

Task H Deliverable
Design Principles for
the New Corpus for Ireland (NCI)
Version 2, 20th March 2004

1. INTRODUCTION	2
TASK	2
DELIVERABLES	2
BACKGROUND: THE ROLE OF THE CORPUS IN CONTEMPORARY LEXICOGRAPHY	2
GENERAL PRINCIPLES	3
2. DATA COLLECTION ISSUES.....	3
SOURCES	3
SAMPLING METHODS	4
COPYRIGHT AND PERMISSIONS	4
3. CLASSIFYING CORPUS TEXTS.....	4
IDENTIFYING TEXTUAL FEATURES	5
PROPOSED ATTRIBUTES AND VALUES.....	5
CLASSIFYING SPOKEN TEXT.....	9
SUMMARY ON TEXT CLASSIFICATION.....	9
4. ACHIEVING BALANCE.....	9
MONITORING CONTENT AS THE CORPUS DEVELOPS	9
TARGET PERCENTAGES.....	10
WHAT IF WE COLLECT ‘TOO MUCH’ TEXT?	11
5. FUTURE DEVELOPMENTS.....	11
TRANSCRIBING RECORDED SPOKEN DATA.....	12
PLANS FOR DIGITISING WRITTEN TEXTS	12

1. Introduction

Task

To design a databank of linguistic evidence (on which the dictionary will be based) appropriate to the dictionary envisaged.

Deliverables

LexMC is contracted to deliver to Foras na Gaeilge the following data, which will collectively make up the New Corpus for Ireland:

- a 30-million-word Corpus of Irish
- a 25-million-word Corpus of Irish English
- a 200-million-word Corpus of British and American English
- a 'reading-and-marking' programme that will deliver benefits during the entire course of the project (Phase 1 and Phase 2)

Every text included in the NCI (with the exception of web-derived texts: see 'Copyright and permissions' in Section 2) will be accompanied by permission from the relevant copyright owner to use the data in the compilation of its English–Irish Dictionary and in any subsequent dictionary projects that Foras na Gaeilge may undertake.

Background: the role of the corpus in contemporary lexicography

In the 21st century, there is a general expectation that any dictionary – and especially one that has English as one of its languages – should be systematically based on the linguistic data supplied by a corpus. A corpus is a collection of texts in digital form, and thus constitutes 'primary' data: it shows language in use in real communicative situations, rather than giving us other people's opinions about how words or phrases are used (or 'should' be used). A good corpus is a small-scale model of the language of a given speech community. The evidence it provides forms the basis for generalisations about the way that units of language function, and these generalisations are the raw materials from which dictionary entries are composed. The better the corpus, the more reliable the generalisations, and hence the better the dictionary.

But what makes a corpus 'good'? Essentially, the answers relate to its *size* and its *content*. When corpus users – in this case, lexicographers – are describing a particular linguistic entity (such as a word, a phrase, a meaning, or a grammatical pattern), they need to see sufficient instances of that entity to give them confidence that they can describe its use accurately. Furthermore, they need to be able to say whether a given entity is typical of the language in general, or characteristic of one particular user-group (such as members of the legal profession, or speakers from a specific age-group or region). Both factors point to the requirement for a corpus that is large (i.e. one that includes a large quantity of *text*) and diverse (i.e. one that includes a wide range of *text-types*).

A large and diverse corpus supports every aspect of the lexicographic process, including:

- decisions about what to include in the dictionary and what to leave out
- the structure of complex entries (for example, the order in which the different senses of a polysemous word are shown)
- the description of meaning – denotative, connotative, and pragmatic
- the explanation of syntactic and collocational behaviour
- the handling of phraseology and multiword expressions of all kinds
- information about a word's 'distributional' tendencies (such as an observable preference for occurring in newspapers or in literary registers)
- the provision of material for illustrative examples

We are confident that the plan outlined here will enable us to build a set of first-class corpus resources. These will be harnessed to state-of-the-art lexical software, providing this project with linguistic tools that are more powerful than any currently available to any dictionary project in the world.

General principles

The model proposed here embodies the application of well-established corpus design principles (as used, for example, in the development of the Longman-Lancaster Corpus and the British National Corpus), with appropriate adjustments to take account of factors special to the Irish situation. (These factors are explained more fully below, especially in Section 4, under 'Target Percentages'.) The 'ideal' corpus is a fully representative sample of the textual output of a given speech community. In reality, however, all corpus-building entails compromises between principle and opportunism – between what we would like to collect if resources were unlimited, and what is practicable and achievable in the real world. Furthermore, corpus developers have tended, in the last ten years or so, to avoid the word 'representative' (and the dubious claims that it implies) in favour of the word 'balanced'. A balanced corpus is one that aims to illustrate – in adequate quantities – the full repertoire of text types in which a language is regularly used (with the exception of the terminology used in highly specialized domains), and as such this is a more realistic goal.

We will start, therefore, by outlining a model that embodies our aspirations for an optimally constructed, well-balanced corpus. As far as possible, this design will shape the text-collection process. But we recognise the need to be flexible, to take advantage of what is available, and to be pragmatic about what is realistically achievable.

2. Data collection issues

Sources

Data for the NCI will be collected in three main ways:

- **using existing corpora:** the NCI will incorporate data from corpus collections developed independently of the present project. These include the British National Corpus (BNC) and texts from the Linguistic Data Consortium (LDC), which will, jointly, be the source of our 200-million-word Corpus of British and American English. No more will be said in this document about these two established corpora, and all subsequent references to the NCI exclude the BNC and LDC data. Aside from these two important sources, by far the most significant existing resource is the *Corpas Náisiúnta na Gaeilge* (CNG), created in the late 1990s by ITÉ, and other texts acquired (and in some cases processed) by ITÉ right up to 2003. But we also hope to include – in some form or other – data from other corpora, including the ICE-Ireland Corpus and the Limerick Corpus of Irish English, and from digital text archives such as those of the Royal Irish Academy.
- **getting data from the web:** the natural language processing (NLP) community increasingly sees the web as a major source of corpus data. In fact, the current edition of the key NLP journal, *Computational Linguistics* (Vol. 29/3, September 2003) is a special issue devoted to the subject of 'The Web as Corpus' and guest-edited by Adam Kilgarriff and Gregory Grefenstette. LexMC has a subcontract with Infogistics Ltd, an NLP company that has pioneered methodologies for automating the collection of internet text for corpus creation. One of the advantages of this approach is that it gives us access to significant text types that would be difficult or impossible to collect by more conventional means, such as discussion lists, bulletin boards, and institutional websites.
- **using standard corpus collection procedures:** since the Brown Corpus was developed in the early 1960s, strategies for gathering corpus text from published and unpublished sources have been progressively refined. These are well documented (for example in Atkins, Clear, and Ostler 1991, and on the BNC website), and essentially involve: (a) establishing general design principles; (b) creating a multi-dimensional matrix of required text-types; (c) identifying specific texts that match these requirements; (d) approaching the owners of the rights in these texts and getting permission to include them; and (e)

acquiring the text in digital form (or if necessary converting it into digital form). The key personnel here are the Corpus Development Officer (CDO), Jo O'Donoghue, and the Corpus Development Administrator (CDA), Paul Atkins. The primary role of the CDO is to identify appropriate texts for the corpus, and to establish contacts with copyright holders and negotiate permissions. The CDA's job is to acquire the text physically, record all relevant textual features (as described in Section 3), and make the text available to the Corpus Processing Manager for incorporation in the corpus. See also Section 4 on proposals for monitoring the content of the corpus as it develops.

Sampling methods

In order to minimise the incidence of 'skewing' (the phenomenon whereby excessive quantities of a single text or type distort the balance of the corpus and reduce the reliability of corpus-derived frequency data), we propose to restrict the size of any individual text to a maximum of 60,000 words. Where a source text is longer than this, a 60,000-word sample will be taken. As a general principle, we would not expect to collect more than one text per author, but we should be prepared to bend this rule in cases where an author's output spans more than one genre, or where the writer in question is especially significant and influential. There is of course no question of *excluding* writers on the grounds that they are not perceived as 'significant' or 'influential'.

Some linguists have argued that texts should not be 'chopped up' in this way, because there are significant variations in discourse structure in different sections of a text. This is unlikely to be a problem for the main users of the corpus (lexicographers), but in order to address these concerns, we plan – whenever feasible – to take samples from the beginning, middle and end of longer texts, in an alternating cycle.

Copyright and permissions

Given that we are collecting very little material dating from before the 20th century, it follows that most of the texts we wish to include in the NCI will be protected by copyright law. We therefore need to ensure that appropriate permissions are obtained from copyright owners for all the constituent texts in the corpus.

For texts collected by LexMC, we will request that donors sign a licence agreement allowing their text to be incorporated into the NCI, which will then be used as linguistic data in the compilation of:

- the present (English–Irish) dictionary
- any subsequent dictionary produced by Foras na Gaeilge

Additionally, this licence gives Foras na Gaeilge the right to sublicense the NCI to third parties for research in lexicography, linguistics, or speech and language technology.

For texts sourced from existing corpora, we will aim whenever possible to 'upgrade' the permissions level so that it conforms to the licence conditions described above. A high proportion of ITÉ-sourced texts already have an appropriate level of permission: donors were required to sign a letter which explicitly mentioned the possible use of their data in future dictionary projects. (We have copies of all the relevant letters, and these will eventually be forwarded to Foras na Gaeilge.)

For texts sourced from the web by Infogistics, the usual copyright issues do not apply. This is a legal grey area but we will keep abreast of any developments.

3. Classifying corpus texts

Identifying textual features

For each of the (non-web-derived) texts and text samples that make up the NCI, we aim to record full bibliographic details, including the date and place of publication, and the name, age, and gender of the author or authors. Beyond this basic information, individual texts will be classified using a system of 'attributes and values'. For example, the attribute 'Medium' indicates the medium in which the text is made available to readers or listeners, and this attribute has a range of possible values, including 'book', 'newspaper', and 'website'. The attributes and values described here enable us to 'define' any corpus text in quite a fine-grained way; this in turn will enable corpus users to see how words behave in different text types and to study the characteristics of different genres. They represent our current view of how the constituent texts of the NCI will be classified and, hence, how the overall balance of the corpus will be monitored. These parameters are based on our corpus-building experience, our reading of the relevant literature, and our discussions with key members of the project. Nothing here is final, however: it may be necessary to make adjustments after consultation with advisers, and for some attributes (notably 'Time'), final decisions about the appropriate values will only be made after further discussion with Foras na Gaeilge.

Proposed attributes and values

These are the main attributes we propose to use

- language
- time
- mode
- medium
- genre and domain
- target readership

The following sections describe them in more detail:

Attribute: Language

The language or dialect in which the text is written (or spoken).

Values

The primary values for this attribute are either 'Irish' or 'Irish-English'.

For texts in the Irish component of the NCI, we will aim to record the following additional information:

- whether the writer/speaker is a native-speaker of Irish.
- whether the text represents one of the three main dialects of Irish
- which county in Ireland the writer/speaker comes from, or – when the data is available – even more precise information regarding an author's origins
- whether the text is an original Irish text or a translation from any other language

For texts in the Irish-English component of the NCI, we will aim to record the following additional information:

- which county in Ireland the writer/speaker comes from
- whether the text is an original Irish-English text or a translation from Irish

All such data will be collected when it is readily available, and recorded when we can be confident of its reliability. Values at the more fine-grained end of the scale will not always be easy to establish, and (in line with normal corpus-collection practice) we will use the 'U' (=unknown/unclassified) tag in cases where the relevant information proves elusive.

Proposals

Our objective is to collect at least 30 million words of Irish and at least 25 million words of Irish-English. We do not propose to set specific targets for the other values of the Language attribute, but they will be monitored carefully to ensure the best achievable balance of

variables. In the case of Irish texts, we recognise that there is a somewhat anomalous situation regarding native-speaker status and its implications for the 'quality' of the text we collect. Much of the available material is produced by writers whose first language is not Irish, and whose Irish will sometimes lack what one commentator has called 'the idiomatic richness and elegance of speech which one might otherwise expect'. Since the data in the corpus itself may well eventually feed back into pedagogical materials, this issue needs careful handling. Our response to this is twofold: first (as noted above) texts will be appropriately tagged so that corpus users can distinguish native and non-native text; and secondly, we will ensure that high-quality native-speaker texts from the whole of the 20th century are well represented in the corpus.

Sources that have been or will be approached in order to gather native-speaker Irish texts include: An Clóchomhar, Sairséal agus Dill (one of the best sources for mid-to-late 20th century material; Caoimhín Ó Marcaigh has already given permission in principle), Mercier (permissions acquired already), An Gúm (which has some native-speaker texts not in the ITÉ holdings), the Cartlann na gCanúintí archive at UCD, and the RIA

Attribute: Time

The date when the text was written (or – in the case of spoken texts – the date when it was recorded or broadcast).

Values

The two extremes here are illustrated by, on the one hand, the Brown Corpus, which consists only of texts 'printed during the calendar year 1961' (<http://helmer.aksis.uib.no/icame/brown/bcm.html>); and, on the other hand, Brigham Young University's 100-million-word Corpus de Español (www.corpusdelespanol.org), which includes texts covering the entire period from the 13th to the 20th century. The former corpus is strictly synchronic, the latter is fully diachronic. Most corpora fall somewhere between these extremes. In the case of the NCI, one could make a case for a broadly synchronic approach: the task is to design a corpus 'appropriate to the dictionary envisaged', and the user-profile for the current (English–Irish) project implies a very low requirement for detailed treatment of archaic or obsolete vocabulary in English, Irish-English, or Irish. Thus a corpus made up entirely of texts from the late 20th and early 21st centuries would – arguably – meet the present requirements of the editorial team. However, the NCI is intended to provide data for a series of future projects that may well include an Irish–English dictionary (and possibly also an Irish grammar), and projects such as this point to the need for a broader range of values for the Time attribute.

Proposals

We propose to treat the two main corpora somewhat differently. For texts in Irish, we suggest three Time values:

- 1883-1959
- 1960-1999
- 2000-

Our understanding is that the Royal Irish Academy has an extensive electronic archive of Classical and Modern Irish texts, and that a CD-ROM with a rich sample of texts dating from the 17th century to 1882 is to be released early in 2004. There seems little point in the NCI duplicating this, and its value for the present project will be minimal. If a future Foras project creates a need to analyse older texts of this type, we presume that an arrangement with the RIA should make it possible to add an 'earlier Irish' component to the NCI with relatively little difficulty. The period from 1883-1959, however, seems likely to yield significant quantities of the kind of 'rich' Irish text – predominantly written by native-speakers, and much of it still read today, especially in educational contexts – referred to in the previous section. This might include, for example, the work of Pádraic Ó Conaire and Séamus Ó Grianna; the well-regarded Irish translations of European literature published by An Gúm during its early years; and the Constitution of Ireland (Bunreacht na hÉireann) of 1937. We have already collected (and been given) a great deal of information about what is available, and we will aim to collect

at least 5 million words from this period. For some text types, there is – arguably – less value in collecting data from earlier periods: older journalism and popular non-fiction are by their nature more ephemeral, while domain-specific academic writing will often be outdated in its content. (There is a precedent here in the design of the Longman-Lancaster Corpus: its diachronic range was from 1900 to 1990, but only fictional texts were collected for the period from 1900 to 1950.) In most cases, therefore, texts from 1883-1959 will contribute to the fictional and (auto)biographical components of the corpus.

With regard to the Irish-English component of the NCI, we see its primary value to the current project as a data-source for contrastively analysing the characteristics of this variety in relation to other dialects of English, especially British English (the variety to which it is – or is assumed to be – most closely related). A corpus designed for underpinning a *diachronic* investigation of Irish-English would have a substantially different make-up. Consequently, we propose for the Irish-English corpus a start date of the founding of the Free State in 1922. Efforts will be made to identify and collect historically or culturally important texts from this point onwards, but we expect the majority of Irish-English texts to be, broadly speaking, contemporary.

Attribute: Mode

The mode of delivery: specifically, whether the text is written or spoken.

Values

There are only two possible values here: spoken and written. For present purposes, we define spoken text quite narrowly, as unscripted interactions in contexts such as meetings, seminars, chat shows, recorded reminiscences, and above all conversations. Fictional or dramatic dialogue, prepared monologues (speeches, lectures etc.), and the emerging text type of ‘chatroom’ interactions will all be regarded as forms of written text in our classification.

Proposals

Good spoken data is difficult to collect. Quite apart from the problem of identifying the target population, the physical process of recording, then transcribing, the data is immensely time-consuming and therefore extremely expensive. (The 10-million-word spoken component of the BNC cost more to collect than its 90 million words of written text.) Speech recognition software is still some way from being able to cope with spontaneous dialogue (given the challenges of ambient noise, overlaps between speakers, and the apparently anarchic syntax of spoken text). Consequently, many large corpora avoid the issue altogether and focus only on written data, and such spoken corpora as exist tend to be much smaller than their written counterparts.

Taking all this into account, and given the current budget and schedule, we propose the following approach to spoken data:

- no *new* spoken text will be collected during Phase 1 of the project
- efforts will be made to acquire text from *existing* spoken corpora
- investigations will be made into the feasibility of transcribing existing recorded text in various archives, with a view to making a costed proposal for Phase 2 of the dictionary project (see Section 5)

Existing spoken corpora include the spoken component of the ICE-Ireland corpus and the recently completed Limerick Corpus of Irish English – both high-quality sources. We are in negotiations with both and are confident that this data will be made available for lexicographic analysis. This will be of great benefit for the project. At the same time, investigations are beginning into the feasibility of transcribing recorded speech held by institutions such as the UCD folklore department and the Raidió na Gaeltachta sound archive. (There is plenty of appropriate material, but the likely costs of transcription are not yet known.)

In light of all these factors, we propose that the NCI is (at this stage) seen as an essentially written corpus, and we plan to base our calculations regarding its content and balance on the volume of written text we collect. We will actively pursue every possible source of spoken

data, but it would be better to view any data that we acquire as complementary (rather than integral) to the NCI. It is, in any case, unlikely, that the owners of laboriously-gathered, pre-existing spoken corpora would be willing to see their data fully absorbed into the NCI.

Attribute: Medium

The medium in which the text appears

Values

There is a wide range of values for this attribute. Our current list is as follows, and more values will be added if necessary:

book	academic journal	ephemera	interview
newspaper	website	broadcast-radio	talk/lecture
magazine	dissertation	broadcast-TV	seminar/meeting
periodical	official publication	conversation	

Proposals

All texts collected for LexMC by the CDO and the CDA will be classified according to these values. All texts sourced from the internet by Infogistics will be classified with the value 'website'. Texts from pre-existing corpora will be classified using these values as far as is feasible.

Attributes: Genre and domain

The specific type of text, or the specific domain or topic that it deals with

Values

None of the existing approaches to these attributes is entirely satisfactory, partly because there is no robust, generally-agreed definition of 'genre', and partly because texts such as novels are inherently different from texts that are 'about' things. Consequently, a one-size-fits-all classification scheme will not always work.

The objective is for us to be able to categorise any individual text fairly precisely, and after considerable discussion we have decided on the following approach:

- top level: a distinction between *imaginative* texts (roughly speaking, creative writing) and *informative* texts (roughly speaking, non-fiction)
- middle level: a set of broad categories for each of these two types. For imaginative texts, the values are more 'genre-like', and include fiction, poetry, and drama. For informative texts, the values are more 'domain-like', and include hard & applied sciences, politics & current affairs, business & finance, and leisure & popular culture
- bottom level: a set of quite specific categories such as (for imaginative texts) short story, soap opera, and literary novel, and (for informative texts) women's studies, rural history, and Irish music

Proposals

All texts, from whatever source, will be classified at the top level. The web-sourced text from Infogistics cannot practically be categorised beyond this level. Most texts collected by LexMC will be classified at the middle level, but this will not be done for texts from journalistic sources. Texts from ITÉ currently have their own classification, which does not map exactly onto the categories we will be using for the NCI. Nevertheless, most of the information we need is already available in ITÉ text headers and in other ITÉ records to which we now have access. We will therefore aim to classify these texts to the middle level also, as far as is feasible.

The most fine-grained categories, at the bottom level, are not generally required for lexicographic purposes (for which relatively broad-brush categories are usually adequate). However, experience with other corpora (notably the BNC) clearly demonstrates the long-term research value of finely categorised data. We will therefore do a pilot study to assess the

feasibility of classifying texts (especially those from ITÉ and those that LexMC collects) to this level of detail.

Attribute: Target readership

The type of reader or hearer that the text is aimed at

Values

Many texts (perhaps a majority) are not aimed at one specific group. For this attribute, therefore, the default value is 'general'. Nevertheless, there are plenty of texts that do have a clear target readership (such as teenagers, schoolchildren, or adult learners), and if such information is readily available, it makes sense to record it.

Proposals

Texts collected by LexMC and texts from the CNG will – as far as is practicable – be classified according to this attribute. Texts from the web are assumed to fall into the default category, unless we have specific information suggesting otherwise.

Classifying spoken text

As noted above, spoken data will be treated as a separate issue. Many of the attributes by which we classify written data can be applied equally to spoken text (such as Language, Time, Medium, and of course Mode). Information about genre and domain is often less relevant, however. Conversely, information about *context* needs to be recorded, and we have established a provisional set of categories to enable us to do this, including 'intimate' (normal conversation), professional (for example, business meetings), oral history, and pedagogical (for example, seminars or school lessons). It is too soon to be clear whether this is a useful set of categories, or whether classifying spoken documents in this way is practically achievable. In the case of any Irish spoken text that is collected, the header documents should differentiate between native and non-native speakers whenever it is possible to do so reliably.

Summary on text classification

It is a general rule of corpus development that the long-term value of the data increases in direct proportion to the 'granularity' of the system used for classifying and annotating it. In principle, therefore, our aim is to categorise all the constituent texts of the NCI as precisely as possible. The information we record about each text will form the basis for the 'document headers' attached to the text, and this will have benefits both for lexicographers (for example, by supporting decisions about labelling) and for researchers (for example, by facilitating the creation of specialised subcorpora). However, some variation in the level of text-classification detail is inevitable, owing partly to the intrinsic nature of some text-types and partly to the range of sources from which we will be gathering our data.

4. Achieving balance

Monitoring content as the corpus develops

The system of attributes and values described above enables us to track very precisely the relative proportions of all the different text-types that the corpus includes. For each of the two main corpora (Irish and Irish-English), a complex Excel spreadsheet will be used for recording the relevant features of every individual text. For monitoring purposes, however, we propose to use a more coarse-grained approach that focuses on a small number of key features, as follows:

- mode: texts will be classified as 'written' or 'spoken'
- medium: texts will be classified as 'book', 'newspaper/periodical', 'website', or 'other'
- genre: texts will be classified as 'imaginative' or 'informative'

- language (Irish corpus only): texts will be classified as ‘native-speaker’ or ‘nonnative-speaker’

These features will be recorded in a Summary sheet, enabling us to get an immediate fix on the balance of the corpus as it develops (and to take appropriate action if any of these key features is either over- or under-represented). It is important to note that it will still be possible to extract more sophisticated counts from the data in the spreadsheets (such as: ‘how much Irish-English fiction by women writers in the period 1960-1999 does the corpus contain?’): the process in these cases would simply be a little more time-consuming.

Target percentages

For the key measures described above we will establish target percentages, and then identify and collect texts so that the final composition of the corpus reflects these targets. This is not an exact science, however, because:

- there are no generally-accepted ‘rules’ about corpus composition, and there is considerable variation in the design of all the major corpora that we know of
- the balance of text-types in a corpus should take account of the needs of the users of the language-products which the corpus will underpin; it thus relates – to some degree – to the user profile for the dictionary (see Task A Deliverable), though attention must also be paid to the requirements of possible future users of the NCI, such as translators, interpreters, theoretical linguists, language teachers and lexicographers working on future dictionaries, including most notably an Irish-English counterpart to the NEID.
- some text-types have special importance within a given speech community: for example, ‘reminiscences’ of various types are a significant feature of the Irish situation (but less significant in British or American English), and corpus developers need to be sensitive to these issues. For the Irish component, the importance of including quality native-speaker texts from the first half of the 20th century has also been discussed (p.5, above).
- a degree of flexibility is always needed when collecting data, because there will inevitably be some differences between the texts that one ideally wants and the texts that are readily available

We propose the following target percentages for our key measures:

- mode: no targets. As stated above, the NCI will be seen essentially as a written corpus, with any spoken data we collect being complementary
- medium: book: 50%; newspaper/periodical 20%; website (Infogistics data): 25%; other (including government/official): 5%
- genre: for Irish: imaginative 50%, informative 50%; for Irish-English: imaginative 40%, informative 60%

Summary of current targets: It should be stressed that these targets are provisional and approximate only: they are subject to adjustment *both* in the light of comments on this document *and* in response to what is actually available in the various categories. Some of the web-derived (Infogistics) data has been slotted into other categories (such as news and official). Some of the imaginative component is subsumed in the broadcast and website categories. Spoken data (unless broadcast) is not included in these counts for reasons given above (p.7: Attribute: Mode)

category	target Irish		target Irish-English	
Books: imaginative	30%	9 m	30%	7.5 m
Books: informative and instructional	20%	6 m	20%	5 m
Newspapers (including news on the web)	15%	4.5 m	15%	3.75
Periodicals/magazines: imaginative	3%	1 m	3%	0.75 m
Periodicals/magazines: informative and instructional	5%	1.5 m	6%	1.5 m
<i>(Newspapers/periodicals total)</i>	<i>23%</i>	<i>7 m</i>	<i>24%</i>	<i>6 m)</i>
Official and government (includes legislation, EU docs (in Irish), Dáil debates, public inquiry proceedings etc	5%	1.5 m	4%	1 m
Broadcast	3%	1 m	3%	0.75 m
Website	18 %	5.5 m	19%	4.75 m
Totals	100%	30 m	100%	25 m.

What if we collect ‘too much’ text?

It is possible, indeed probable, that we will be able to collect more text than the task actually requires. This is especially likely in the case of the Irish corpus, because ITÉ alone is making available up to 20 million words of text, and Infogistics will supply at least a further 10 million words. Should this situation arise, we propose that the two main components (30 million words of Irish, 25 million words of Irish-English) be seen as the ‘core’ corpora: for these, we aim to achieve the best possible balance of text-types, and the highest possible level of permissions, ‘header information’, and linguistic annotation. Any additional material would then be archived (with minimal annotation) for possible future use, and Phase 2 of the project could include provision for upgrading the levels of permission and documentation to conform with the standards of the core corpora.

5. Future developments

This document describes our plans for developing corpora during Phase 1 of the English–Irish dictionary project. During this phase, a number of other avenues will be (briefly) explored, with a view to possible further corpus-development activity during Phase 2. These include:

☐ **Transcribing recorded spoken data**

Ireland is blessed with large archives of recorded speech dating back over 70 years, and this represents an extremely valuable linguistic (and cultural) resource. Very little of this material has been digitally transcribed, however, and it would clearly be of great benefit for the study of Irish history and culture (as well as language, of course) if a transcription programme could be put in place. Investigations are in progress regarding material in the Raidió na Gaeltachta sound archive, led by the project's Chief Irish Editor (Dónall Ó Baoill) and by Peadar Mac an Iomaire of NUI Galway. We hope these will lead to a costed proposal for transcribing some sound data later in the project

☐ **Plans for digitising written texts**

There are a number of significant texts (including both originals and Irish translations), that were published by An Gúm in the early-to-mid-20th century, and that currently exist only in printed form. An Gúm plans a pilot project to assess the likely costs of either scanning this material or having it keyboarded

LexMC's directors would like to acknowledge the valuable contribution made to this document by members of the consortium, notably Dónall Ó Baoill, Jo O Donoghue, Elaine Uí Dhonnchadha, and Eoghan Mac Aogáin. We also thank Dónall Ó Riagáin, who wrote a detailed and extremely useful report on an earlier draft. Any errors in the final document, however, are LexMC's responsibility.