EURALEX 2010 Reports on Lexicographical and Lexicological Projects

Database of ANalysed Texts of English (DANTE): the NEID database project

B. T. Sue Atkins, Adam Kilgarriff, Michael Rundell
Lexicography MasterClass Ltd.

Overview

The project in question is the source-language analysis stage of the New English-Irish Dictionary (NEID: http://www.focloir.ie). The task is to build a fine-grained lexical database of English. The database is being developed for Foras na Gaeilge, Dublin (FnaG: http://www.forasnagaeilge.ie/), the official government body responsible for the promotion of the Irish language, and their primary aim is to use it as the starting point for the NEID. However, because the database is target-language-neutral, it has the potential to function as: a 'starter pack' for any bilingual dictionary where English is the source language (SL); a resource for updating or enhancing any English monolingual dictionary; and, thanks to the machine-readability of many of its fields, a source of linguistic information for software and research institutions. Consequently, Foras na Gaeilge intend to market the DANTE database worldwide.

This database is currently being built by the Lexicography MasterClass (http://www.lexmasterclass.com) and their 15-strong lexicographic team, managed by Valerie Grundy, Managing Editor; the project administration is in the hands of Diana Rawlinson, Project Administrator. We describe the project and introduce the resource, looking especially at those aspects which contributed most to the efficient production of a comprehensive corpus-based account of the language (a process described in detail in Atkins and Rundell 2008).

Entries in the database include the lexical fields which would be expected in a resource of this type, recording data about meaning, grammar, combinatorial behaviour, colligational preferences and text-type information. While most of these elements are familiar, there are a number of ways in which DANTE is unique – both as a product, and in terms of the project which created it. Ultimately, what makes it special is that an existing methodology has been applied systematically, across the whole lexicon, and at a level of detail which we believe to be unprecedented. In order to achieve this, we made a number of significant innovations in the areas of project management and software. We believe these innovations have the potential to benefit other lexicographic enterprises.

Starting point

The project we describe here (NEID Phase 2a) forms part of a much larger lexicographic programme being managed by Foras na Gaeilge. The NEID has a target date of 2012 for completion, and the broader programme may ultimately include other English-Irish and Irish-English bilingual and Irish monolingual dictionaries.

Phase 2a began with the following resources created in NEID Phase 1:

1. a 1.7 bn word lexicographic corpus;

2. the Sketch Engine corpus query system with the corpora loaded;
3. the IDM DPS with a working document type definition (DTD) loaded and project-management data set out (scheduling, textflow, budgets etc.);
4. a user profile
5. a set of headword selection principles and a headword list;
6. a list of linguistic labels for marking register, style, domain etc.;
7. a working style guide for the analysis process, to be fleshed out as necessary;
8. 50 'template' (model) entries, for specific lexical sets;
9. 100 sample dictionary entries covering the full range of entry types;
10. a detailed description of the proposed entry structures needed for the dictionary;

Our team of 15 skilled editors included several American and Irish lexicographers, the remainder being from the UK. All worked from home, equipped with dual monitors to facilitate corpus consultation during the compiling process.

In the account that follows, we focus on

- entry structure, with particular reference to grammatical and collocational information
- software for analyzing and recording linguistic data
- project management and quality-control mechanisms

Entry structure

We refer here to the sample entry for *scorn* (see other file), a relatively short and simple entry, yet complex enough to show our lexicographic principles in action and give a flavour of the database as a whole.

The first lexical unit (LU), indicated by a 'Framework Sense Container'[1] or FWKSENCNT, has a POS label, followed by a GRAM tag: this field is used for recording secondary grammatical characteristics (such as countability or reciprocal use) and colligational preferences (such as preferred position, mood, or number). The GRAM field allows us, for example, to categorize any adverb as 'manner' (the default), 'degree' (*astronomically expensive*), 'viewpoint' (*Astronomically, the evidence shows*…), or 'sentence' (*Frankly, this doesn't interest me*). Most of the LUs in the sample also include a 'structure container' (FWKSTRCNT), and this is where the bulk of the syntactic information is shown. The second LU, for example, shows a verb which can either take a simple noun phrase object (NP) or be used in the pattern *scorn someone for something* (NP_PP_X, with the preposition 'for'). For each main word class, a drop-down menu offers a wide choice of structures, with 42 available for verbs alone (of which only three appear in the sample entry).

The database structure provides a range of options for handling multiword expressions of various types, including compounds, phrasal verbs, phrases (see *pour scorn on* in the sample), support verbs, collocations (as shown in the fourth sense of *scorn*), itemisers, and recurrent chunks.

[1] In the context of the NEID bilingual dictionary development, the DANTE database was known as 'the Frameworks', hence this element name.

Detailed information about text-type preferences is also provided: each of the six categories of label – for register, domain, style, evaluation (or speaker attitude), region, and time – has a number of attributes. The domain set is especially rich, including over 150 types of label. (The sample word does not include any labels.)

It is a key feature of DANTE that, as the entry for *scorn* illustrates, every linguistic feature we identify is exemplified by at least one (and in most cases three or more) sentences from the corpus. This necessarily short overview of DANTE's entry structure should give an idea of the granularity of the database.

Software

The software aspects we deal with include:

- the use of the Sketch Engine (http://www.sketchengine.co.uk/) corpus query software, with a corpus of 1.7bn words. The Sketch Engine's functions – especially its 'word sketches' – are well known (e.g. Kilgarriff et al. 2004), but in this case the software was customised for the project. The grammatical relations in the word sketches were tailored to match the syntactic coding in DANTE (based on the theory of lexicographic relevance: Atkins et al. 2003), so that the evidence in each of the word sketch columns (whether for unary or binary relations) corresponded to the codes and data-fields used in DANTE.
- the use of the 'GDEX' algorithm (Kilgarriff et al. 2008) for detecting and foregrounding the 'best' examples sentences in the corpus, combined with a seamless interface with the project's dictionary-writing system. It is a feature of the database that every linguistic fact recorded is accompanied by an average of three full corpus sentence(s) illustrating its use in text. The process of selecting examples and transferring them to the relevant field in the dictionary database was thus streamlined for maximum efficiency;
- the use of IDM's Dictionary Production System (DPS: http://www.idm.fr/products/dictionary_writing_system/27/), not only for managing textflow and project administration, but for running sophisticated data-searches in order to ensure high levels of quality.

Project management

With a long track record of running complex dictionary projects, DANTE's editorial management team brings a good deal of experience to this assignment. Nevertheless, we took advantage of the available software to introduce a a number of innovations in the area of project management. These included:

- improving the reliability of schedule and workflow by classifying, before the compiling started, over 50,000 headwords according to type and complexity. Using information drawn from corpora and from earlier projects, we assigned every headword to one of 16 categories, from the simplest single-sense words to the most complex items (such as 'light' verbs and major function words). We then established provisional timings for each type, and this has helped ensure that the compilation schedule remains on track.
- the systematic use of 'template' entries: the benefits of this type of proforma entry are discussed in Atkins & Rundell 2008 (123-128). Originally used in

the *Oxford-Hachette English-French Dictionary* (OHFD) and the *Macmillan English Dictionary for Advanced Learners* (MEDAL), they are applied more systematically in the DANTE database. We developed 68 of these outline entries, and they were pre-loaded into the database at every relevant headword, and thus already in place at point when entries came to be compiled. Thus, for a significant proportion of the lexicon, the compilation process was streamlined, and entries within a given semantic category show high levels of consistency.

- a novel and highly effective approach to quality control, combining conventional entry-editing by senior team members with the use of complex search scripts that list all entities of a specific type and allow rapid checking for accuracy. The 'SkXml' search function in IDM's DPS program allows users to construct complex search strings which can pinpoint any information category in the database. To give an example, the string:

  <FwkStrCnt:(%<strv@code=(NP AVP)),<hwd:(^[a-d].*)

  searches for every appearance, in the range A to D, of the verb structure ('strv') NP ADV (a verb followed by an object and an adverbial): this proved useful because there is a known risk of editors using a plain AVP code (without the NP) when an object is implicitly present (as in a sentence like *The book was favourably reviewed*). Similarly, we have applied search strings to identify any noun with either 'mass' or 'uncount' in the GRAM tag. Earlier in the project, this was a common area of confusion, but identifying all the relevant entries allowed us not only to correct errors but also to refine policy guidelines for the editorial team. For scheduling purposes, the lexicon has been divided into nine large alphabetic chunks. As each of these batches is completed, we run a set of 187 search strings on the text and tidy up any anomalies. By complementing 'traditional' editing techniques with this more systematic approach, we have ensured high levels of quality in the lexicographic data.

Apart from the meaning explanations, all the significant information is machine-readable: it should be possible to program a computer to map our grammar codes to those of other projects where grammar is recorded; our collocates can be directly linked by computer to actual lexical items in the corpus; and similarly our examples (each attached to a specific linguistic fact) to corpus sentences.

DANTE is a lexicographic project where the end-product is not a dictionary but an in-depth analysis to be used for creating one or more dictionaries. The users of DANTE are not the dictionary-using public but the lexicographic teams who will take this on to dictionary status. There is no need to compromise precision for them. The database is thus a rare, possibly unique, beast: a rich and comprehensive lexicographic analysis on linguistic principles, prepared on a substantial budget by a large team of professional lexicographers, and uncompromised by the needs of accessibility to non-linguist users.

References

Atkins, B. T. Sue, Charles Fillmore & Christopher Johnson (2003) "Lexicographic relevance: selecting information from corpus evidence", in *International Journal of Lexicography*, guest editor Thierry Fontenelle, Oxford, OUP: 16:3 251-280

Atkins, B. T. Sue and Valerie Grundy (2006) "Lexicographic profiling: An aid to consistency in dictionary entry design". In *Proceedings of the Twelfth EURALEX International Congress, EURALEX 2006*, Alessandria Italy: Edizioni dell'Orso. 1097-1107.

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

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). 'The Sketch Engine', in Williams and Vessier (2004). 105-116. Reprinted in Fontenelle (2008).

Kilgarriff, A., Rundell, M., & Uí Dhonnchadha, E. 2007. 'Efficient corpus development for lexicography: building the New Corpus for Ireland', *Language Resources and Evaluation* 40:2. 127-152.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychly, P. (2008). 'GDEX: Automatically Finding Good Dictionary Examples in a Corpus', in Bernal, E. and DeCesaris, J. (Eds) Proceedings of the XIII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra: 425-433.