Covering All Bases: Regional Marking of Material in the New English-Irish Dictionary

Cathal Convery, Pádraig Ó Mianáin, Muiris Ó Raghallaigh
Foras na Gaeilge, Dublin

The New English-Irish Dictionary is a government-sponsored project that began in 2000 and is due for completion in 2012. The aim is to produce a modern bilingual dictionary containing c. 40,000 headwords which is to be published in both printed and electronic formats. When published this dictionary will be the first major dictionary published for Irish in over 40 years. The project is currently at the translation phase, and this paper focuses on the approach taken to attempt cover dialect variations in the modern spoken language. The methodology employed was divide the headword list into three distinct categories, each requiring a different level of translation.

Given the time and budgetary constraints of the project it was decided that only the 1000 (approx) most frequently occurring lemmas could receive a full dialectal profile. Translators from each of the three main dialects translate each entry, passing the entry on to a translator from the next dialect as they complete their part of the process. This translation work is carried out without reference to written sources. Once a translator from each main dialect has completed their work the entry is checked for completeness against set sources and labelled accordingly.

The main advantages of this process are as follows.

- *It captures current translations that may not be covered in existing out-dated sources.*
- *It provides a dialectal profile of words, phrases and usages.*
- *It enables an element of dialectal marking in the final product, particularly in the electronic version.*
- *It enables the option to customise the electronic version, fronting any particular dialect.*
- *A given dialect may be selected as the default pronunciation in the electronic version.*
- *It enriches the bilingual database creating a useful research resource for other academic research projects.*

1. Background

Irish (*Gaeilge* or Gaelic) is the indigenous language of Ireland since approximately 500 BC. It is officially recognised under Article 8.1 of the Irish Constitution (1937) as the national language and first official language of Ireland. Irish also gained official recognition in the North of Ireland with the Good Friday Agreement in 1998, while it has also been adopted as a working language in the EU since 2003.

Irish is now a minority language in Ireland, a consequence of centuries as a *lingua non grata* in the eyes of the ruling ascendancy, and its survival to this day is due largely to a vibrant language revival movement towards the end of the 19th century that coincided with the nationalist movement for political autonomy in the early 20th century. When Saortsát Éireann, the Irish Free State, was founded in 1922, the Irish language was seen as the cornerstone of the cultural ideology of the new nationstate, and although the language was already extinct in most of the country, substantial legislative measures were put in place by the new indigenous regime in an attempt to arrest the decline of the language where it was still being spoken and to promote the language throughout the state. The language continued to decline, however, particularly in the regions along the west coast known as 'Gaeltacht' areas where Irish has survived as a living community language, but the outlook became more positive in the latter half of the last century with the latest census figures in 2006 showing over 1.8 million people (out of a population of 5.7 million on the island of Ireland), North and South, professing some knowledge of the language, though only about 60,000 active speakers in the official Gaeltacht areas[1].

---

[1] Central Statistics Office Ireland. Census 2006 Principal Demographic Results. Central Statistics Office Ireland. Census 2006, Amended Principal Demographic Results 2006. 29-30.

With the overthrow of the Gaelic regime in the 17th century, the language lost the stabilising and standardising stewardship of its literary élite, and from this point onwards the defining and distinguishing characteristics of the main dialects begin to come to the fore (Williams 1996: 447- 478). Since the 19th century, the spoken language has largely split into the three major dialects spoken in the Gaeltacht areas in three provinces, Ulster, Connaught and Munster (Ó Giollagáin & Mac Donnacha 2008). With its new status of official state language came an urgent need for the standardisation of Irish, and so in the 1950s the Official Standard of the Spelling and Grammar of Irish was set out (Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil. 1958). For a variety of complex reasons, however, the Official Standard failed to gain widespread acceptance among the native speaker population in any of the main dialects, and the ensuing credibility gap keeps widening as the living language is constantly changing while the Official Standard has remained frozen for the last sixty years (Ó Háinle: 1996: 791-793).

2. Modern Lexicography in Irish

Since Irish was granted official language status, dictionaries have played a key role in the normalisation of the language in the 20th century as they were the main platforms for the standardisation of grammar, spelling and terminology. Given the proportion of non-native speakers among the Irish-speaking population and persisting issues around literacy in the language among the native-speaker population, major bilingual dictionaries have become the 'weapons of mass dissemination' of the language.

There have been three such dictionaries since the foundation of the state: Dinneen's Irish-English Dictionary in 1927, de Bhaldraithe's English-Irish Dictionary in 1959, and Ó Dónaill's Irish-English Dictionary in 1977. Each of these dictionaries were milestones in the rehabilitation and development of Irish as a national language, and though inevitably dated, they remain to this day the essential reference tools of the language.

Unfortunately, work in the field of modern Irish-English bilingual lexicography at this level ground to a halt after the publication of Ó Dónaill's dictionary in 1977. It is now over fifty years since the last major English-Irish dictionary was published and it is universally accepted that the canonical dictionaries of Irish are long out of date.

3. The New English-Irish Dictionary (NEID) project

Provision for the Irish language was included in the Good Friday Agreement between the Irish and British governments in 1998, and subsequently in December 1999 Foras na Gaeilge was founded as part of The Language Body, a cross-border implementation body, with responsibility for the promotion of Irish throughout the island of Ireland. One of the stated functions of Foras na Gaeilge is 'developing terminology and dictionaries' and steps were taken immediately to initiate its first major lexicographical project. The following year, in 2000, work started on NEID, a modern bilingual dictionary containing c. 40,000 headwords to be published in both printed and electronic format, and due for completion at the end of 2012. The entire project is supported by IDM's DPS system (www.idm.fr); its use on this project is intended to facilitate the production of other dictionaries in the future, as well as NEID.

Because the field of bilingual lexicography in Irish had lain dormant for so long, a considerable emphasis was placed on planning not only for the current project but also for future requirements such as new projects or maintenance of completed projects. Consequently, the NEID project was divided into three major phases:

3.1. Phase 1 (Planning and Technical Design)

The contract for this phase was won by [Lexicography MasterClass](). The work was completed in 2006, and included significant elements of the project, such as:

- The New Corpus for Ireland (30 million words of Irish and 25 million words of Hiberno English);
- A draft list of 50,000 English headwords for NEID[2];
- The Document Type Definition (DTD);
- A comprehensive Style Guide.

3.2. Phase 2 (compilation, writing and editing)

Phase 2 began in 2007 and is scheduled to be completed in mid-2012. It involves the compilation, writing and editing of the Dictionary entries. Phase 2 has been divided into three subphases:

- *Phase 2A: The English Entry Frameworks*. The contract for this subphase was won by Lexicography MasterClass, and is on schedule to be completed in July 2010. For each headword, an 'entry framework' is being created; an entry framework is a database entry which records the various meanings and uses of a given word along with relevant grammatical and functional information and examples from the project's 1.7 billion word LexMCI English corpus. The entry frameworks are designed to create a complete profile of each English headword for the translators working in Phase 2b (Rundell, M. Atkins, S, & Kilgarriff, A. 2010).[3]
- *Phase 2B: Translation*. Phase 2B started in September 2009 and is scheduled to be completed by December 2011. It will produce a rich bilingual corpus reflecting current usage and dialectal variations as well as covering existing sources.
- *Phase 2C: Editing*. Phase 2C is scheduled to start in January 2011. In this subphase, material for the final dictionary entries will be selected from the translated entry frameworks and the actual entries will be compiled and edited.

3.3. Phase 3 (Publication)

This phase is scheduled to commence in January 2012, and the dictionary is scheduled to be published by December 2012. This phase will involve the preparation of the printed and electronic versions and the final proofreading.

4. The translation challenge: currency and coverage

Bearing in mind the project's aims to produce a modern comprehensive dictionary, up-to-date in its content in both English and Irish, the issues of currency and dialectal coverage were identified as the foremost challenges facing the NEID project in the translation stage. The absence of a comprehensive corpus of contemporary spoken Irish which would underpin this aim poses significant difficulties, as does the internal dialectal division within the language; the issue is further complicated by the failure of the official standardised language to gain acceptance among the general native speaker population.

---

[2] The original plan was for NEID to contain 50,000 headwords, but this has been scaled back to 40,000 due to budget and time constraints.

[3] See also Atkins & Rundell (2008) chapter 8 for description of the process used. 261-309.

4.1. Currency

NEID has two main target groups. The obvious target group would be learners or non-native speakers of Irish who want to know how to express themselves in correct, 'natural' Irish. Learners in general tend to be more committed than the average native speaker to learning terminology, so their main requirement would be this 'natural' use of language that distinguishes the fluent speaker from the efficient speaker. The other, less obvious, target group would be the native or fluent speakers of Irish who would have an instinctive grasp of general vocabulary and syntax but whose lexicon is increasingly influenced by the language of global culture and media. They require an authoritative dictionary that provides them not only with the translation of modern terminology but also intuitive translations of the English phrases, idioms and turns of phrase from modern culture that have become a staple part of daily language. As it is envisaged, then, that both main target groups would prioritise Irish as a language for communication rather than a sterile or austere intellectual language, capturing and fronting current usage of Irish is of critical importance to the NEID project.

Unfortunately, the existing bilingual dictionaries do not constitute a ready-made platform on which to base the NEID. The last English-Irish dictionary was published over fifty years ago and the last Irish-English dictionary over thirty years ago, leaving them both out of date from the point of view of any words or phrases that have been coined in either language in the intervening period. Both dictionaries appeared before the advent of the CD-ROM or web-based dictionary, and consequently they couldn't accommodate a full range of translations. In some instances where a word or phrase in either dictionary in one language may well be very relevant today, its translation would fail the test of currency; also, both dictionaries would be generally seen to contain a high degree of formal language as opposed to common colloquial language, compared to recent trends in dictionaries where a much more democratic view is taken of what level of any language merits inclusion in a dictionary. In any case, there are no workable electronic versions of either dictionary that could be incorporated into NEID.

The absence of compatible lexicographical resources which would provide a starting platform is mirrored in the absence of a viable corpus of spoken Irish, either monolingual or bilingual. Although, as mentioned previously, a 30-million word Irish corpus was delivered in Phase 1 of the NEID project, it does not contain any spoken Irish. A new initiative in this field has been announced recently, but though it will be a valuable asset to future developments in Irish lexicography, it will not be in a position to facilitate the current NEID project.

4.1.1. Working solution

In order to capture current usage of Irish which may not be in existing sources, all material in the NEID except terminology or technical language is being translated initially without any reference to printed or other sources; rather the translators are directed to translate the material *ex tempore*, and their entries can be checked against sources later to complement the translated frameworks. By marking the translations that occurred to the translator *ex tempore*, the translations recorded for 70% of the NEID headwords will have a significant element of currency, and will inform the decision-making process at the editorial stage (see below). This would also enable a function in an electronic version of the dictionary whereby users could customise their view to front those translations that are labelled as current.

4.2. Dialectal coverage

Even more important for NEID is to ensure reasonable coverage of the three main dialects. This is primarily a measure to capture as many translations as possible for any given word or phrase in order to provide a wide range of options for the editorial process. As mentioned

previously, size limitations meant that the existing bilingual dictionaries were limited as to the range of translations that could be included. With the capacity of modern electronic dictionaries, however, it is possible to include as many translations as would be deemed accurate, and it is also possible to label each translation for its dialect of origin.

Depth of coverage is not the only reason for trawling the three main dialects, however, for if the NEID is to win over the language community it must be seen to treat all dialects on an equal basis. For many of the Irish-speaking community, particularly though by no means exclusively in Gaeltacht areas, their linguistic allegiance would be to their native or adopted dialect, and the other dialects would be almost akin to a different language. It is worth noting that there are some significant differences between the dialects in terms of pronunciation, vocabulary, semantics, grammar and syntax. Those who have learned or studied the language in an academic way often don't appreciate the difficulties of fluent but linguistically untrained speakers when faced with a different dialect. In 1932, ten years after the foundation of the new state, in the preface of his seminal work on Irish dialects, *Irish Dialects Past and Present*, O'Rahilly stated: 'Irish to-day lives only through its dialects; the obliteration of dialect would mean the extinction of the language' (: 10), and few would argue that the situation has changed much in the intervening period, much to the chagrin of those who wished to see standardised Irish established as the accepted 'national' version of the language.

For various reasons, the native-speaker population and a significant proportion of non-native speakers have never warmed to the state-sponsored Official Standard; instead it is often disparaged by them as 'book-Irish' or 'Dublin-Irish', or even a 'sell-out' to other dialects perceived as enjoying the favour of the establishment. Although, from a numbers point of view, the Gaeltacht native speakers may only account for a mere fraction of those who profess to know Irish, most learners of Irish tend to align themselves closely with a particular dialect and the fervour of the converted leads some of them to be even more protective of that dialect that the actual native speakers. In instances where a reader finds an Irish word or phrase prevalent in their own dialect to be absent in any of the existing dictionaries, the normal reaction is to view it as yet another instance of dialectal bias rather than to consider the necessary limitations of any printed dictionary.[4]

There is only so much that can be done, however, to cater for the diverse dialects, without making the project overly complex and impractically long. The project has a fixed target date and finite resources available, and while questions of spelling and grammar can be passed on to the writing and editing stage, it was decided to attempt to capture as many current translations as possible from the three main dialects. Of course, it would be impossible to process all 40,000 NEID headwords like this within the project's timeframe, and having considered various options, it was decided to divide the headword list into three categories according to their translation requirements, as follows:

- Category A: the 1000 most common lemmas in English. These headwords will be translated into each of the three main dialects, all translations being labelled for each dialect, then checked against existing sources before being passed on to the editing stage.
- Category B: any headwords not in Category A above or Category C below. Each headword is initially translated by a single translator and without reference to sources (except for any related technical senses). They are not marked by dialect.

---

[4] O'Rahilly also commented: 'We possibly have among us some speakers of Irish whose natural affection for their own dialect is apt to be distorted into a feeling of hostility towards dialects other than their own.' (1932: x).

- Category C: relatively uncomplicated headwords such as technical terms, proper names, rare words, etc. The translation process for Category C consists largely of checking against sources, and the source information is always included for each translation.

Ideally the whole dictionary, or at least the non-technical content, would be covered in each dialect, but time, resources and budget constraints required a less ambitious target, although the flexibility of the dictionary production system would easily allow us to upgrade the Category B headwords in the future. The limit of 1000 headwords for Category A was chosen because current research indicates that the 1000 most common lemmas in English cover c. 80% of normal usage in the language, whereas a further 1000 would only cover an additional 5% or so (Davies 2005. 108-111).[5]

The translation process involves each of these headwords being fully translated *ex tempore* by a translator from one dialect, then passed on to translators from the other dialects in turn. The translators are encouraged to insert as many translations as come to mind, regardless of whether or not the translations pertain to their own dialect. At a later stage, as the initial translators are working without reference to published sources, the whole entry is checked against existing dictionaries in case significant translations have slipped through the net. In actual processing, then, the full translation process entails four separate steps:

- Step 1: The entry is translated by a translator from Dialect A who labels each translation to indicate if it is used in Dialect A;
- Step 2: The entry is then checked by a translator from Dialect B who adds any additional translations and also labels each of the translations from Dialect A, indicating whether or not it is also used in Dialect B;
- Step 3: The entry is checked by a translator from Dialect C, who adds any additional translations and also labels each translation to indicate if it is used in Dialect C;
- Step 4: The entry is checked against set sources for other significant translations. Any new translations not generated in the *ex tempore* steps will be labelled accordingly.

The entry is then ready to be passed on to the editorial stage.

The flexibility of the DPS system enables us to attach attributes to each entry at various levels. For dialectal marking, we have added an attribute to allow each translation to be marked for each of the 3 main dialects. This marking is done in a DIAL field, where each translator is required to set the attribute for his/her own dialect to C (= Connacht), M (= Munster) or U (= Ulster), as appropriate, followed by one of the following codes:

1 (in this specific sense of the English headword) this is an Irish word/phrase I would use in my own dialect

---

[5] Table reproduced from Davies (2005) showing the percentage of normal usage in English, Spanish and German captured by the 2000 most frequently occurring tokens.

	Non-fiction			Fiction			Oral		
	Span	**Eng**	**Germ**	**Span**	**Eng**	**Germ**	**Span**	**Eng**	**Germ**
1st thousand	76.0	74.5	64.7	79.6	82.3	72.0	87.8	84.3	82.6
2nd thousand	8.0	4,7	7.2	6.5	5.1	5.4	4.9	6.0	4.4
First 2000	84.0	**79.2**	71.9	86.1	**87.4**	77.4	92.7	**90.3**	87.0

0 (in this specific sense of the English headword) this is an Irish word/phrase I would not use in my own dialect

A (in this specific sense of the English headword) I am not sure whether or not I would use this Irish word/phrase in my own dialect

N (in this specific sense of the English headword) this is a dialect-neutral word/phrase, i.e. I would use it in my own dialect even though it does not pertain to the dialect per se, for example: ríomhaire (= computer), bus (= bus), litearthacht (= literacy), etc.

Entry: 'at once' (= immediately)

Entry: 'must' (= have to)

Figure 1. Sample entries from the NEID database showing the dialectal marking after a translator from each dialect has contributed.

Obviously, as four translators are involved in the processing of each entry, control measures are required to minimise the ensuing impact on time and resources, and as a result of a period of testing, we have identified three ways of substantially reducing the workload, i.e.

As a general rule, only the node (keyword or phrase) is being translated rather than entire sample sentences or phrases (see Figure 2 below).

The dialectal marking is done only at the first instance of a given translation in the entry, unless there is reason to change the dialect code for that translation in a subsequent sense within the entry.

As a general rule, all material is being captured in standard spelling to avoid multiple variations of the same word/phrase.

Figure 2. Part of the entry for 'minute'.
As a general approach, only the node is translated rather than whole sentences.

5. Initial results

For the purposes of this paper test translations were carried out on a small number of headwords from Category A, these were specially chosen because they contained no technical senses. After each of the individual translators had gone through the translation process the number of translations and dialect fields were counted for each entry. The number of translations includes repeat uses of the same translation, whereas because the dialect field is only inserted for the first instance of any given translation in the entry, the number of dialect fields indicate roughly the number of distinct translations. The results are given in Figure 3 below: By July 2010 the paper will include statistical facts and figures from the first 10 months of the translation phase.

Headword	No. of senses	No. of translations after Translator 1	No. of translations after Translator 2	No. of translations after Translator 3	% increase in translation fields	No. of DIAL fields after Translator 1	No. of DIAL fields after Translator 2	No. of DIAL fields after Translator 3	% increase in DIAL fields
many	18	149	175	219	47%	81	102	124	53%
maybe	11	34	43	57	68%	30	35	48	16%
minute²	2	10	11	12	20%	7	8	8	14%
model	17	37	49	56	51%	20	25	30	50%
much	34	139	170	186	34%	105	128	129	23%
must	14	97	105	112	15%	36	44	50	39%
nature	15	58	87	102	76%	40	60	73	83%
need	15	98	163	188	92%	47	59	76	62%
once	27	88	102	120	36%	53	67	81	53%
particular	10	51	56	62	22%	31	34	26	-16%
person	14	32	46	51	49%	29	42	47	62%
Total/ Average	**177**	**793**	**1007**	**1165**	**47%**	**479**	**604**	**692**	**44%**

Figure 3. Initial results of the dialect-based translation process: total number of translations (including repeat uses) and total number of distinct translations.

6. Challenges

There remain a number of issues that are challenging to the process outlined above, such as:
- Material added by translators B and C will not be marked by translator A, nor translator C by translator B; however, a computer search string can identify translations missing a component of the dialectal profile and assign them as a batch to suitable translators from the relevant dialects.
- There are significant sub-dialects within each of the three major dialects, and useful material may not be captured.
- There are non-Gaeltacht Irish-speaking communities, many of whose first language would be Irish, whose use of the language also merits consideration, particularly as they are often more likely to use Irish terms and phrases for certain spheres of modern life (urban or technological, for instance) than the traditional native speaker.
- Translators are at times loath to pronounce on whether or not a given translation pertains to their own dialect.

7. Conclusion

Notwithstanding the challenges outlined above, the process of dialectal marking holds a number of significant advantages for the project, such as:
- It captures current translations that may not be covered in existing out-dated sources.
- It provides a dialectal profile 'from the horse's mouth' of words, phrases, etc. which is invaluable in the absence of dialectal corpora.
- It enables an element of dialectal marking in the final product, particularly in the electronic version.
- It enables the option to customise the electronic version so that a given dialect can be fronted in the dictionary.

- It enables the option of selecting a given dialect as the default pronunciation in the electronic version.
- It enriches the bilingual database, which may then be used as a research resource for other academic research projects.

References

Atkins, B.T. and Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press

Central Statistics Office Ireland. (2006). 'Census 2006 Principal Demographic Results'. In Central Statistics Office Ireland. Census 2006, Amended Principal Demographic Results 2006. 29-30.

Davies, Mark. (2005). 'Vocabulary Range and Text Coverage: Insights from the Forthcoming Routledge Frequency Dictionary of Spanish', In Eddington, David. (ed.) *Selected Proceedings of the 7th Hispanic Linguistics Symposium*. Somerville, MA 2005. 106-116.

De Bhaldraithe, T. (1959). *English-Irish Dictionary*. Dublin. An Gúm.

Dinneen, P. (1927). *Foclóir Gaedhilge agus Béarla: An Irish-English Dictionary*. Dublin: Irish Texts Society.

Kilgarriff, A. (1996). http://www.kilgarriff.co.uk/bnc-readme.html#lemmatised.

Kilgarriff, A. (1997). 'Putting frequencies in the dictionary'. In *International Journal of Lexicography* 10 (2). 135-155.

Kilgarriff, A. Rundell, M. Uí Dhonnchadha, E. (2006). 'Efficient corpus development for lexicography: building the New Corpus for Ireland. Language'. In *Resources and Evaluation Journal* 40 (2). 127-152.

Mac Amhlaigh, Liam. (2008a). 'The role of Foclóir Gaeilge-Béarla Néill Uí Dhónaill in Irish language lexicography in the twentieth century'. In Bernal, Elisenda and DeCesaris, Janet. (eds.). Proceedings of the XIII Euralex International Congress. Barcelona 2008. 1149-1154.

Mac Amhlaigh, Liam. (2008b). *Foclóirí agus Foclóirithe na Gaeilge*. Dublin: Cois Life.

McCone, K. McManus, D. Ó Háinle, C. Williams, N. Breatnach, L. (eds.). (1996). *Stair na Gaeilge*. An Sagart: Maynooth.

Nic Pháidín, C. Ó Cearnaigh, Seán (eds.). (2008). *A New View of the Irish Language*. Dublin: Cois Life.

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

Ó Giollagáin, C. Mac Donnacha, S. (2008). 'The Gaeltacht Today'. In Nic Pháidín, C. Ó Cearnaigh, Seán (eds.). (2008). *A New View of the Irish Language*. Dublin: Cois Life.

Ó Háinle, C. (1996). 'Ó Chaint na nDaoine go dtí an Caighdeán Oifigiúil'. In McCone, K. et al. (eds.). *Stair na Gaeilge*. Maynooth: An Sagart: 745-793.

O'Rahilly, T. (1932/1972). *Irish Dialects Past and Present*. Dublin: Dublin Institute for Advanced Studies.

Rialtas na hÉireann. (1958). *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil*. Baile Átha Cliath Oifig an tSoláthair.

Rundell, M. Atkins, S. & Kilgarriff, A. (2010). 'Database of ANalaysed Texts of English (DANTE) (the NEID database project)'. *Proceedings of the 14th Euralex International Congress*. Leeuwarden.