Automatic Collocation Dictionaries

Adam Kilgarriff, Milos Husák

Lexical Computing Ltd., UK

An Automatic Collocation Dictionary (ACD) is a dictionary listing the collocations for the words of a language, created automatically from a corpus. Other optional components include

* example sentences for the collocations
* the grammatical relation between headword and collocation
* the commonest string that the collocation appears in.

We prepared our first ACD, for English, in 2008, and since then it has been available at <http://forbetterenglish.com>. Recently, as part of the Meta-share initiative and in projects for publishers and institutes, we have produced ACDs for ten further languages (available at <http://www.sketchengine.co.uk> ).

From a technical point of view, an ACD is a set of pre-compiled word sketches (Kilgarriff et al 2004) for a language (with or without GDEX examples (Kilgarriff et al 2008) and commonest strings (Kilgarriff et al 2012)), so from that point of view there is little new to say. However from a user’s point of view – particularly a user who does not know what a corpus is, that is, 99% of them – they are a radically new proposition: a dictionary not written by a person.

Most dictionary users are not linguists, and do not want to be confused by unfamiliar terms like ‘corpus’ in the course of answering their information needs. Yet is will often be a corpus-based report that can best meet their needs. The solution may well be to disguise the corpus as a dictionary.

For those of us who do know what a corpus is, ACDs can be seen as a data point in the space between corpora and dictionaries. It seems likely that a part of the future of lexicography lies in this territory, where user needs are met with responses prepared by lexicographers (where suitable content written by a lexicographer exists), or automatically-generated ones otherwise. The presentation will explore the ACD as one kind of object in the territory between corpus and dictionary.

References

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, David Tugwell 2004. [The Sketch Engine](http://www.kilgarriff.co.uk/Publications/2004-KilgRychlySmrzTugwell-SkEEuralex.rtf)  Proc. Euralex. Lorient, France

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý 2008.  [GDEX: Automatically finding good dictionary examples in a corpus](http://www.kilgarriff.co.uk/Publications/2008-KilgEtAl-euralex-gdex.doc). Proc EURALEX, Barcelona, Spain.

Adam Kilgarriff, Pavel Rychlý, Vojtech Kovar, Vit Baisa 2012. Finding multiwords of more than two words.  Proc. Euralex. Oslo, Norway.