The DANTE Database (Database of ANalysed Texts of English)

Cathal Convery, Pádraig Ó Mianáin, Muiris Ó Raghallaigh, Foras na Gaeilge
Sue Atkins, Adam Kilgarriff, Michael Rundell, The Lexicography MasterClass Ltd.

This database (www.webDante.com) was designed and created for Foras na Gaeilge by the Lexicography MasterClass and their 15-strong team led by Valerie Grundy (Managing Editor); textflow is managed by Diana Rawlinson (Project Administrator). The corpus of 1.7 bn words of current English, custom-built in 2007, was queried using the Sketch Engine (www.sketchengine.co.uk/), and the database was compiled in IDM's Dictionary Production System (DPS: www.idm.fr). The present volume contains a fuller description of this project (Atkins, Kilgarriff and Rundell Database of ANalysed Texts of English (DANTE): the NEID database project) and of its use in a bilingual dictionary (Convery, Ó Mianáin and Ó Raghallaigh Covering all bases: Regional Marking of material in the New English-Irish Dictionary).
The 95,000 or so DANTE entries cover approximately 50,000 headwords and 45,000 compounds, idioms and phrasal verbs, using over 40 datatypes. The lexical entry is subdivided into lexical units, each a sense of a single- or multi-word lemma. Almost every linguistic fact recorded is accompanied by full corpus sentences illustrating its use in text. Apart from the definitions and the corpus-derived example sentences, all the significant information is machine-retrievable. Functionality demonstrated here includes simple and complex searches over various combinations of datatypes and the automatic insertion of empty translation fields for use in dictionary building.
DANTE was created as the initial stage of compilation of the New English-Irish Dictionary. Its long-term potential is much more far-reaching: it offers publishers world-wide a comprehensive launchpad for bilingual dictionaries with English as the source language or the draft stage of a learners' dictionary of English; a source of updating material for an existing dictionary, etc. It offers software developers, universities and other research institutions a resource for improved word sense disambiguation, the creation or enhancement of online lexicons, and other uses in software applications such as machine-assisted translation, information retrieval systems, etc. More details from info@webDante.com.

DANTE is, we believe, unique in the depth of its analysis and its comprehensive recording of lexicographically relevant facts (as defined in Atkins et al 2003). This demo complements two papers presented to Euralex 2010: *Database of ANalysed Texts of English (DANTE): the NEID database project* by Atkins, Kilgarriff and Rundell; and *Covering All Bases: Regional Marking of Material in the New English-Irish Dictionary* by Convery, Ó Mianáin and Ó Raghallaigh.

The demo presents this monolingual, target-language-neutral database (www.webDante.com) which was designed and created for Foras na Gaeilge by the Lexicography MasterClass and their 15-strong team, led by Valerie Grundy (Managing Editor); textflow was managed by Diana Rawlinson (Project Adminstrator). The Lexicography Editor for Foras na Gaeilge was Pádraig Ó Mianáin and the Project Manager was Cathal Convery. The corpus of 1.7 bn words of current English, built for this project in 2007, was queried using a customized version of the Sketch Engine (www.sketchengine.co.uk/) , described in Kilgarriff et al (2004)), and which included a number of key enhancements, among them the GDEX functionality described in Kilgarriff et al (2008). The database was compiled in IDM's Dictionary Production System (DPS: www.idm.fr).

The 95,000 or so entries cover approximately 50,000 headwords and 45,000 compounds, idioms and phrasal verbs, using over 40 datatypes to record their behaviour in the corpus. The process is based on the approach to corpus analysis described in Atkins and Rundell (2008). The lexical entry is subdivided into lexical units, each being a sense of a single- or multi-word lemma. Almost every linguistic fact recorded is accompanied by full corpus sentences illustrating its use in text. Apart from the definitions and the corpus-derived example sentences, all the significant information is machine-retrievable.

In the short term DANTE was created in order to serve as the initial stage of compilation of the *New English-Irish Dictionary*. Its long-term potential is much more far-reaching: it offers publishers world-wide a comprehensive launchpad for any bilingual dictionary where the source language is English, the draft stage of a learners' dictionary of English, or a source of updating material for an existing dictionary, etc. It offers software developers, universities and other research institutions a resource for improved word sense disambiguation, the creation or enhancement of online lexicons, and other uses in software applications such as machine-assisted translation, information retrieval systems, etc. More details from info@webDante.com .

The functions to be demonstrated include the retrieval of data based on various datatypes and combinations of datatypes, down to very fine-grained levels; the automatic insertion of empty translation fields for use in dictionary building; and the customized link from the Sketch Engine to the DPS for instant copying of example sentences.

The information recorded where relevant for each lexical unit includes

- wordclass
- secondary grammar (inherent properties of headword)
- informal definitions
- syntactic constructions and arguments of the headword
- lexical collocates based on corpus frequencies
- support verb constructions
- support prepositions
- domain /subject field
- regional variety
- speaker/writer attitude
- time
- register
- style
- full example sentences (from corpus)
- variant forms
- derived forms
- cross-reference

References

Atkins, B. T. Sue; Fillmore, Charles; Johnson, Christopler (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 Rundell, Michael (2008). *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press.

Kilgarriff,Adam; Rychlý,Pavel; Smrz, Pavel; Tugwell, David (2004). 'The Sketch Engine' In *Proceedings of the XI EURALEX International Congress*. Lorient, France, July: 105-116. Reprinted in *Lexicology: Critical concepts in Linguistics*. Hanks, editor. Routledge, 2007

Kilgarriff,Adam; Husák,Milos; McAdam, Katy; Rundell, Michael; Rychlý, Pavel (2008). 'GDEX: Automatically finding good dictionary examples in a corpus' In *Proceedings of the XIII EURALEX International Congress*. Barcelona, Spain.