's status, and special efforts were therefore made to target native-speaker texts and record details of the author's origins.

2.2.2 Dialect

There are, broadly speaking, three main dialects of Irish: Connacht, Munster, and Ulster. Again, information was only likely to be available for books (with the provenance of local newspapers providing a clue for newspaper text). Our objective was that the corpus should represent all three dialects as evenly as possible, and we would aim to record as much information about the authors as we could reasonably discover. In the event, thanks largely to the encyclopedic knowledge of our Corpus Development Manager (see section

[6] While this is clearly also true of English worldwide, it is a lesser consideration for English produced in Ireland, where English is the mother tongue of an overwhelming majority of the population.

6, below), we were able to establish place-of-birth and place-of-residence for most of the authors in the corpus. While information at this level of detail goes beyond the usual needs of lexicographers, it nevertheless meets the longer-term goal of developing rich linguistic resources for Irish.

2.2.3 Diachronicity, and "high quality" Irish

A tension that arose in relation to the design of the Irish component concerned the issue of "high quality" Irish. As with many languages which have experienced falling levels of use, there is an argument that the truest form of the language is best represented by its use before the collapse set in. (For Irish, the date may be set between the Irish famine of 1844-45 and the First World War). An associated concern is that many of the documents that are produced in Irish today, and readily available in electronic form, are translations, usually from English, produced by organizations which are required (by legislation or political considerations) to supply documents in Irish as well as in English. The document may not have been translated well, and may not have been translated by a native speaker of Irish.

These factors make up the case for filtering potential corpus documents to accept only "high quality Irish". The case against has both theoretical and practical aspects. The first argument is simply that the selection of documents according to a criterion of quality is precisely the kind of subjective and value-laden process that corpus linguists have always sought to avoid. Who should judge what is good or bad Irish, and according to what criteria? It is likely to be people whose concerns lie with the literary heritage of the language, so the evolving, living language may simply be deemed "low quality" and thereby excluded from the corpus – an outcome that sits uncomfortably with the broad range of uses expected of the new dictionary.

A related argument concerns the descriptive ethos inherent in most modern corpus-building initiatives. It is desirable that a general-purpose lexicographic corpus includes the full repertoire of text-types in a language, not just a subset. While to Samuel Johnson it was an "obvious rule" that his citations should be drawn from "writers of the first reputation" (Johnson 1747), this approach was superseded a century later by Chenevix Trench in his seminal paper *Of some deficiencies in our English dictionaries* (1857). In characterising the lexicographer as "an historian, not a critic", Trench laid the foundations for modern lexicography and ushered in the rigorously descriptive methodology on which the *OED* was based.

A further argument against a carefully selected corpus of high-quality Irish concerns the lexicographic process. In the current project – an English-to-Irish bilingual dictionary – it is the source language (SL) that is to be described in detail, so the target-language (TL) corpus has a secondary role. It is largely there for checking whether candidate translations, as produced by the human translators, are "natural". (As yet there is limited computational support for this process, though see e.g. Janes (2004).) The TL corpus is only useful to the extent that it shows how arbitrary source-language phrases might be translated, so it needs very wide coverage. It has to be able to help the translator whether the expression is low-brow or high-brow, literary or mundane: "boot the computer",

"asylum seekers' hostel" and "air-freight it to Sudan" as well as "the babbling brook". The corpus should therefore be as large as possible and as broad in its coverage as possible. A quality filter is likely to compromise both goals.

We took the view that the Irish component of the corpus should include a very wide range of text-types, selected on standard corpus-gathering principles, but that we should make special efforts to describe each constituent text in sufficient detail to enable lexicographers to make informed decisions and (if appropriate) to create subcorpora of, for example, native-speaker Irish, and that where there was a choice of which texts to use to represent a text type, we would where possible choose well-written native-speaker texts.

The Irish corpus is clearly a resource for many purposes beyond the preparation of the NEID; within the project reported on here, the needs of NEID lexicography were foremost. All being well, the corpus development programme will continue, and at other stages, literary and historical studies may well move centre-stage.

The Irish to be covered by the NCI was language produced during the period from 1883 to the present day, though most of the earlier texts (written before1960) would be largely in the "imaginative" genres (fiction, drama, and reminiscences). The start date was chosen to fit with an electronic archive project at the Royal Irish Academy, which has an end date of 1882 (see http://www.ria.ie/projects/fng/index.html).

2.3 Delivery formats

One design question concerned encoding and delivery formats. For longevity, and as an interchange format, it was clearly appropriate that the corpus be delivered in XML, and in a standard corpus-encoding formalism. For the purpose, the corpus was to be delivered in the XML Corpus Encoding Standard, XCES (see http://xces.org).

However, for the corpus to be usable, an XCES corpus was only one part of what was required. The corpus also had to be loaded into a corpus-querying system (CQS). Any particular CQS will have encoding conventions more specific than those imposed by XCES, which dictate which searches can be made easily and efficiently. The tool adopted for this project was the Word Sketch Engine (Kilgarriff, Rychly, Smrz and Tugwell 2004; http://www.sketchengine.co.uk). The project included the delivery of a version of the corpus loaded into the Word Sketch Engine, in a set-up in which the type of queries a lexicographer would regularly need to make could be made quickly and efficiently.

Some of these queries involve grammar, and most involve lemmas (*initiate (v)*) rather than word forms (*initiates, initiating*). To this end the corpus was to be lemmatized (that is, with the lemma for each word specified; also known as morphological analysis) and part-of-speech tagged. While software for lemmatizing and part-of-speech tagging is widely available for English, the situation for Irish is less advanced, so a key goal of the project was the development of these tools for Irish.

A particular feature for the Word Sketch Engine is its analysis of the grammatical constructions that each word (and, more specifically, each collocate) occurs in. An input for this is a set of definitions of grammatical relations for the language. One had already been prepared for English; again, it was part of the project to develop one for Irish.

2.4 Targets

In the design stage, we set targets for the proportions of different types of text. These are presented in Table 1.

Text category	targets for Irish		targets for Hiberno-English	
	Percentages	Words	Percentages	Words
Books-imaginative	30%	9,000,000	30%	7,500,000
Books-informative	20%	6,000,000	20%	5,000,000
[Books total	*50%*	*15,000,000]*	*50%*	*12,500,000]*
Newspapers	15%	4,500,000	15%	3,750,000
Periodicals	8%	2,500,000	9%	2,250,000
[News+Per. total	*23%*	*7,000,000]*	*24%*	*6,000,000]*
Official/Govt	5%	1,500,000	4%	1,000,000
Broadcast	3%	1,000,000	3%	750,000
Websites	18%	5,500,000	19%	4,750,000
TOTALS		**30,000,000**		**25,000,000**

Table 1: New Corpus for Ireland by Text Type: target figures

3. Data Collection

Three corpus collection strategies were used:

- incorporating existing corpora
- contacting publishers, authors, newspaper companies etc. to request permission to use their texts
- collecting data from the web.

The budget did not support a scanning programme. No texts which were not already in electronic form were used.

3.1 Existing resources

Irish was one of the languages of the EU PAROLE project, and as part of that project, an 8-million-word corpus of Irish had been developed at ITÉ (Institiúid Teangeolaíochta Éireann, the Linguistic Institute of Ireland). ITÉ had continued its data collection programme after the end of the PAROLE project and had several million further words of

Irish text in its archive, with varying levels of copyright clearance. This formed the core of the Irish corpus.

For English, as mentioned above, the NCI was supplemented by the BNC and Gigaword texts. We also learned that there were two corpora of transcribed Hiberno-English speech already in existence: the 1-million-word Limerick Corpus of Irish-English[7] and the 400,000-word Northern Ireland Corpus of Transcribed Speech (NICTS) from Queen's University Belfast. Both were, with the kind permission of the corpus collectors, incorporated into NCI+.

3.1.1 Duplication in Gigaword

We had assumed that material we received from other corpora would already be well-behaved, from a corpus linguistics perspective. So we were taken aback when, on loading NCI+ into the Word Sketch Engine for the first, trial, run, we found high levels of duplicate text.

The Gigaword data is taken from four newswire services. These services provide bulletins of news stories up to several times a day. The distributor of the Gigaword, the Linguistic Data Consortium, had taken the full set of these bulletins, transformed them into minimally-marked-up XML, and packaged them as the English Gigaword corpus.

The duplication arose because successive bulletins often contained the same news story – either word-for-word identical, or modified, perhaps because there had been some new development. We applied de-duplication strategies as developed for the web, as discussed below (3.3.1).

3.2 Contacting publishers, authors, newspaper companies

Our Corpus Development Manager, who has extensive contacts in the publishing industry in Ireland, got in touch with as many publishers and other copyright-holders as possible. Potential text-donors were given a short document outlining the nature of the project and its importance for Ireland's heritage and future, and explaining (for a mainly non-corpus-aware audience) how donated text would actually be used in the dictionary-making process. They were asked to contribute to the project by sending electronic copies of texts, and signed copyright letters which allowed the texts to be used as part of a lexicographic corpus.

The fate of such letters and emails, as of so many unsolicited materials, was frequently the bin, and many hours were spent with follow-up letters, phone calls and meetings, in the enormous effort required to coax publishers into allowing their texts to be used. Our experience with the BNC and other corpora had prepared us for this: the publishing business is based on the sale of copyright material, so it is not surprising that the default response from the publisher, when asked to give texts for free, is "no". Also it takes time to explain what a corpus is and to convince publishers that it does not threaten their

[7] www.ul.ie/~lcie/

business, and publishers are frequently busy and see no reason why they should spare the time. High levels of charm and persistence are required, and it was a large task.

Gathering copyright agreements could be awkward for a number of reasons. In the closing days of the project, as we were finalizing the set of signed copyright agreements, a last check with one publisher resulted in him discovering, to his dismay, that the copyright in a set of texts he had given us had recently reverted from him, the publisher, to the authors. We then had to recommence the wooing process with those authors – two of whom responded with a very definite "no". We would note that, those two cases notwithstanding, the response was in the main very positive, with most copyright-owners pleased to be associated with the project.

3.2.1 Text delivery and pre-processing

Once we had agreement-in-principle, we needed to actually acquire the text. Sometimes it was sent on CD or other media, sometimes it was received by email. Occasionally, despite promises, it took further charm and persistence before (a) the signed copyright permission form and (b) the text itself, were in our hands. For some texts, the process was not complete within the time limits of the project.

As expected, text arrived in a wide range of formats, including proprietary forms such as Quark, so the first step was to reduce everything to the same plain-text format. Further steps are covered in section 4, below.

3.3 **Web data**

The web offers enormous possibilities for corpus development, for language of all varieties (Kilgarriff and Grefenstette 2003) and for 'smaller' languages in particular (Jones and Ghani 1999). Following earlier successful collaborations, we worked with Infogistics Ltd., a company with expertise in computational linguistics, web crawling and large scale data transformation.

Infogistics ran some experiments to determine how much Irish there might be on the web, using the method presented in Grefenstette and Nioche (1990): identify some words which are common in Irish but do not occur in other languages; find their frequencies in a known corpus of Irish (the PAROLE corpus); find their frequencies on the web (using a search engine such as Altavista); and scale up. They felt confident they would be able to find 15 million words. They undertook to deliver 15 million words of Irish and 20 million words of Hiberno-English, processed into XCES-compliant XML. The text was to be as varied as possible, from a wide range of websites. They delivered the data in three iterations, and at each turn, we inspected it and reported back on any problems we encountered, which they addressed prior to the next iteration.

We first briefly discuss some recurring themes of corpus development from the web, then duplication, and then how we found Irish and Hiberno-English material.

Input formats (e.g. .txt, html, pdf, rtf, MS-Word, postscript): How many different document formats can be converted to plain text and used in the corpus? We used all of those listed. We avoided "dynamic" pages, which are generated when the user calls them up, as they introduce assorted complications such as highly repetitive boiler-plate text, or text generated by computer, and would not increase the spread of the corpus.
***Formatting*:** the corpus collector's default model is continuous uninterrupted text, but on the web, frames and pages are often used to split up a text, and text is often split across different, short web pages. Documents which are "split" in ways which do not respect linguistic structure (such as sentence-boundaries and paragraph-boundaries), must be either rebuilt, so that the result reconstructs the correct linguistic structure, or rejected. We rebuilt in the straightforward cases and rejected in the others.
Character representations: Standard Irish uses only the Latin-1 character set,[8] so the problem was limited; nonetheless there were various complications. For example, pdf files represented accented characters in different ways, depending on the software used to generate the pdf and the kind of source text the pdf was generated from. We also noted that only a small proportion of web pages declare character sets in the WWW-approved way.
Navigational material: text like "click here" "next page" "further details" is specific to web genres, and will distort the statistics if left in a lexicographic corpus. Common navigational phrases and constructions were identified and removed, for both Irish and English.
Lists: the web contains many lists: price lists, product lists, the players in a sports team, the companies in a business sector, local councillors, and so on. It is not obvious where lists should be included in a corpus, and where excluded, and much will depend on the uses to which the corpus will be put: if it is to be used as a source of names, then lists will be very useful, and if as a source of technical terminology, then product lists may be particularly valuable. Also some lists contain noun phrases, others may contain full sentences or more. For our (lexicographic) purposes, the rule of thumb was that we most wanted language when it occurred in sentences, and lists which displayed no sentence-like characteristics were rejected. We checked to see whether strings of texts included items we recognize as verbs. (The issue interacts with unit-size and duplication, see below.)
Linguistically-aware spam: there is an ongoing "arms war" between spammers and the search engines (notably, as market leader, Google). Google and others want to point users to the most relevant websites, and spammers aim to inveigle themselves into that process so that Google directs users to their websites. Search engines work through words as search terms (amongst other things – Google also uses links), so text is one of the battlefields. The spammers invent new stratagems, which the search engine teams strive to detect and counteract, in an ongoing process. The manoeuvres include adding thousands of words into web pages, in the same colour as the background, so they are visible to search engines but invisible to users. Google counteracts by ignoring lists of words, maybe drawn from a dictionary, that do not look like continuous text, and the spammers counteract by making their spam look more text-like. We paid heed to the

[8] The one much-used character not covered by Latin 1 was the Euro sign €; others have encountered this issue before, and developed Latin-0, roughly Latin 1 + €. We standardized on Latin-0 (officially ISO 8859-15, also known as Latin-9).

known manoeuvres of spammers and developed strategies for excluding 'text' with spam-like characteristics.

3.3.1 Duplication

Duplication is pervasive on the web, for a wide range of reasons, from caching to quotation and plagiarism. Sometimes the duplication is exact, sometimes approximate. Web corpora which have not been "de-duplicated" are highly problematic, and any statistics derived from them are likely to be misleading. The level of duplication in the "Irish web" was substantially higher than first estimated, leading to some concerns as to whether we would achieve the 15-million-word target.

Duplicates present a theoretical question: what is the textual unit for identification of duplicates? If the unit is set too large, lots of duplicates will remain, but if the unit is set too small, as, say, a sentence, then common sentences like "How do you do?" will be rejected as duplicates, throwing out the linguistic fact that this is a very common expression and destroying the integrity of documents from which it has been excised.

The algorithm developed by Infogistics considered units at both the sentence level and the text level, and rejected texts where x% of the sentences were duplicates, as follows:

1. order texts, from longest to shortest.
2. set sentence-db to empty
3. for each text
    a. set sentence-count and duplicate-sentence-count to 0 and empty the buffer
    b. break into sentences
    c. for each sentence over 25 characters long
        i. normalize:
            1. delete all non-alphanumeric characters and characters above ASCII 127
            2. convert all characters to lower case
        ii. if normalized sentence is in sentence-db (using an exact match), increment duplicate-sentence-count; else add normalized sentence to buffer
        iii. increment sentence-count
    d. if duplicate-sentence-count > x% of sentence-count reject text; else accept text, add sentences in buffer to sentence-db.

The normalization means that different variants of a text (where, for example, one is derived from a Word version, another from html, and a third from pdf) will be mapped to the same normalized version. The reason for ordering the texts is to address the case where one text is a part of another. We wish to keep the whole and reject the part, which is achieved by considering texts in length order. Values of x% of 60% and 80% were explored. The value made little difference to the number of texts rejected, confirming the validity of the approach. A 60% value was selected.

Where texts contain very few sentences, one would expect the method to be less reliable. In common with others using the web as a corpus, we found that very short pages (and also very long pages) tend not to contain usable text. But since we in any case rejected

web pages which did not contain a reasonable number of sentences, the issue did not arise.

Our use of the corpus to date shows the method to have been fully effective. No unwanted duplication has been encountered.

3.3.2 Irish

Our two strategies for gathering Irish were (1) going to known Irish-language sites and downloading the whole site, and possibly also pages linked to from that site; and (2) entering a set of Irish words in Google and harvesting the pages that Google found. In both cases, it was necessary to check whether each page was Irish. For this a high-accuracy language-identifier was developed, using the PAROLE corpus as a sample of Irish to start from.

One issue which was not fully resolved in the PAROLE corpus was mixed-language text. A proportion of documents "in Irish" also have sections, or sentences, or phrases in English, and in fact about 5% of the text in the PAROLE Irish corpus is in English. Web pages often included quotations of English in otherwise Irish text or vice versa, or mixed-language dialogue, or "bitexts" where tables had an Irish column and an English column, or a paragraph followed by its translation.

We used "paragraphs" as the unit for language identification. Approximations to paragraphs were identified using low-level cues, predominantly line-breaks and corresponding html markup. We developed an Irish-language-identifier based on Irish-language-only words and letter sequences, and applied it to the paragraphs, accepting them only if the identifier deemed them Irish. Some units thus identified were however too small to accept or reject without looking at the context. We accepted a paragraph as Irish if it was a long paragraph which the language-identifier identified as Irish, or a short paragraph which was associated with Irish long paragraphs.

Scots Gaelic and Irish are in many respects very similar, and hard for language identification algorithms to discriminate. We used the heuristic that text from a website with an .ie domain was far more likely to be Irish than Scottish, and rejected non-.ie pages unless they contained keywords such as place names which indicated Ireland.

A complication arose in relation to spelling conventions. Irish has had official spelling rules in place since 1958. Moves to standardize Scots Gaelic are more recent, with policies dating from 1981. One particular website, based in Scotland, had a large quantity of Irish material, but the words were spelled using Scottish-style conventions, so this data was rejected.

We also encountered several websites dedicated to the teaching of the Irish language. These presented an acute form of the mixed-language problem, with phrases, sentences and paragraphs of Irish mixed with explanations and instructions in English. This material, too, was rejected.

3.3.3 Hiberno-English

The obvious question for Hiberno-English was: how could it be distinguished from other varieties of English? While an optimal answer might depend on internal evidence, it was an output, not an input, of the project to identify what was characteristic about Hiberno-English. We considered various strategies for identifying Hiberno-English websites. The one we used was this: assume that the English on a website is Hiberno-English, if there is also Irish on the website. It seems plausible that most sites with content in Irish will be produced in Ireland by Irish people, so the English on those sites can be assumed to be Hiberno-English. Using this heuristic, there was no shortage of Hiberno-English web data available.

3.3.4 Newspapers

In terms of collection strategy, newspapers turned out to be intermediate between "web collection" and "ask the publisher" collection. When we asked newspaper publishers and they gave us permission to use their text, they told us the easiest way for us to acquire the texts was from their websites, and this is what we did. (This situation only arose for Hiberno-English newspapers.) For classification purposes, text from printed newspapers was categorized as "Newspaper" (see Table 1, above) even if we collected it by downloading from the web.

3.3.5 Web text types

The questions, "what types of text are there on the web, and in what proportions?" are large, hard, and under-researched (Kilgarriff and Grefenstette 2003). To give an idea of the range and variety of texts gathered for Irish in this project, we list in Table 2 a dozen websites from which we took substantial quantities of text, along with the types of document found in each.

Name	***Organization type***	***Document types include:***
FUTA FATA	Magazine	Reviews of, and extracts from Irish novels, books of poetry
Galway County Council	County Council	Policy statements, application forms
University College Galway	University	Policy statements, statements of objectives, reports
Department of Community, Rural and Gaeltacht Affairs	Government Department	Speeches and press-releases from the Minister, news reports
Údarás na Gaeltachta	Regional Development Agency	Announcements, forms, policy statements, grant schemes
Ógras	Irish-language Youth Organisation	Activities, competitions
Sinn Féin	Political party	History, policy, events
Gaelport/Comhdháil Náisiúnta na Gaeilge	Umbrella Irish-language organisation	Electronic newsletter
Rondomondo	Magazine	Arts, music, drama
Irish Army/Navy	Armed forces	missions, career descriptions
Raidió na Gaeltachta	Radio station	Notices, news
Aran Mór College	College	Advertising, programmes, activities

Table 2: Sample of websites and text types for Irish web corpus collection

Text types requiring particular consideration include chatroom, email, bulletin boards and discussion lists. They are sociolinguistically interesting as they are new genres, native to the web and distinct from pre-existing genres. However they are hard to use in the same way as more traditional textual material. There are large numbers of abbreviations, reduced forms and spelling mistakes, and any Irish material found in them tends to be freely mixed with English. This causes problems for the corpus developer and for the lexicographer, for example when they want to find all examples of a word: occurrences with non-standard spellings and spelling errors will be missed. For these reasons, for the time being, these genres have not been included in the NCI.

3.3.5 Web text selection

At 15 million words for Irish and 20 million for Hiberno-English, our goals for web text collection were much higher than required for the NCI: as Table 1 shows, our NCI targets were just 5.5 million for Irish and 5 million for Hiberno-English. This gave us a large surplus of web data.

For Hiberno-English, we carried out a careful inspection of the downloaded data in order to identify recurrent problems. The results were fed back to Infogistics, who refined their search algorithms accordingly (and so on in an iterative process). Once we were satisfied that the data was, broadly-speaking, of good quality, we then took a random sample (while retaining the full range of domains, to keep the corpus as broad as possible.). We followed a similar approach for the Irish web data, but here the checking process also gave us the opportunity to filter according to "quality of Irish" and thereby to pay heed to the concerns about low-quality translations as discussed in Section 2.2.3. Our senior Irish linguist studied a sample of each of the main websites we had used as sources of data, and declared them "good", "OK", or "bad". According to his stringent criteria, there was just enough "good" and "OK" text to meet our needs, so these were the texts we used.

3.4 Actual corpus composition, compared with targets

The table below shows the composition of the final corpus, compared with our original targets

Text category	Irish		Hiberno-English	
	Words: actual	Words: target	Words: actual	Words: target
Books-imaginative	7,600,000	9,000,000	6,000,000	7,500,000
Books-informative	8,400,000	6,000,000	7,000,000	5,000,000
[Books total	*16.000,000*	*15,000,000]*	*13,000,000*	*12,500,000]*
Newspapers	4,500,000	4,500,000	5,300,000	3,750,000
Periodicals	2,600,000	2,500,000	700,000	2,250,000
[News+Per. total	*7,100.000*	*7,000,000]*	*6,000,000*	*6,000,000]*
Official/Govt	1,200,000	1,500,000	1,000,000	1,000,000
Broadcast	400,000	1,000,000	0	750,000
Websites	5,500,000	5,500,000	5,000,000	4,750,000
TOTALS	**30,200,000**	**30,000,000**	**25,000,000**	**25,000,000**

Table 3: New Corpus for Ireland: target figures and actuals

For the most part, our *a priori* targets could be met. The biggest disparity is in the Books category, where, it transpired, imaginative texts were harder to find (for both languages) than originally anticipated. No formal targets for dialect or native-speaker provenance had been established but nonetheless it is good to note that almost half of the text in the Books category of the Irish corpus can be reliably attributed to Irish native-speaker authors and around 80% is categorized as belonging to one of the three major dialects.

4. Encoding

Once a set of documents has been collected, a number of choices must be made and acted on before it is in an optimal state for use by linguists and lexicographers. We call this stage 'encoding'.

Encoding needs to be a goal-driven process. It is the goal of the exercise that defines what counts as a job well done. Our goals were to support (1) the lexicography for the NEID, and (2) research in Irish and Hiberno-English in general, with, as noted above, the corpus delivered both in XML and within the Word Sketch Engine. (The content of both versions would be the same.)

For the English side of the corpus, the relation between NCI and NCI+ (i.e. NCI plus BNC and Gigaword) was a particular challenge (see Fig. 1). The XML delivery related only to the NCI, but, for the Word Sketch Engine delivery, it had to be straightforward for users to query both Hiberno-English-only, and the whole English-language component. The encoding of the various components of the NCI, the BNC and the Gigaword needed unifying.

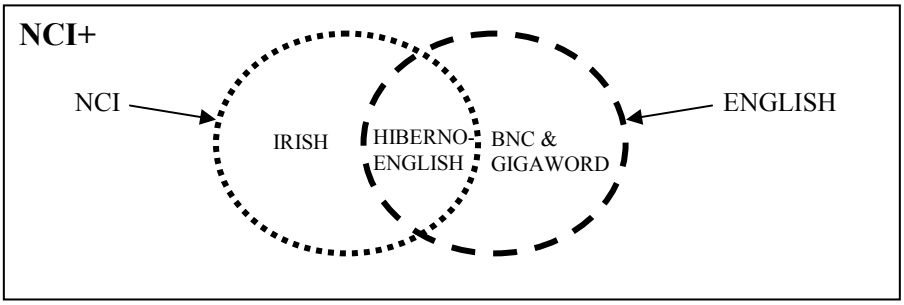

Fig. 1: NCI+ comprises the NCI (which has both Irish and Hiberno-English components) and additional English corpora

In this section, we first discuss the organization of the task, the encoding of the text; and then, the gathering and encoding of meta-information about each text, its 'header'.

4.1 Organization: intranet, document IDs, documents, files, file systems

4.1.1 Project intranet

Corpus development work was distributed across teams in Ireland, Scotland, the Czech Republic and England. At an early stage of the project, a webserver was set up with an intranet, so there was a definitive place for the corpus to be stored, and for texts to be placed when one team had processed them, so that the next team could pick them up for the next stage of processing.

4.1.2 Document Ids, filenames and directory structure

Corpus development involves very large numbers of documents. It is easy for documents to get lost. In other corpus projects, we had witnessed all too much effort expended on looking for lost files, so it was a priority to set up a system which minimized the risk.

Our strategy was to assign to each document, at the earliest possible stage in the process, a document identifier and a specification for where in the file system the document was stored: the structure of the file system would map directly onto the document identifier. Identifiers would be:

- unique: different teams would be collecting different parts of the corpus, so it was essential to preclude the possibility of different teams assigning the same ID to different documents
- short and informative
- not subject to change at any later stage[9]

This implies a rigid "one file per document, one document per file" convention. However, different *versions* of the same file could be stored with the same core filename but a different suffix, such as .txt or .xml.
The identifiers were to have eight characters. They, and the corresponding file system, were constructed as follows:

- at the highest level, NCI+ had two components, Irish and English. Thus
  - the first letter of the document ID was either ***i*** or ***e***
  - there were two top-level directories,[10] `irish` and `english`
- at the next level, the material was either collected by Lexicography MasterClass directly from copyright holders, or from the web, or newspapers, or was part of the Limerick corpus, or NICTS, or the PAROLE corpus (CNG) or other holdings of the ITÉ, or from the BNC or the American-English Gigaword. Each of these was given a single-letter mnemonic, and appropriate subdirectories were created under the top-level directories.

- the LexMC, CNG and ITE, newspaper and American material was gathered from a range of publishers and news services; each was allocated a two-letter code, which became the fourth and fifth characters of the document ID and was associated with the next level of subdirectory. For web material, each of the 25

[9] When exploring the relation between spoken-BNC audio material and published text, Kilgarriff and colleagues unearthed no less than seven layers of numbers for the "same" conversation, and an extensive piece of detective work was needed to link tapes to published files. Each renaming introduces an additional possibility of confusion, at each stage of the corpus preparation process.
[10] In some operating systems, 'directories' are called 'folders'.

websites contributing most text to the corpus was assigned a letter from a to z (with the remaining letter for "other"); this was the fourth character, and the fifth was always 0.

- The remaining characters were either a number, or an identifier taken from the source corpus. Each team used its own numbering system; this part of the document ID was information-bearing for the Gigaword and some periodicals where it signified the month and year, or issue number, or for Irish legislation the year and Act number. For other components it was not information-bearing, except that, for the BNC, Limerick Corpus and NICTS, it included the document identifier for the document in the source corpus, for ease of cross-reference.

Thus document IDs can be read as follows:

- ***exbn0006*** is English, collected by Lexicography MasterClass (***x***), from Brandon Books (***bn***), file number 0006
- ***eb000a04*** is BNC file A04
- ***eaap0594*** is English, from the Gigaword (***a*** for American), from the Associated Press newswire (***ap***), stories for May 1994
- ***itgm0074*** is Irish(***i***), from PAROLE collection (***t***), from publisher An Gúm (***gm***) and has number 0074
- ***icco1039*** is Irish collected by ITÉ (***c***), from publisher Coiscéim (***co***), with number 1039.

The conventions also gave us convenient and well-defined ways for talking about subcorpora. The ***ex*** subcorpus was the English subcorpus collected by LexMC; the ***iw*** was the Irish web subcorpus. We found this useful as these different subcorpora frequently had a range of distinct characteristics and needed encoding differently. It was helpful to be able to tell immediately which file belonged where, and was likely to exhibit which patterns.

While details of file-naming conventions are of no theoretical interest, we believe they have contributed significantly to the speed with which we have been able to develop the corpus, through the avoidance of confusion and lost files; it may prove a useful model for others to follow.

4.1.3 File=document

As noted above, our system requires a one-to-one mapping between files in a computer's file system and documents. In XML Corpus Encoding Standards (XCES) terms, this meant that there was a single `cesDoc` element, containing one `cesHeader` and one `text` in each file. This imposed the constraint that all the text in one document, or file, had to share the same header information: it would be incorrect to put together items by

two different, known, authors in the same file, as then there would be two conflicting values for 'author'.

While 'file' is a well-defined concept, 'document' is open to interpretation. A book is a single document, and all the text from a book forms a single document. But for newspapers, magazines and web material, it is less clear what a document is. A narrative on the web may be spread across several web pages (where each web page has its own URL, or address) while having the same claims to coherence and single-document-hood as any book.

Our definition of 'document' was pragmatic, and was dictated by the appealing simplicity of the one-file-per-document principle. While files can be of any size, it is convenient for a range of purposes if there are not too many of them (hundreds or thousands, but not tens of thousands) and they are not too big (more than a megabyte is often inconvenient). That imposed constraints on files, and we would gather text together into 'documents' to suit.

In the published form of the Gigaword corpus, there are very large numbers of small files, each containing a number of stories (where a story is something that may be used as a newspaper article). We ran together the stories from the same news service in the same month. The only header fields available were (1) the newswire service it was taken from, (2) the date, and (3) the facts implicit in it being American English journalism All these facts were common to all the stories (with the date shared, down to the level of detail of the month) so a single header could be written for all the text. XCES provides several mechanisms for grouping texts together in a single document: we used the structure:

```
<cesDoc>
   <cesHeader>  …  </cesHeader>
   <text>
     <group>
       <body>  (first story)  </body>
       <body>  (next story)  </body>
       …
     </group>
   </text>
 </cesDoc>
```

For the web material, tens of thousands of web pages were used so it was not convenient to give each its own file. There were extensive discussions about how to organize the web pages, since

- one file per web page gave too many files
- one file per website gave some overlarge files, and
- there was no other obvious intermediate level of structure that provided any semblance of putting similar things together.

Our strategy was this: web pages from the same website, and, as far as possible, from the same directory within the website, were put together in the same document (using the XCES 'group' mechanism as above). There was no guarantee that directory structure on a website corresponds in any way to logical structure. While the different parts of a single narrative are usually in the same directory, they need not be, and also they may be mixed in with any number of other files. Putting successive components of a single narrative in the correct, uninterrupted order, while desirable, was not critical for lexicography, and was done only for straightforward cases. The fact that a large number of (possibly disparate) web pages would now all share one header was not a practical problem since headers for web material were in any case minimal.

4.2 Text encoding

4.2.1 Text cleaning, paragraph markup

For each of the documents we collected from publishers, once it had had its ID assigned and had been saved as raw text with matching filename in the appropriate place in the file system, we examined the text in an editor. We counted words, and deleted parts of the text which were not suitable for a lexicographic corpus. The 'unsuitable parts' included, for books:

> title pages, tables of content and other tables, figures and diagrams, footnotes and endnotes, indexes, page headers and footers including running titles, mathematical and scientific formulae, extensive quotations and other sections in other languages, e.g. non-English for the English corpus and non-Irish for the Irish corpus.

From newspaper and magazine text we also removed:

> crosswords, TV listings, isolated names and addresses dates from advertisements, racing results, lists of team members etc.

Paragraph tags were then added, semi-automatically. Poetry and plays were identified, and XCES markup suitable to them was inserted. Symbols in the text such as "&", "<" and ">" which would interfere with XML validation were converted to XML entities, becoming "&", "<" and ">" respectively. Once this XML markup had been added, the document was 'topped and tailed' with suitable start- and end-tags, and then validated against the XCES DTD in an XML editor. The validation process often uncovered character-encoding issues, which were then fixed. (A similar 'cleaning' process for the web data is described above.)

To our surprise, the 'cleaning' removed an average of a third of the words in a text.

4.2.2 Linguistic markup

A corpus is more useful if it is morphologically analyzed and part-of-speech-tagged. For English, we used existing tools. Although the BNC is published complete with part-of-speech tags, they are CLAWS-5 tags and we chose to standardize on the widely-used Penn tagset so we re-tagged the BNC as well as the other 125 M words.

Irish linguistic tool development was a substantial contribution of the project and is described in Section 5.

4.3 Header encoding

The headers needed to give whatever information a lexicographer might need about a text, including feature-values which would potentially be used in corpus queries. They had to deal with all the very different NCI+ components in a single, consistent form, so the lexicographer did not need to remember that, for example, what the BNC called 'subject', the NCI called 'topic'. There were, of course, pragmatic constraints on how much detail could be provided about each text, given the number of documents and the scope and budget of the project.

In this section we first discuss the header design, then, how the values for each feature were identified for each corpus component, and then show how header information can be used in the Word Sketch Engine.

4.3.1 Header design

Within XCES, a document header (`cesHeader` element) is structured. In the input format for the Word Sketch Engine, it is an unstructured set of feature-value pairs. While NCI headers are XCES-compliant, and nomenclature is taken from XCES, we do not discuss the structure or other XML/XCES issues (or other bookkeeping features) here.

Header fields are of two kinds: 'free text' ones, and ones with a fixed set of possible values. The former are:

> *h.title, h.author, publisher, pubPlace, pubDate, author-birthplace, author-dob, author-residence*

h.title and *h.author* are drawn from XCES and are the features XCES uses for simply stating the author and the title. They, and publication details, are standard bibliographic information. The three last features were only filled in for Irish books.

The fields with a fixed set of possible values are specified, with their possible values, in Table 4. For most features, values will not be specified for some documents, which is equivalent to them being given the value 'u' or 'unknown'.

Feature	**Values**	**Note**
language	ga en	ISO 639 Language Codes
langvariety	ie br am	Hiberno/British/American: applies

		to English only
docid	unique 8-character document IDs	(see details above)
nativesp	y n u	applies to Irish only
nativesp-dialect	connacht munster ulster u	applies to Irish only
ie-region	n s e w u	applies to Hiberno-English only
translation	y n	applies to Irish only; default is 'n'
time	1883-1959 1960-1999 2000-on u	applies to Irish only
biog	yes no auto	applies to Irish only; default is 'no'
mode	written spoken	
medium	book newspaper magazine periodical acad-journal website-news website-other email-webchat dissertation official-govt unpublished ephemera broadcast-radio broadcast-tv conversation interview lecture meeting unknown	Used in defining target proportions; see Table 1; several values (e.g. email-webchat, dissertation) were unused.
genre	inf imag	All documents to receive a basic classification for genre. Used in defining target proportions; see Table 1.
genre2	fiction poetry drama non-fiction information instruction official unknown	A more fine-grained genre classification.
topic	hard-applied-science social-science govt politics history religion-philosophy business-finance arts-culture leisure geography health news legislation unknown	
targetreaders	general schools academic teenagers children adult-learner unknown	

Table 4: NCI header fields with fixed sets of possible values.

4.3.2 Populating the headers

Once the header fields were defined, the next task was to establish the value for each, for each document. To record these details we set up a web database. The interface had a text-input box for each free-text field and a menu for each fixed-value-set field. The Document IDs served as primary keys.

A large mapping table was produced which stated, for each of the eleven corpus components (identified using their two-letter codes; ***iw, ic, it, ix, ew, ex, el, eq, eb, ea, en***) how each field was to be filled. For the books gathered from publishers, the instruction was usually just "use manually-input data". For some fields, the mapping was implicit in the component name: for all the ***i*** components, *language* was set to *ga* and for

all the ***e*** components, to *en*; for all ***ew*** and ***iw*** documents the value for *medium* was *website*.

For the 'books' component of the NCI, header fields were filled manually; for the other parts, it was largely automatic. The database eventually held almost four thousand records. Approximately 400 Irish and 300 English were entered manually, the remainder automatically generated.

For the CnG, BNC and Gigaword components, the task was one of identifying where, if anywhere, the information required to fill an NCI+ header field was to be found in an existing corpus header.

It was necessary to fall back on 'defaults' and 'unknown' in various cases, particularly for the web and Gigaword material. However the basic information that, for example, Gigaword always had *lang=en, langvariety=a*, *genre=inf, medium=newspaper* is a large part of what is useful for lexicography.

The online database allowed all team members to check on a document at any time and records could easily be updated. This was particularly useful where details relating to the author and the text, such as author age and place of birth, only became available after further investigation. Updates to the permission status from copyright holders were maintained in the same way. The database provided a range of reports, which were critical for monitoring progress.

Procedures were written to transform database contents into XCES-compliant XML headers. The methodology thus combined using XML for data exchange with a relational databases and the SQL query language for distributed data input, progress-tracking, and the ability to perform bulk updates.

4.3.3 Subcorpora in the Word Sketch Engine

The Word Sketch Engine has a 'Create Subcorpus' function. Once the user has created and named a subcorpus, they can specify it and then search within it. Thus, in the NEID project, where lexicographers have a suspicion that an English word behaves differently in Ireland to elsewhere, they will be able be able to set the corpus to "Hiberno-English only" and examine its behaviour there. If they wish to contrast an Irish word's use pre- and post-1960, they can do this by first setting up two subcorpora and then searching each in turn.

The Word Sketch Engine interface for creating a subcorpus, as it appears when the corpus is the English component of the NCI+, is shown in Fig. 2.

The numbers given are numbers of words in each component, and are relative to the specified corpus which has been selected, so if a subcorpus (like Hiberno-English) has been selected, then the numbers will be the numbers of Hiberno-English words in each component.

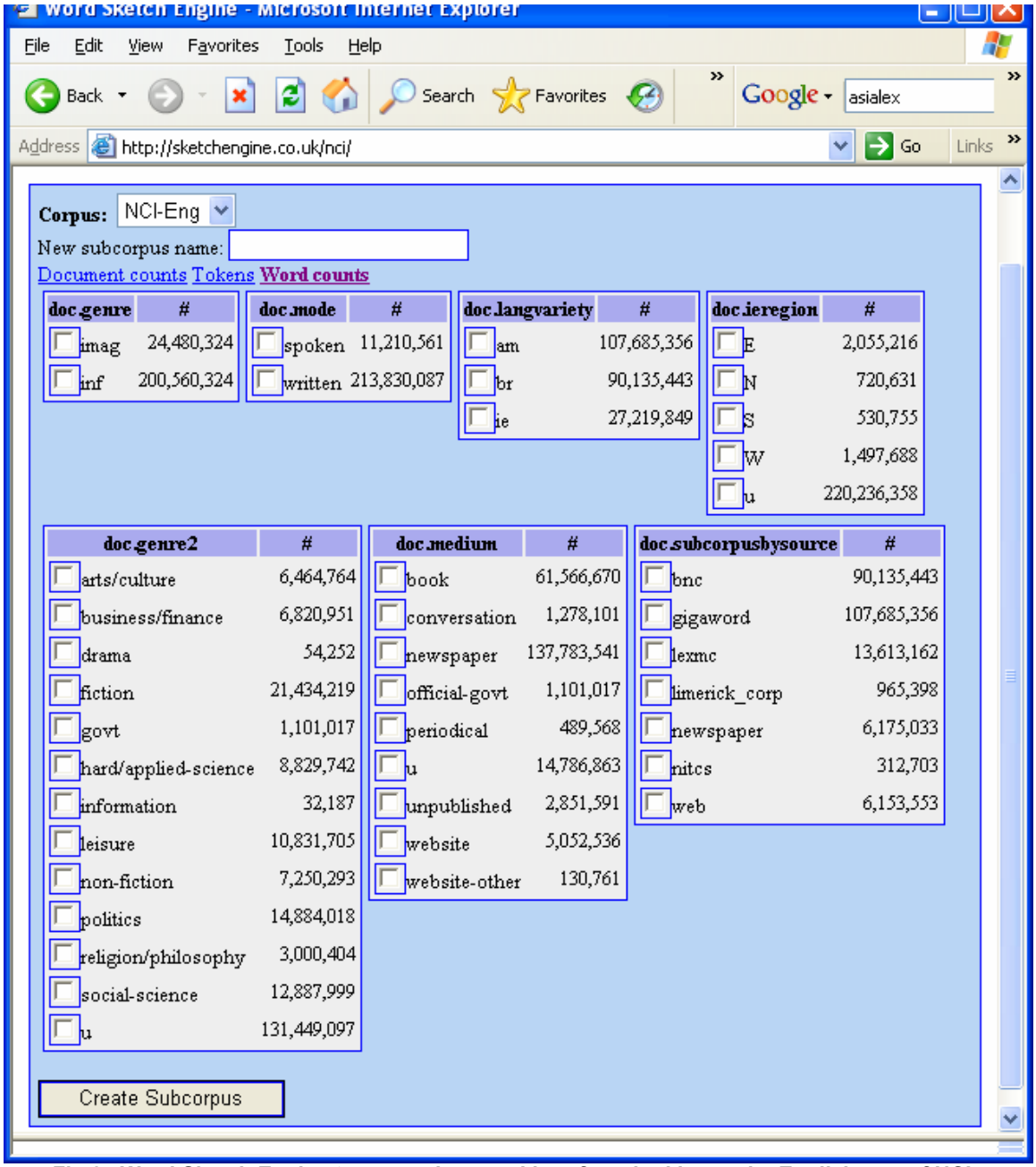

Fig 2. Word Sketch Engine 'create subcorpus' interface, looking at the English part of NCI+

5. Irish linguistic tools

In order to linguistically annotate Irish, we needed a morphological analyser and a part-of-speech tagger. For Irish word sketches, we would in addition need to specify grammatical relations for Irish.

In choosing a tagging methodology, we considered existing resources and how best to use them.

- A tagset for Irish had been developed within the PAROLE project, by members of the NCI team (http://www.ite.ie/corpus/pos.htm)
- A finite-state tokenizer and morphological transducer for Irish had already been developed (Uí Dhonnchadha 2002; Uí Dhonnchadha et al 2005).
- We established that a constraint based tagger[11] was available to us

The approach would all be finite state. We would perform morphological analysis on the text. The morphological analyser outputs all of the possible lemma and tag combinations for a particular token. Constraint Grammar rules would then be applied to this output in order to choose the appropriate analysis for the wordform based on its context in the sentence.

Irish has complex morphology. It is an inflectional language in which nouns have gender and are inflected for number and case. Adjectives agree with nouns in terms of gender, number and case, and verbs are inflected for tense, mood, person and number. There are morphosyntactic dependencies whereby the initial phoneme of a word mutates depending on the previous word. In example (1) we see that *bean* changes to *bhean* following the definite article. Following the article this particular mutation only occurs in the case of feminine nouns. Example (2) shows a similar mutation occurring when a verb form is preceded by a negative particle.

(1) a. ***b****ean* "a woman" (there is no indefinite article in Irish)
b. *an* ***bh****ean* "the woman"

(2) a. ***c****eannaím* OR ***c****eannaíonn mé* "I buy"
b. *ní* ***ch****eannaím* OR *ní* ***ch****eannaíonn mé* "I do not buy"

Irish also contains consonant harmony whereby a broad suffix goes with a broad stem and slender suffixes with slender stems. In some cases (3a & b) the suffix varies and in others (3c) the stem changes to preserve this harmony. This is shown orthographically by the vowels accompanying the consonants.

(3) a. *carr* "a car" (singular), *carr****anna*** "cars" (broad pl. suffix)
b. *méid* "an amount" (sg.), *méid****eanna*** "amounts" (slender pl. suffix)
c. *rud* "a thing" (sg.), *ruidín* "a little thing" (stem is slenderised to accommodate slender suffix.)

The Parole tagset, in which tags comprise up to nine characters each representing a linguistic feature, allows for the expression of all linguistic features, which are salient for Irish morphology. In (4) the Parole tag for "bháisteach" is Ncfsc where N=noun, c=common, f=feminine, s=singular and c=common case.

(4) `<w ctag = "Ncfsc" base = "báisteach">bháisteach</w>`

[11] Constraint Grammar vislcg downloadable at http://www.sourceforge.net

Internally, the formalism used was a fuller and more explicit notational variant of the Parole tagset in which each feature is represented by a short name, as in the example (5) below.

(5) "<bháisteach>" "báisteach" Noun Fem Com Sg Len

5.2 Tool development

Table 4. shows the sequence of processing stages. We then describe the development of tools for each step for Irish.

PROCESSING STAGE	*OUTPUT*
1. TOKENIZATION	Tokenised Text
2. MORPHOLOGICAL ANALYSIS	Multiple Lemma/Tag choices
3. CONSTRAINT GRAMMAR DISAMBIGUATION	POS and Lemmatised Text
4. XML FORMATTING	XCES POS and Lemmatised Text
5. BINARY ENCODING FOR CORPUS QUERY SYSTEM	Binary corpus data
6. GRAMMATICAL RELATIONS FOR CQS	Word Sketches

Table 4: Text processing steps

5.2.1 Tokenization and Morphological Analysis

The existing tokenizer and morphological analyser/generator for Irish (Uí Dhonnchadha, 2002) was built using Xerox Finite-State Tools (Karttunen and Beesley, 1992; Beesley and Karttunen, 2003). This lexical transducer implemented all of the inflectional rules for Irish and contained a test lexicon of approximately 1500 lemmas, which included the 1000 most frequently occurring word-forms in the CNG corpus. Its recognition rate was on average 85% on unrestricted text.

In order to achieve accurate POS tagging the recognition rates needed to be increased substantially. This was achieved by

- increasing the lexicon
- adding derivational morphology rules and
- implementing morphological guessers.

The lexicon was increased by semi-automatically converting a 15,000 word pocket Irish-English dictionary (An Gúm, 1980) to Xerox *lexc* format. As newspaper and web texts in particular contain a high proportion of proper nouns, lists of names and places were also scanned and incorporated into the lexicon (Uí Dhonnchadha *et al*, 2005). Average recognition rates increased to 95% on unrestricted text.

As many words are derived by affixing prefixes and/or suffixes to existing stems, the lexical transducer was augmented by including approximately 150 common prefixes and some derivational suffixes which can be concatenated to nouns, verbs and adjectives as appropriate. New rules were included for the morphological changes which occur at affix-stem junction.

A lexicon of approx 20K Irish items is still modest, and a method was also needed for dealing with unrecognised words. This function was implemented as a series of morphological guessers (Beesley & Karttunen, 2003, p444) which make use of the distinctive suffixes, syllable structure, initial capitals and particular characters in the token to identify verbs, adjective, proper nouns, nouns and foreign words. The guessers were applied in order, first testing an unknown word to see if it could be a verb, and if that failed, then, an adjective, and so on until a possible analysis applied. This provided a high degree of accuracy in selecting the part-of-speech. However the lemmas tended to be unreliable due to the changes which most stems undergo when combined with an affix. Further work in this area could prove fruitful.

5.2.2. Constraint Grammar Disambiguation

The following is a sample of output after tokenization and morphological analysis has been applied to the phrase "*Tháinig an bháisteach"* (The rain came).

(6)
```
"<Tháinig>"
      "tar" Verb PastInd Neg Len
      "tar" Verb PastInd Len

  "<an>"
      "an" Art Sg Def
      "an" Part Vb Q Cond
      "an" Part Vb Q Fut
      "an" Part Vb Q Past
      "an" Part Vb Q Pres
      "is" Cop Pres Q
      "is" Cop Pres Dep Q

   "<bháisteach>"
      "báisteach" Noun Fem Voc Sg Len
      "báisteach" Noun Fem Com Sg Def
      "báisteach" Noun Fem Com Sg Len
      "báisteach" Verbal Noun Rel Len
```

As we can see, each token is ambiguous in that more that one morphological analysis is possible. Constraint Grammar (CG) is applied to the output of the morphological analyser[12] and by applying language specific rules it endeavours to select the correct analysis, as in the following sample.

[12] Some minor reformatting is required. Thank you to Trond Trosterud, University of Tromsø and Anssi Yli-Jyrä, University of Helsinki for sharing their perl script.

markup. Some units thus identified were however too small to accept or reject without looking at the context.

(7) "<Tháinig>"
"tar" Verb PastInd Len
"<an>"
"an" Art Sg Def
"<bháisteach>"
"báisteach" Noun Fem Com Sg Def

CG operates at sentence level, whereby a sentence is described in terms of cohorts, readings and tags. Each token in the sentence has a cohort, which consists of all the possible readings (morphological analyses) for that token. Each reading consists of tags, which include the lemma and morphological tags and grammatical function tags if present. Example (8) shows a cohort for the token *dá* with three possible readings, which include conjunction and preposition.

(8) "<dá>"
"dá" Conj Subord
"do" Prep Poss
"de" Prep Poss

CG has two basic types of rule; 'select' and 'remove' (see Tapanainen, (1996) for details). The input is disambiguated by either *selecting* one reading from a cohort based on the context to the left and/or right of the token or by *removing* impossible readings based on the context. Example (9a) shows a rule where the noun reading is selected if the previous token is an article, and in (9b) the verb reading is removed if the previous token is an article.

(9) a. `SELECT (Noun) IF (-1C (Art));`
b. `REMOVE (Verb) IF (-1C (Art));`

The rule syntax is straightforward to encode. When specifying a rule the correct order of individual tags in a reading is important but not every tag must specified, e.g. "Noun Sg" will include nouns of all genders and cases even though these tags are interleaved between `Noun` and `Sg` , e.g. "`báisteach`" `Noun Fem Com Sg Len`

(10) `REMOVE (Noun Sg) IF (-1C (Art Pl));`

There are several tokens in Irish such as *don* "to the" which is a contraction of the preposition *do* "to" and the article *an* "the". The morphological analyser tags *don* as *"do" Prep Art*. A separate rule does not need to be written to handle these cases as (9a) above will suffice.

This means that CG rules are independent of many changes to the morphological analysis module, e.g. new tags can be introduced and as long as the sequence of existing tags is maintained the CG rules will be unaffected.

5.2.3. XML formatting of linguistic markup

The disambiguated output is then converted to XCES format using the word tags <w> and the *ctag* and *base* attribute/value pairs.

The following shows the XCES markup for our earlier snippet of text in (7):

(11) <w ctag = "Vmip" base = "tar">Tháinig</w>
<w ctag = "Td" base = "an">an</w>
<w ctag = "Ncfsc" base = "báisteach">bháisteach</w>

5.2.4. Grammatical Relations

Grammatical relations are specified using the CQP query language developed at the University of Stuttgart (Schulze and Christ 1994). This is an extended regular expression formalism, which supports regular expressions both at the level of the character and at the level of the word. Associated with each word there may be additional fields of information (for example, the lemma and the part-of-speech tag) and these can be accessed in Boolean combinations with the wordform.

Complex queries can be built from simpler ones, by first assigning names to simple expressions and then using these names to build more sophisticated ones: for this we use the m4 definition language. For example in (12) "any_noun" is defined as the set of tags starting with N and followed by at least 1 and up to 6 characters. In (13) verb forms which are inflected for person and number are characterised as having tags starting with V followed by 3 characters and having a person indicator 1, 2 or 3 in the fifth position (and, optionally, contain up to 4 more characters).

(12) define('any_noun', '"N.{1,6}"')
(13) define('verb_incl_subj', '"V.?.?.?.[1-3].{0,4}"')

Irish has verb-subject-object (VSO) word order and adjectives follow nouns. The following is an example of grammatical relation for expressing the relation object of verb.

=object
1:verb_incl_subj any_adv{0,1} 2:np

Here, the first argument of the grammatical relation called *object* is the item prefixed by "1:" and the second if the item prefixed by "2:". The main line of the definition then reads: "wherever we find a *verb_incl_subj*, followed by 0 or 1 *any_adv*s and then an *np*, we have identified a grammatical relation of type object, first argument *verb_incl_*subj and second argument *np*.

6 Project team and resources

Developing such a substantial set of resources requires a range of talents. We list here the different roles, with a very brief note of responsibilities and, as a guideline to others planning comparable projects, the total amount of time spent on the project.

Role	responsible for	approx time spent
Corpus Development Manager	identifying and acquiring texts and permissions; bibliographic data.	9 person-months
Corpus Processing Manager	General; Irish linguistic tools	9 person-months
Infogistics Ltd: web specialists	Collecting and encoding web corpus	6 person-months
Senior Irish linguist	Reviewing Irish web data and linguistic tools	1 person-month
Student interns; corpus 'cleaners'	Manual text cleanup, header input	18 person-months
Systems administrator	Intranet, web database etc	0.5 person –months
Computational Linguist	Corpus encoding	3 person months

Michael Rundell was in overall charge of design and collection issues, while Adam Kilgarriff oversaw the text-processing and encoding operations. This represented a total of around six person-months of management input.

7 Further plans

As currently configured, the NCI is a well-balanced and well-annotated corpus, representing a wide range of text-types, and we believe it will form an excellent basis both for the English-Irish dictionary and for Foras na Gaeilge's longer-term publishing programme. There are, additionally, a number of opportunities for further enhancing these resources in the coming months and years, in terms both of data and linguistic annotation. These include:

- "classic" literary sources: a significant number of books by important and highly-regarded Irish-language writers do not currently exist in electronic form (having been published mainly during the first half of the 20th century): a scanning programme to capture this body of literature would add valuable new data to the NCI.
- untapped spoken data: Ireland is blessed with large archives of recorded speech dating back over 70 years but, to date, very little of this material has been transcribed. One such archive alone, that of Raidió na Gaeltachta, has many hundreds of hours of recordings. This represents a very valuable linguistic (and cultural) resource, which it would be desirable to add to the NCI.

- improved linguistic tools for Irish: the time available in the current project for developing and refining the Irish linguistic tools was necessarily limited. While current performance figures are satisfactory for lexicographic purposes, they could be further improved. We hope that resources will be made available to do this, and that any improvements will be fed back into the NCI through re-lemmatizing and POS-tagging the Irish data with improved tools

There is planned to be a new Irish-to-English dictionary in due course, and we would hope that that project would be associated with a re-examination of corpus requirements. Extensive coverage of Irish literature is of limited significance to an English-to-Irish dictionary, but would play an important role in the analysis of the Irish language required for an Irish-to-English one.

8 Conclusion

The project has successfully gathered a high-quality corpus of substantial size from a wide range of sources, in just over a year and with modest resources. The corpus was designed primarily to meet the lexicographic requirements of an English-to-Irish dictionary, but with an eye to the resource being used more widely, by scholars of Irish and Hiberno-English. Three routes were followed for collecting data: (1) using data from existing corpora, (2) approaching copyright holders, and (3) harvesting the web. Each raised assorted issues, and each plays an important role in the resulting corpus. The most innovative was the use of the web, which we have described in some detail.

We established and implemented policies for data encoding, and in this paper we address in some detail questions such as

- which parts of web pages and newspapers should be retained?
- how should duplication be addressed?
- how should the constituent documents be organized?

We have shown how the encoding of the corpus feeds into lexicography. Lexicographers are best supported by a linguistically-aware corpus query tool, and that will require a linguistically-annotated corpus. Such tools are readily available for English, but were not, at the outset of the project, for Irish, so, we developed and extended tools for the morphological analysis and part-of-speech tagging of Irish within the project: we would encourage others, when working with a language where tools are currently limited in scope or non-existent, to do likewise.

We believe that many of the procedures outlined here can be applied in order to rapidly and inexpensively gather corpora for other smaller languages.

Corpus access

All enquiries regarding access to the corpus should be addressed to Foras na Gaeilge, 6 Merrion Square, Dublin 2, Ireland.

Acknowledgements

In addition to the authors, the main corpus-development team comprised Steve Finch, Eamon Kegan, Eoghan Mac Aogáin, Mark McLauchlan, Lisa Nic Shea, Jo O'Donoghue, Paul Atkins, Pavel Rychly and Dan Xu, all of whom deserve our heartfelt gratitude. We would also like to thank Seosamh Ó Murchú, Foras na Gaeilge's Project Manager for the NEID, for his supportive role; Josef van Genabith of Dublin City University, for arranging the student internships; Dónall Ó Riagáin for helpful advice at the corpus design stage; John Kirk of the Queen's University, Belfast, for permission to use NICTS; and Anne O'Keefe and Fiona Farr of the University of Limerick, for permission to use the Limerick Corpus of Irish English.

References

An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta, 2003. *Plean Straitéise 2003-2004* p49-50

Atkins, B. T. S. (2002). Then and now: competence and performance in 35 years of lexicography. In Braasch and Povlsen (Eds.) *Proceedings of the Tenth Euralex Congress*, University of Copenhagen, Denmark: 1-28.

Atkins, B. T. S., Clear, J. H., and Ostler, N. (1992). Corpus design criteria. *Journal of Literary and Linguistic Computing*: 1–16.

Beesley K. & Karttunen L., (2003). *Finite State Morphology*. CSLI Publications: California.

Census of Ireland, 2002. *Volume 11 Irish Language*. Tables 7A and 31A http://www.cso.ie/

Chanod J-P., & Tapanainen, P., (1995a). *Tagging French – comparing a statistical and a constraint-based method.* In Proceedings of Seventh Conference of EACL, Dublin 1995, p149-156.

Chanod J-P., & Tapanainen, P., (1995b). *Creating a tagset, lexicon and guesser for a French tagger*. In Proceedings of ACL SIGDAT workshop From Texts to Tags: Issues In Multilingual Language Analysis, Dublin 1995, p58-64.

de Bhaldraithe, T. 1959. *English–Irish Dictionary* Baile Átha Cliath: An Gúm.

Grefenstette, Gregory and Julien Nioche. (2000). Estimation of English and non-English Language Use on the WWW. Proc. RIAO (Recherche d'Informations Assistee par Ordinateur), Paris.

Grefenstette G., Schiller A., & Ait-Mokhtar S., (2000). *Recognizing Lexical Patterns in Text*, In: van Eynde, F. & D. Gibbon: Lexicon Development for Speech and Language Processing. Dordrecht: Kluwer Academic Publishers

Janes, A, (2004). *Bilingual comparable corpora for bilingual lexicography.* MSc Dissertation, University of Brighton.

Johnson, S. (1747). *The Plan of an English Dictionary*.

Jones, R. and R. Ghani. (2000). Automatically building a corpus for a minority language from the web. 38th Meeting of the ACL, Proceedings of the Student Research Workshop. Hong Kong. Pp 29-36.

Karttunen L. & Beesley K. (1992). *Two-Level Rule Compiler*. Technical report, Xerox PARC.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine, in Williams and Vessier (Eds.) *Proceedings of the Eleventh Euralex Congress*, UBS Lorient, France: 105-116.

Kilgarriff, A. and Grefenstette, G. (2003) Web as Corpus: Introduction to the Special Issue. *Computational Linguistics* 29 (3) 333-347.

Schulze, B. and O. Christ (1994). *The IMS Corpus Workbench*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Tapanainen, P. (1996). *The Constraint Grammar Parser CG-2*. Publication No. 27, University of Helsinki.

Trench, R. Chenevix. (1857). *On some deficiencies in our English dictionaries.* London: The Philological Society. (reprinted at http://www.oed.com/archive/paper-deficiencies/)

Uí Dhonnchadha, E. (2002). *An Analyser and Generator for Irish Inflectional Morphology using Finite State Transducers*. Unpublished MSc Thesis: Dublin, DCU.

Uí Dhonnchadha, E., Nic Pháidín, C. Van Genabith, J. (2005 forthcoming). *Design, Implementation and Evaluation of an Inflectional Morphology Finite-State Transducer for Irish*. In: MT Journal - Special Issue on Finite State Language Resources and Language Processing. Kluwer