A Detailed, Accurate, Extensive, Available English Lexical Database

Adam Kilgarriff
Lexical Computing Ltd
Brighton, UK
adam@lexmasterclass.com

Abstract

We present an English lexical database which is fuller, more accurate and more consistent than any other. We believe this to be so because the project has been well-planned, with a 12-month intensive planning phase prior to the lexicography beginning; well-resourced, employing a team of fifteen highly experienced lexicographers for a thirty-month main phase; it has had access to the latest corpus and dictionary-editing technology; it has not been constrained to meet any goals other than an accurate description of the language; and it has been led by a team with singular experience in delivering high-quality and innovative resources. The lexicon will be complete in Summer 2010 and will be available for NLP groups, on terms designed to encourage its research use.

1 Introduction

Most NLP applications need lexicons. NLP researchers have used databases from dictionary publishers (Boguraev and Briscoe, 1989; Wilks et al., 1996), or developed NLP resources (COMLEX (Macleod et al., 1994), XTAG (Doran et al., 1994)) or used WordNet,(Fellbaum, 1998) or have switched to fully corpus-based strategies which need no lexicons. However the publishers' dictionaries were pre-corpus, often inconsistent, and licencing constraints were in the end fatal. COMLEX and XTAG address only syntax; WordNet, only semantics. Also these resources were not produced by experienced lexicographers, nor according to a detailed, stringent 'style guide' specifying how to handle all the phenomena (in orthography, morphology, syntax, semantics and pragmatics, from spelling variation to register to collocation to sense distinction) that make lexicography complex. Unsupervised corpus methods are intellectually exciting but do not provide the lexical facts that many applications need.

We present DANTE (Database of Analysed Texts of English), an English lexical database. For the commonest 50,000 words of English, it gives a detailed account of the word's meaning(s), grammar, phraseology and collocation and any noteworthy facts about its pragmatics or distribution.

In outline this is what dictionaries have been doing for many years. This database is of more interest to NLP than others (for English) because of its:

- quality and consistency
- level of detail
- number of examples
- accountability to the corpus
- purity: it has been created only as an analysis of English, and has not been compromised by publishing constraints or other non-lexicographic goals
- availability, on licencing terms that promote its research use and also the re-use of enhanced versions created by NLP groups.

2 The Project

The overall project is the preparation of a New English Irish Dictionary, and is funded by Foras na Gaeilge, the official body for the (Gaelic) Irish language.[1] The project was designed according to a

[1] FnG was set up following the Good Friday Agreement of 1998 on Northern Ireland, between the Governments of the Re-

Proceedings of the NAACL HLT 2010: Demonstration Session, pages 21–24,
Los Angeles, California, June 2010.

model where the first stage of the production of a blingual dictionary is a target-language-neutral monolingual analysis of the source language listing all the phenomena that might possibly have an unexpected translation. (The next stages are then translation and 'finishing'.) The 2.3 MEuro contract for the analysis of English was won by Lexicography MasterClass Ltd in 2007.[2] The lexicographers are working on the letter 's' at time of writing and the database will be complete in Summer 2010.

3 Lexicography

Writing a dictionary is a large and complex undertaking. Planning is paramount.

In the planning phase, we identified all the aspects of the behaviour of English words which a full account of the lexicon should cover. We then found words exemplifying all aspects, and prepared a sample of one hundred model entries, where the hundred words chosen covered all the principal phenomena (Atkins and Grundy, 2006). A detailed style guide and corresponding DTD were written. We created the New Corpus for Ireland (NCI) (Kilgarriff, 2006), and set up a corpus query system (Lexical Computing's Sketch Engine; http://www.sketchengine.co.uk) and dictionary editing system (IDM's DPS: http://www.idm.fr) for the project to use. 50,000 headwords were identified and each was classified into one of eighteen categories according to type and complexity. This supported detailed planning of lexicographers' workloads and hence, scheduling, as well as adding to the richness of the data. Template entries (Atkins and Rundell, 2008, pp123-128) were developed for 68 lexical sets and for words belonging to these sets, the template was automatically inserted into the draft dictionary, saving lexicographer time and encouraging consistency.

We identified forty syntactic patterns for verbs, eighteen for nouns and eighteen for adjectives. Lexicographers were required to note all the patterns that applied for each word sense.

The lexicographers were all known to the management team beforehand for their high-quality work. They were trained in the dictionary style at two workshops, and their work was thoroughly checked throughout the project, with failings reported back and progress monitored.

A typical short entry is *honeymoon* (shown here in full but for truncated examples). Note the level of detail including senses, subsenses, grammatical structures and collocations. All points are exemplified by one or usually more corpus example sentences. (The style guide, available online, states the conditions for giving one, two or three examples for a phenomenon.)

honeymoon

- *n* holiday after wedding
  Following the wedding day, Jane and . . .
  Upon your return from **honeymoon** *. . .*
  Lee and Zoe left for a **honeymoon** *in . . .*
  SUPPORT VERB spend
    They now live in Cumbernauld after spending . . .
    Their **honeymoon** *was spent at Sandals . . .*
  SUPPORT VERB have
    I hope that you have an absolutely fantastic . . .
    The reception was held at the local pub and . . .
  SUPPORT PREP on
    I have a ring on my left hand which Martha . . .
    The groom whisked the bride off on **honeymoon** *. . .*
    This particular portrait was a festive affair, . . .
  STRUCTURE N_premod
    destination hotel suite holiday night couple
    Classic **honeymoon** *destinations like the . . .*
    We can help and recommend all types of . . .
    We were staying in the **honeymoon** *suite . . .*
    A magical **honeymoon** *holiday in the beautiful . . .*
    Our honeymoon packages offer a wide range of . . .
    It is the favourite of our many **honeymoon** *couples.*
- *v* spend one's honeymoon
  STRUCTURE Particle (locative)
    They'll be **honeymooning** *in Paris (ooh, la la).*
    Mr and Mrs Maunder will **honeymoon** *in . . .*
    The couple spent the early part of their . . .
    A Dave Lister from five years in the future is . . .
- *n* period of grace
  VARIANT FORM **honeymoon period**
    Since his May 1997 landslide election, Blair has . . .
    The UN and Europe were pan national organisations
  CHUNK the honeymoon is over
    VARIANT the honey moon period is over
    The shortest post-election **honeymoon** *is over.*
    Could the **honeymoon** *period be over that quickly?*

public of Ireland and the UK. FnaG is an institution of the two countries.

[2] Lexicography MasterClass had also previously undertaken the planning of the project.

4 Corpus strategy and innovation

The project team combined expertise in corpora, computational linguistics and lexicography, and from the outset the project was to be solidly corpus-based. In the planning phase we had built the NCI: by the time the compilation phase started, in 2007, it was evident not only that the NCI would no longer capture current English, but also that the field had moved on and at 250m words, it was too small. We appended the Irish English data from the NCI to the much larger and newer UKWaC (Ferraresi et al., 2008) and added some contemporary American newspaper text to create the project corpus, which was then pos-tagged with TreeTagger [3] and loaded into the Sketch Engine.

The distinctive feature of the Sketch Engine is 'word sketches': one-page, corpus-driven summaries of a word's grammatical and collocational behaviour. The corpus is parsed and a table of collocations is given for each grammatical relation. For DANTE, the set of grammatical relations was defined to give an exact match to the grammatical patterns that the lexicographers were to record. The same names were used. The word sketch for the word would, in so far as the POS-tagging, parsing, and statistics worked correctly, identify precisely the grammatical patterns and collocations that the lexicographer needed to note in the dictionary.

As is evident, a very large number of corpus sentences needed taking from the corpus into the dictionary. This was streamlined with two processes: GDEX, for sorting the examples so that the 'best' (according to a set of heuristics) are shown to the lexicographer first (Kilgarriff et al., 2008), and 'one-click-copying' of sentences onto the clipboard (including highlighting the nodeword). (In contrast to a finished dictionary, examples were not edited.)

5 XML-based dictionary preparation

The document type definition uses seventy-two elements. It is as restrictive as possible, given that accuracy and then clarity take priority. Lexicographers were not permitted to submit work which did not validate. Wherever there was a fixed range of possible values for an information field, the list was included in the DTD as possible values for an attribute and the lexicographer used menu-selection rather than text-entry.

The database was also used for checking potential problems in a number of ways. For example, there are some word senses where examples are not required, but it is unusual for both senses of a two-or-more-sense word not to need examples, so we routinely used XML searching to check lexicographers' work for any such cases and scrutinised them prior to approval.

6 None of the usual constraints

Most dictionary projects are managed by publishers who are focused on the final (usually print) product, so constraints such as fitting in limited page-space, or using simplified codes to help naive users, or responding to the marketing department, or tailoring the analysis according to the specialist interests of some likely users, or features of the target language (for a bilingual dictionary) usually play a large role in the instructions given to lexicographers. In this project, with the separation of the project team from the publisher, we were unusually free of such compromising factors.

7 Leadership

Many lexicographic projects take years or decades longer than scheduled, and suffer changes of intellectual leadership, or are buffeted by political and economic constraints, all of which produce grave inconsistencies of style, scale and quality between different sections of the data. A consistent lexicon is impossible without consistent and rigorous management. The credentials of the managers are an indicator of the likely quality of the data.

Sue Atkins, the project manager, has been the driving force behind the Collins-Robert English/French Dictionaries (first two editions), the COBUILD project (with John Sinclair), The European Association for Lexicography (with Reinhart Hartmann), the British National Corpus, the Oxford Hachette English/French dictionaries (assisted by Valerie Grundy, DANTE Chief Editor) and with Charles Fillmore, FrameNet. She has co-published the Oxford Guide to Practical Lexicography with Michael Rundell, another of the project management

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

team, who has been Managing Editor of a large number of dictionaries at Longman and Macmillan.

8 Licencing

In the late 1980s it seemed likely that Longman Dictionary of Contemporary English (LDOCE) would have a great impact on NLP. But its star rose, but then promptly fell. As a Longman employee with the task of developing LDOCE use within NLP, the first author investigated the reasons long and hard.

The problem was that NLP groups could not do anything with their LDOCE-based work. They could describe the work in papers, but the work itself was embedded in enhanced versions of LDOCE, or LDOCE-derived resources, and the licence that allowed them to use LDOCE did not alow them to publish or licence or give away any such resource. So LDOCE research, for academics, was a dead end.

A high-quality dictionary represents an investment of millions so one cannot expect its owners to give it away. The challenge then is to arrive at a model for a dictionary's use in which its exploration and enhancement is encouraged, and is not a dead end, and also in which the owner's interest in a return on investment is respected.

DANTE will be made available in a way designed to meet these goals. It will be licenced for NLP research for no fee. The licence will not allow the licencee to pass on the resource, but will include an undertaking from the owner to pass on the licencee's enhanced version to other groups on the same terms (provided it passes quality tests). The owner, or its agent, will also, where possible, integrate and cross-validate enhancements from different users. The owner will retain the right to licence the enhanced data, for a fee, for commercial use. The model is presented fully in (Kilgarriff, 1998).

9 DANTE Disambiguation

'DANTE disambiguation' is a program currently in preparation which takes arbitrary text and, for each content word in the text, identifies the DANTE patterns it matches and thereby assigns it to one of the word's senses in the DANTE database. It is designed to demonstrate the potential that DANTE has for NLP, and to undertake in a systematic way a piece of work that many DANTE users would otherwise need to do themselves: converting as many DANTE data fields as possible into methods which either do or do not match a particular instance of the word. The program will be freely available alongside the database.

Acknowledgments

Thanks to colleagues on the project, particularly the management team of Sue Atkins, Michael Rundell, Valerie Grundy, Diana Rawlinson and Cathal Convery.

References

Sue Atkins and Valerie Grundy. 2006. Lexicographic profiling: an aid to consistency in dictionary entry design. In *Proc. Euralex*, Torino.

Sue Atkins and Michael Rundell. 2008. *Oxford Guide to Practical Lexicography*. OUP, Oxford.

Bran Boguraev and Ted Briscoe, editors. 1989. *Computational lexicography for natural language processing*. Longman, London.

Christy Doran, Dania Egedi, Beth Ann Hockey, B. Srinivas, and Martin Zaidel. 1994. Xtag system: a wide coverage grammar for english. In *Proc. COLING*, pages 922–928.

Christiane Fellbaum, editor. 1998. *WordNet, an electronic lexical database*. MIT Press.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating UKWaC, a very large web-derived corpus of English. In *Proc*ẆAC4, *LREC, Marrakesh*.

Adam Kilgarriff, Milos Husak, Katy McAdam, Michael Rundell, and Pavel Rychly. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*, Barcelona.

Adam Kilgarriff. 1998. Business models for dictioanries and NLP. *Int Jnl Lexicography*, 13(2):107–118.

Adam Kilgarriff. 2006. Efficient corpus development for lexicography: building the new corpus for ireland. *Language Resources and Evaluation Journal*.

Catherine Macleod, Ralph Grishman, and Adam Meyers. 1994. The comlex syntax project: the first year. In *Proc*Ḣuman Language Technology workshop, pages 8–12.

Yorick Wilks, Brian Slator, and Louise Guthrie. 1996. *Electric words: dictionaries, computers, and meanings*. MIT Press, Cambridge, MA, USA.