

From DANTE to Dictionary: The New English-Irish Dictionary

Pádraig Ó Mianáin, Cathal Convery
Foras na Gaeilge, Dublin
pomianain@forasnagaeilge.ie, cconvery@forasnagaeilge.ie

Abstract

Most major bilingual dictionary projects tend to largely involve adapting an existing bilingual dictionary, by adopting or amending the existing source language material as required and then supplying the necessary target language material. This paper describes the innovative approach followed in the production of the New English-Irish Dictionary (NEID) project. The online version of NEID was launched in 2013 (www.focloir.ie) and the project is to be completed when a printed version is produced in 2016. The English language content of NEID is based on the Database of ANalysed Texts of English (the DANTE database), a ground-breaking corpus-based lexical database developed specifically for the project. Attention is drawn to how the DANTE entry frameworks evolved through the various translation and editing processes to the final entries now available in the online. The paper also discusses the development of the Irish-language material and details some of the challenges, practical and editorial, encountered in the course of the production of the dictionary, and the working solutions that were developed.

Keywords: DANTE; bilingual; Irish-language

1 The New English-Irish Dictionary

The New English-Irish Dictionary (NEID) is being produced and funded by Foras na Gaeilge, the inter-governmental body with responsibility for the promotion of the Irish language, with a project budget of €6m. It is the first major English-Irish dictionary since de Bhaldraithe's English-Irish Dictionary (1959) and the first major bilingual dictionary in Ireland since Ó Dónaill's *Foclóir Gaeilge-Béarla* [Irish-English Dictionary] (1977). The first version of NEID was launched online in January 2013 with new material and revisions being uploaded at regular intervals until the envisaged completion of the online edition in 2015. The NEID will then be published in printed format in 2016.

2 Project Timeline and Phases

The NEID project started in 2000 and is on target to be completed at the end of 2016. The project is supported by IDM's DPS platform along with the Entry Editor interface. The online version of NEID

will eventually contain c. 130,000 sense units (c. 40,000 headwords) by December 2015, with a print version to follow in 2016. The project is divided into three major phases:

- **Phase 1: Planning and design** (started 2000, completed 2006). The planning and design phase was carried out by Lexicography Masterclass and delivered key elements such as an overall project plan, the English-language and Irish-language corpora which underpin the entire project, as well as sample entries, headword lists, draft style guides for each phase etc.
- **Phase 2: Compilation, Writing and Editing** (started 2008, to be completed 2016). This phase concerns the actual writing of the dictionary, and is discussed in detail below.
- **Phase 3: Publication** (started 2012, to be completed 2016). The online and mobile platforms were launched in January 2013 (www.focloir.ie) with about 30% of the final content of the dictionary; the online content is being added to on an incremental basis. The dictionary will also be made available as an app in 2014.

A number of separate databases were created to facilitate the progress of entries from the DANTE database through the various translation and editing stages to the online entry. The main databases are shown in Figure 1 below along with their purpose and some of the more significant changes between databases:

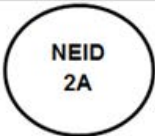
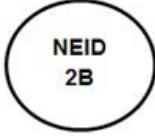
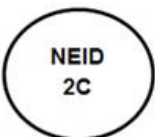
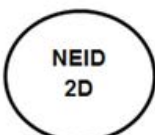
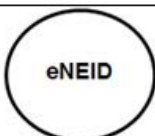
DATABASE	DESCRIPTION	PROCESSED BY
 NEID 2A	The source-language phase	English-language lexicographers
 NEID 2B	The target-language phase Changes automatically implemented: <ul style="list-style-type: none"> • Compounds extracted as full HWDs • FREQUENCY field inserted • TRID (translator ID) field inserted • TrCnt (translation container) inserted at set positions within entries 	Translators
 NEID 2C	The editing phase (at sense unit level) <ul style="list-style-type: none"> • Initial editing • Senior editing Changes automatically implemented: <ul style="list-style-type: none"> • EDID and SENED (editor & senior editor ID) fields inserted • reporting of style guide non-compliance 	Editors
 NEID 2D	The final entry phase <ul style="list-style-type: none"> • Sense units merged where possible • Disambiguators inserted • Streamlining of DOMAINS etc Changes automatically implemented: <ul style="list-style-type: none"> • Sense units grouped in blocks by POS • Links to grammar and sound files automatically inserted 	Senior Editors
 eNEID	Content uploaded to website Changes automatically implemented: <ul style="list-style-type: none"> • All fields not to be displayed are deleted (MEANING, FREQ, EDID etc) 	IDM

Figure 1: The main databases involved in the phased transition between the DANTE database (designated NEID 2A within the context of the NEID project) and the final online entries.

3 The Source-Language phase: the DANTE database (began 2008, completed 2010)

Lexicography Masterclass also supplied the English-language lexical database on which the source language content of the dictionary is based. It has subsequently been made available as a resource for various other lexicographical and linguistic projects as the Database of Analysed Texts of English or the DANTE database. The entire database can be viewed and explored at www.webdante.com.

DANTE is based on systematic analysis of a 1.7-billion-word corpus of English supported by the Sketch Engine corpus query package and the GDEX feature for optimising the corpus example selection. The final package is a highly-structured 16-million-word lexical database of English that can be used not only by lexicographers but also by linguists, researchers and teachers. From a purely lexicographical point of view, DANTE is target-language-neutral in the sense that, despite its origins as the foundation for a bilingual dictionary, its sole focus is on describing and evidencing the English language without regard for any potential companion language (the Irish language specifically in the NEID project). As such, DANTE is equally a primary resource for monolingual and bilingual English dictionaries.

The headline statistics of the DANTE database give an indication of its coverage: 42,000 headwords (62,000 if compounds are counted as separate headwords), 12,000 phrases and 3,000 phrasal verbs, 149,000 sense units in total, over 600,000 relevant corpus examples and 16 million words in all. However the real value of the database is seen in the fine detail of its coverage within the entries, for instance:

- 42 different grammatical structures for verbs, 16 for nouns and 15 for adjectives (see Table 1 below);
- all headwords classified by 12 levels of complexity;
- entry templates;
- over 150 domain categories;
- regular proformas for closed set entries; etc.

From a purely NEID perspective, the DANTE database more than adequately served its purpose of providing the translation team with a comprehensive and detailed lexical profile of each English-language headword which was a candidate for inclusion in the final dictionary.

4 The target-language phase: adding the Irish-language content (began 2009, completed 2012)

The aim of the translation phase of the NEID project was to maintain the structure and detail of the English-language entry frameworks while providing the editing team in the next phase with a comprehensive and detailed translation database, by adding as many relevant Irish-language equivalents

as possible to the English-language frameworks. It was at this stage that Foras na Gaeilge's own lexicography team took over the writing of the dictionary, as the translation and editing phases would be entirely driven by the Irish-language requirements.

The most efficient way of implementing this phase was to clone the DANTE database as a separate translation database and, as part of the cloning process, to automatically insert translation-specific fields in specified positions within the entry frameworks. Also, all compound entries were extracted from their mother entries (as they are in DANTE) and were promoted to full headword status, not least to facilitate the distribution of work batches in a more efficient manner. The end product of the translation phase is a rich database of bilingual entry frameworks consisting of the unabridged DANTE, plus 4 million words in Irish in over 600,000 translation fields.

Given the size and nature of the DANTE database, the translation task posed significant challenges at a number of levels, particularly given the lack of experienced Irish-language lexicographers available to the project. There were three main challenges:

- **The size of the translation task.** It would have been impractical and quite pointless to translate the entire English-language database, particularly as the specific remit of the translators was to provide as many relevant translations as possible in Irish. The translators were directed to concentrate on the key element or node of the word or phrase in question, and not to translate any of the surrounding text in the supporting examples unless it impacted on the translation; this was facilitated by automatically highlighting the node as part of the cloning procedure. Also, as each lexicographically significant fact in DANTE is contained in a specific structure container with corpus examples to match, there was no need to add translations to each individual example; to minimise such duplication, translation fields were automatically inserted and highlighted at the head of each structure container as a guide to the translators:



Figure 2: An example of how translation fields (highlighted in green) were automatically inserted at specified positions within the DANTE entries for the translation phase: the first sense unit of *expect*.

- The nature of the translation task.** Given the fifty-year gap since the previous English-Irish dictionary, a significant proportion of the English content of the NEID (approximately 30%) had to be translated into Irish from scratch, and a lot of the content that could be sourced in pre-existing dictionaries was dated or even obsolete in both English and Irish in terms of language, register etc. It was consequently decided from the outset to instruct the translation team to work *ex tempore* and NOT to consult published sources, except for technical terms. This self-reliance allied to the requirement to record as many Irish-language equivalents as they could think of came as a culture shock to translators who normally work from reference sources using one Irish-language word or phrase to translate a given English-language word or phrase. In addition to the translation challenges, the most frequent 1,000 headwords in English (20% of the sense units in DANTE) were each translated by three translators, one from each of the main dialects, who also recorded the relevance of each translation to their own dialect. Only technical terminology was translated from existing sources, with the translators recording those sources to facilitate the decision-making process at the editing stage.
- The layout and structure of DANTE.** To the uninitiated – and bearing in mind that the translators were primarily Irish-language specialists with no experience of lexical databases – DANTE is a

daunting beast. The hierarchical structure, consisting as it does of headwords, multiword entries, senses, structures and patterns, took some getting used to, and some of the grammar labels (or underlying structures) are less than transparent to all but the linguistically trained. The early solution here was to instruct the translators to ignore the grammar or structural information and concentrate on the corpus examples as a prompt for their translations in those cases.

5 The editing phase (began 2011, ongoing)

The initial step in the editing phase is to clone the completed translation frameworks to provide the editors with the full English-language profile plus the accompanying translations. This also ensures that the macrostructure of the edited content remains secure while the editors are drafting the final entry. Working on a sense-unit by sense-unit basis, the editors first decide on the English-language content of the entry, then the matching Irish-language content. Examples in the edited entries are included according to strict guidelines depending on the type and level of entry or sense involved, and the editors can either amend one of the corpus examples from DANTE or compose a new example entirely. Finally, the remainder of the framework is discarded once the English and Irish content of the entry has been decided.

It is also during the editing phase that new sense units are added to the dictionary database for senses not present in the original DANTE database. This may occur

- if a single sense unit in DANTE requires significantly different translation solutions in Irish (for instance *bassist* as a single sense in DANTE covers the ‘double base’ and ‘bass guitar’ but each instrument requires a different translation in Irish):



Figure 3: The entry *bassist* as one sense unit in DANTE (left) and how it was split in the editing database (right).

- if a word sense in usage in Ireland is not covered in DANTE (for instance *hurl* as verb = ‘to play hurling’ and as noun = ‘hurling stick’);
- if senses have come to the fore in the intervening years (for instance *to friend sb* in a social network context).

Also as part of the final streamlining of metalanguage etc, most of the grammar and structure labels are either removed entirely or converted to the smaller subset in use in the published dictionary; for instance, the current dictionary entries show only three verb structure labels (intransitive, transitive and modal) where the underlying DANTE database has forty-two:

Field	Examples	Range of fields present in English frameworks (DANTE)	Range of fields present in final entry
DOMAIN labels	<i>agri, food, ornith</i>	156	98
GRAM (grammar labels)	<i>abbrev, c_u, proper</i>	28	3
STRADJ (adjective structures)	<i>AVP_premod, that_0</i>	17	0
STRN (noun structures)	<i>AJ_pert, PP_X</i>	18	1
STRV (verb structures)	<i>AJP, Part, that_0_cond</i>	42	3
STYLE labels	<i>child, euph, pc, tech</i>	21	9

Table 1: Examples of the reduction in the number of fields between DANTE and the final entry.

The editors can also recommend and mark entire entries or individual sense units for omission from the final dictionary, but as this is the prerogative of the senior editing team at the next phase no entries or sense units are discarded at this stage.

Though still not at a point where it could be published, the edited database is now a much leaner version of the translation database (for instance, the word count in the 19,000 frameworks edited by April 2014 was 2.2 million compared to 8.5 million in the same translation frameworks).

6 Publication phase (began 2012, ongoing)

In this phase the final dictionary entries are arranged and prepared for uploading to the website (and in 2016 for the print version). The most obvious change is the re-ordering of all sense units by part of speech: in the DANTE database, associated senses are clustered together regardless of part of speech, but NEID follows the traditional POS-based order. This re-ordering is done automatically, reflecting the order of precedence of the sense units in DANTE, and occasionally requires minor adjustment where the order of precedence of senses under one POS may not mirror another POS. In the case of *strip*, for instance, the first sense as a verb is ‘to remove clothes’, but as a noun the sense ‘narrow band

of sth’ would be much more common than ‘an instance of removing clothes’, and thus was brought to the beginning of the noun section:

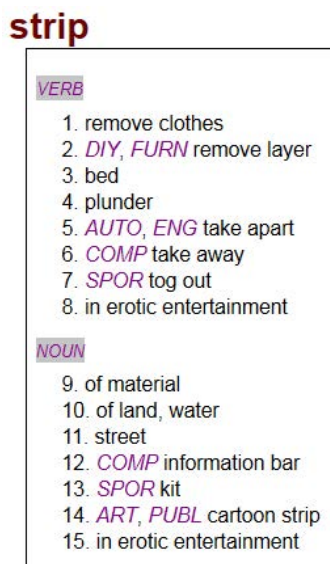


Figure 4: The manually adjusted sense order under *strip* where the noun sense associated with the first verb sense was manually transferred to a lower position in the noun order.

Another instance where the automatic re-ordering needed to be manually tweaked was *turf* as a noun, where the sense ‘peat’ is much more common in Ireland than the sense ‘sod’ and was brought to the top of the order. It is also at the publication phase that entries or senses are marked for exclusion from the published dictionary though they are not deleted from the dictionary database.

Sense units are then merged where possible in order to avoid the user having to scroll unnecessarily through numerous senses with similar translations. For instance, the Irish equivalent remains the same for seven DANTE senses of the word *studio* as ‘work area’, so those seven senses were amalgamated into one single sense in NEID with examples added to indicate the breadth of coverage. Similarly four sense units under *coffee* in DANTE became one in NEID with pertinent examples:

<p>NOUNBLK</p> <ul style="list-style-type: none"> FWKSENCNT [1] <ul style="list-style-type: none"> POS [N] <ul style="list-style-type: none"> LABELGP <ul style="list-style-type: none"> DOMAIN [ART] DOMAIN [PHOT] DOMAIN [TV-RAD] DOMAIN [ETC] MEANING a room where an artist, sculptor or photographer works MEANING a room or building where television and radio programmes are recorded or broadcast MEANING a place where cinema films are made or produced MEANING a room where sound or music recordings are made MEANING a room where dancers, actors, people doing exercise etc can practise MEANING a film or television production company MEANING a place in a company where new products are designed 	<p>studio</p> <p>1 <i>NOUN ART, PHOT, TV-RAD, ETC</i> <i>stiúideo masc4</i> 🗣️ C M U aerobics studio <i>stiúideo aeróbaice</i> art studio <i>stiúideo ealaíne</i> film studio <i>stiúideo scannán</i> photographic studio <i>stiúideo grianghrafadóireachta</i> recording studio <i>stiúideo taifeadta</i> television studio <i>stiúideo teilifíse</i></p> <p>2 <i>NOUN studio flat</i> <i>árasán stiúideo</i> 🗣️ C M U</p>
<p>NOUNBLK</p> <ul style="list-style-type: none"> FWKSENCNT <ul style="list-style-type: none"> POS [N] <ul style="list-style-type: none"> LABELGP <ul style="list-style-type: none"> DOMAIN [DRINK] DOMAIN [FOOD] MEANING the liquid as a drink MEANING in its dried form, for making into a drink MEANING in the form in which it grows on plants MEANING a unit or containerful of the drink 	<p>coffee</p> <p>1 <i>NOUN DRINK, FOOD</i> <i>caife masc4</i> 🗣️ C M U a cup of coffee <i>cupán caife</i> I like strong coffee <i>is maith liom caife láidir</i> he takes two spoonfuls of coffee <i>glacann sé dhá spúnóg chaife</i> she ordered two coffees <i>d'ordaigh sí dhá chaife</i> black coffee <i>caife dubh, caife gan bhainne</i> decaffeinated coffee <i>caife gan chaiféin</i> instant coffee <i>caife ar an toirt</i> white coffee <i>caife bán</i> coffee jar <i>próca caife</i></p> <p>2 <i>ADJECTIVE COL</i> <i>ar dhath an chaife</i></p>

Figure 5: Multiple senses of *studio* and *coffee* merged into a single unit within the 2D database (left) and how the final entry appears in the online dictionary (right).

This approach, which would not be possible in a decoding dictionary, is facilitated by the fact that virtually all users of NEID are fully fluent in English and have an intuitive understanding of the English content.

To complete the editing process, sense disambiguators and domain labels are added or removed as required, and finally, when the text content of the entry is finalised, grammar and sound files are attached to the translation fields in preparation for publishing online.

7 Practical challenges

The project to produce a modern bilingual English-Irish dictionary faced a number of significant practical challenges. Some of these challenges may apply to similar projects in any language, some may apply in particular to other lesser-used languages, while others stemmed from the previous gap in bilingual lexicography in Irish and the consequent problems of finding suitably qualified and experienced staff at editorial and managerial level. The main such challenges were:

- **An innovative approach to dictionary compilation.** The approach followed in the NEID project is ground-breaking in that the final content of the dictionary is entirely derived from a lexical database based on systematic corpus analysis. This required all translators and editors to exercise a higher level of judgement throughout.
- **Technical challenges.** The technical working environment of modern lexicography posed a significant challenge, both at the organisational and at the individual level, where continuous training and monitoring was required.
- **Management of staff and processes.** The varied nature of the project and the overlapping of the various phases and sub-phases required careful organisation and management, particularly as a lot of the processes in the translation and editing stages were being developed from scratch.
- **Training and upskilling of staff.** The project required a significant and continuous programme of training and monitoring as the processes for each phase of the project were being tested and implemented, and this burden was exacerbated by the shortage of experienced people at all levels in the project.

8 Conclusion

Notwithstanding the challenges posed by undertaking such an innovative model for this dictionary project and the particular challenges arising in relation to Irish lexicography, the fundamental benefit to NEID is that DANTE enabled the project team to produce a dictionary which they can claim is uniquely Irish, in as much as every word of both English and Irish in the dictionary is there by the editors' choice. The customisations and additions to the DANTE database at the subsequent three stages of the compilation process of NEID show that DANTE is a very flexible resource which can be adapted to the requirements or wishes of any English-language dictionary project.

9 Further information

The Database of Analysed Text of English (DANTE): www.webdante.com

The New English-Irish Dictionary: www.focloir.ie

10 References

- Atkins, B. T. S. (2010). The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data. In G.-M. de Schryver (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis: A Festschrift for Patrick Hanks*. Kampala: Menha Publishers (www.menhapublishers.com).
- Atkins, B. T. S., Kilgarriff, A. & Rundell, M. (2010). Database of ANalysed Texts of English (DANTE): the NEID database project. In *Proceedings of the Fourteenth EURALEX International Congress*. Leeuwarden: Fryske Akademy, pp. 549-556.
- Atkins, B. T. S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Convery, C., Atkins, B. T. S., Kilgarriff, A., Rundell, M., Ó Mianáin, P., & Ó Raghallaigh, M. (2010). The DANTE Database (Database of ANalysed Texts of English). In *Proceedings of the Fourteenth EURALEX International Congress*. Leeuwarden: Fryske Akademy, pp. 293-295.
- Convery, C., Ó Mianáin, P., & Ó Raghallaigh, M. (2010). Covering All Bases: Regional Marking of Material in the New English-Irish Dictionary. In *Proceedings of the Fourteenth EURALEX International Congress*. Leeuwarden, pp. 609-619.
- Kilgarriff, A. (2010). DANTE: A Detailed, Accurate, Extensive, Available English Lexical Database. In *Proceedings of a meeting of the North American Association for Computational Linguistics (NAACL-HLT)*, Los Angeles, June.
- McCarthy, D. (2010). DANTE: a new resource for research at the syntax-semantics interface. In *Proceedings of Interdisciplinary Workshop on Verbs, Pisa*.
- Ó Mianáin, P. (2013). The New English-Irish Dictionary. In Stickel, G. & Varadi, T. (eds.) *Lexical Challenges in a Multilingual Europe: Contributions to the Annual Conference 2012 of EFNIL in Budapest*. Frankfurt um Main: Peter Lang, pp. 111-114.
- Rundell, M. & Atkins, B. T. S. (2011). The DANTE database: a User Guide. In *Proceedings of eLex 2011*. Trojina: Institute for Applied Slovene Studies, pp. 233-246.
- Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end?. In Meunier, F., De Cock S., Gilquin G. & Paquot M. (eds) *A Taste for Corpora: A tribute to Professor Sylviane Granger*. Benjamins, pp. 257-281.

