Using the FrameNet approach in a professional lexicography project (the DANTE database)

Sue Atkins
Lexicography MasterClass Ltd.
Sue.Atkins@lexmasterclass.com

This talk describes important aspects of the Database of Analysed Texts of English (DANTE), devised according to the principles of lexicographic relevance underlying the FrameNet project, as described in Atkins et al (2003). The database was commissioned by Foras na Gaeilge, Dublin[1] to serve as a launchpad for their *New English-Irish Dictionary*.[2] It is being built by the Lexicography MasterClass[3] and their 15-strong lexicographic team, using customized software, IDM's Dictionary Production System.[4] Its purpose is to assemble from a text corpus all the lexicographically relevant data for each word in the wordlist. It is currently 64% compiled, and will be completed in July 2010, when it will hold approximately 50,000 lexical entries, covering more than 90,000 lemmas (single-word and multiword). The corpus of 1.7 billion words created for this project is queried using a customized version of the Sketch Engine software.[5]

The database records corpus data in over 40 datatypes, as relevant to the headword (see Atkins & Grundy 2006). Apart from the definitions and the corpus-derived example sentences, all the significant information is machine-retrievable. Cross-tabulation is fast and easy, and the DPS permits the retrieval of data based on various datatypes and combinations of datatypes, down to very fine-grained levels. Customized functions for this project include the prioritizing by the Sketch Engine of 'good' corpus examples via the GDEX program (Kilgarriff et al 2008) together with the direct ('one-click') import of complete corpus sentences into the example field; and, by IDM for use in subsequent stages of a bilingual dictionary project, the automatic insertion in specified locations of empty translation fields.

The DANTE entries complement the FrameNet data in that they record for each LU facts which are outside the remit of FrameNet.

The database design reflects Fillmore's definition of lexicographic relevance (Fillmore 1995):

> *The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions (complements, modifiers, adjuncts, etc.)*

[1] http://www.forasnagaeilge.ie/

[2] http://www.focloir.ie

[3] http://www.lexmasterclass.com The project was designed by and is the responsibility of the LexMC directors, Sue Atkins, Adam Kilgarriff and Michael Rundell. The team of lexicographers is led by Valerie Grundy, Managing Editor, and the project is managed by the Project Administrator, Diana Rawlinson. The Project Manager at Foras na Gaeilge is Cathal Convery (cconvery@forasnagaeilge.ie).

[4] http://www.idm.fr/products/dictionary_writing_system/27/

[5] http://www.sketchengine.co.uk/

which the word can have in such constructions, in so far as the occurrence of such accompanying elements is dependent in some way on the meaning of the word being described.

The lexicographic team was trained to identify the maximal projection of the lemma in corpus contexts, and to record as many of the relevant facts as possible. The technique used is that set out in Atkins and Rundell (2008).

The entry is subdivided into lexical units (LUs): these are senses of the headword (as in FrameNet) and/or multiword expressions in which the headword figures, such as compounds, idioms, phrasal verbs etc. Significant datatypes recorded where relevant for each LU include:

- an informal definition;
- part of speech;
- secondary grammatical information;
- labels indicating subject field, regional variety, attitude, time, register, style;
- syntactic constructions,
- corpus-derived collocates
- variant forms;
- derived forms;
- support verbs;
- support prepositions and
- salient patterns found in the corpus for a specific LU.

With the exception of encyclopedic headwords and technical vocabulary items, every fact is illustrated by an average of three full unedited corpus sentences.

We shall briefly introduce the principal datatypes, illustrating each one with an extract from a DANTE lexical entry. The concept of 'template' (reusable) entries will be explained and exemplified.

The FrameNet and DANTE entries for the verb *observe* will serve as a case study, in which the types of information found in FrameNet and those in DANTE will be discussed and compared. The comparison will show where the two databases overlap, and where each complements the other.

The DANTE database was specifically designed to be reusable, to serve as a resource for the enhancement of online lexicons in both commercial and research environments. The final section of the presentation will consider the issue of how the FrameNet and DANTE databases may contribute to the automatic enrichment of the other, an aim which of course depends for its fulfilment on the possibility of mapping DANTE LUs to their counterparts in FrameNet.

References

Atkins, B. T. S., C. J. Fillmore & C. R. Johnson (2003) "Lexicographic relevance: selecting information from corpus evidence", in *International Journal of Lexicography*, Oxford, OUP: 16:3 : 251-280.

Atkins, Sue & 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.

Fillmore, C. J. (1995) "The Hard Road from Verbs to Nouns", in: M. Chen and O.Tzeng (eds.) *In Honor of William S-Y. Wang* . Taipei, Taiwan: Pyramid Press. 105-129.

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