# Tickbox Lexicography

Adam Kilgarriff[1], Vojtěch Kovář[2], Pavel Rychlý[3]
Lexical Computing Ltd, Masaryk University

**Abstract**

Corpus lexicography involves, first, an analysis of a word, and then, copying of collocations and examples from corpus to dictionary. In a large project, there are hundreds of thousands of items to be copied across. Most modern lexicography takes place on a computer, with both corpus and dictionary editor being software applications and the copying done by re-keying or copy-and-paste. This can be inconvenient and time-consuming. We present a more efficient approach, where the user selects collocations and examples by clicking on tickboxes and the material is automatically structured and formatted according to the particular dictionary's requirements, ready for pasting into the dictionary editor.

**Keywords**: corpus lexicography, dictionary editing

## 1. Corpus Lexicography

In corpus lexicography we:

- identify the senses for the word

and then, for each sense:

- identify the key patterns, collocations and phrases
- find example sentences.

This process is the core of the lexicography for a language. Once it has been completed for the full vocabulary, the resulting database is a base analysis of the language which will serve for the development of a range of dictionaries, monolingual and bilingual (where the language analysed is the source language, and the analysis will form the basis whatever the target language) (Atkins 1994; Atkins and Rundell 2008: 97-101).

In a large project, there are hundreds of thousands of items to be copied across from corpus to dictionary editor. Most modern lexicography takes place on a computer, with both corpus and dictionary editor being software applications and the copying done by re-keying or copy-and-paste. This can be inconvenient and time-consuming.

[1] Lexical Computing Ltd, UK, adam@lexmasterclass.com
[2] Masaryk University, Brno, Czech Republic, xkovar3@fi.muni.cz
[3] Masaryk University, Brno, Czech Republic, pary@fi.muni.cz

We present a more efficient approach, where the user selects collocations and examples by clicking on tickboxes and the material is automatically structured and formatted according to the particular dictionary's requirements, ready for pasting into the dictionary editor.

Following a brief note on our corpus application, the Sketch Engine (Kilgarriff *et al.* 2004), we describe TickBox Lexicography (TBL). We then give an account of how it is being used in two large-scale projects, at Macmillan Publishing in the UK and at the Institute for Dutch Lexicology (INL) in the Netherlands.

## 2. The Sketch Engine

The Sketch Engine is a leading corpus query tool, in daily use for lexicography at publishing houses such as Oxford University Press, Cambridge University Press, Collins and Macmillan in the UK, INL in the Netherlands and Cornelsen in Germany, and for language research and teaching at a number of universities worldwide. It operates as a ready-to-use online service, with large corpora available to all customers for most of the world's major languages.

There are two functions which the Sketch Engine offers which support the process:

- 'word sketches', one-page summaries of the key collocations (and sometimes phrases) for the word, in a table organised by grammatical relations (*e.g. object, modifier, modified*) (cf. Figure 1).
- the 'Good Dictionary Example eXtractor', GDEX, a function for finding good dictionary examples (Kilgarriff *et al.* 2008)

GDEX is far from perfect and we cannot assume that the 'best example' according to GDEX is good enough to go straight into a printed dictionary. But it does greatly improve the chances that the lexicographer will find a useable sentence (often needing some further editing) amongst the first few concordance lines for a collocation.

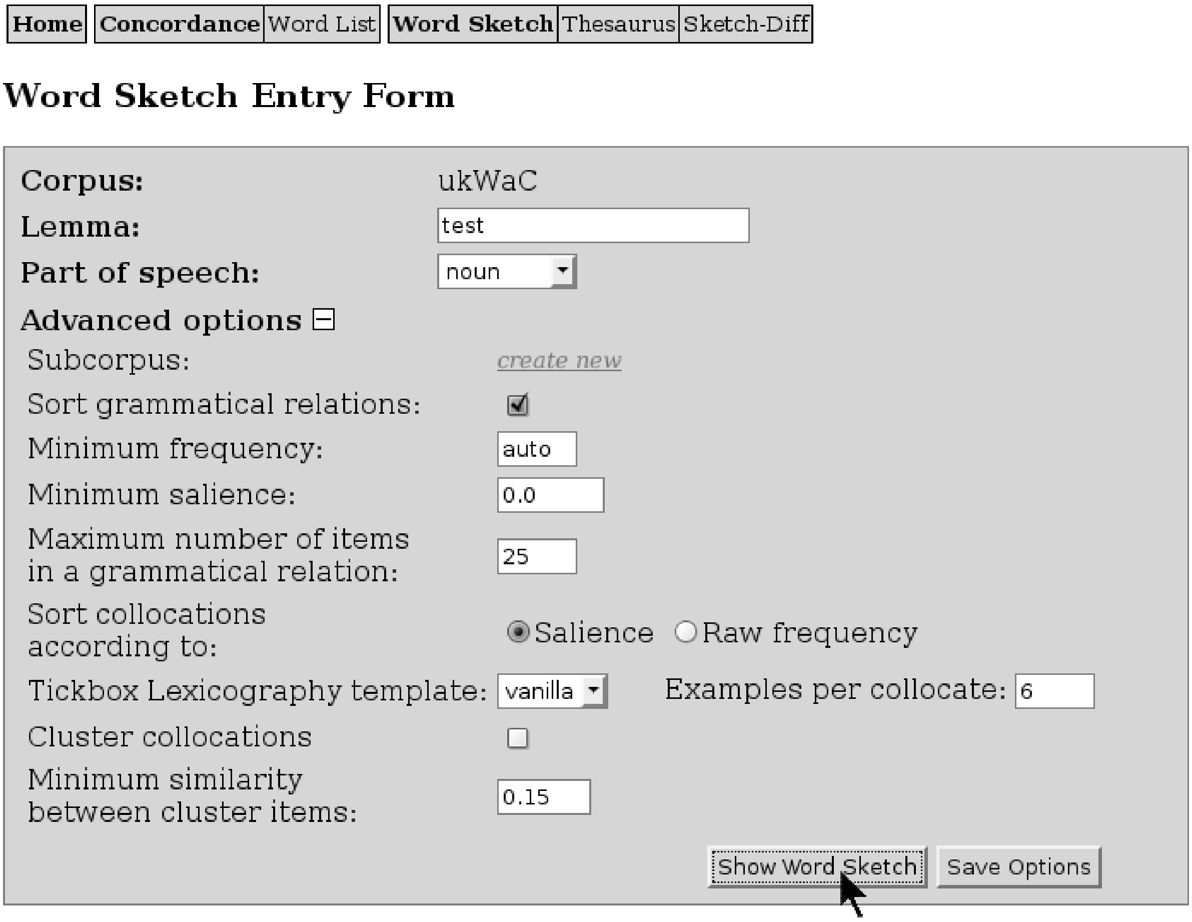
## 3. Tickbox Lexicography

Tickbox Lexicography is a variety of corpus lexicography with intensive computational support in which the lexicographer selects aspects of the automated analysis of the word for inclusion in the dictionary by ticking boxes, and then pastes their selections into the editing interface.

The process is as follows:

- the lexicographer sees a version of the word sketch with tickboxes beside each collocation.
- for each sense and each grammatical relation, they tick the collocations they want in the dictionary (see Figure 2).
- they click a 'next' button

- they then see, for each collocation they have ticked, a choice of six (by default) corpus example sentences, chosen by GDEX, each with a tickbox beside it: they tick the ones they like (see Figure 3).
- they tick a "copy to clipboard" button.



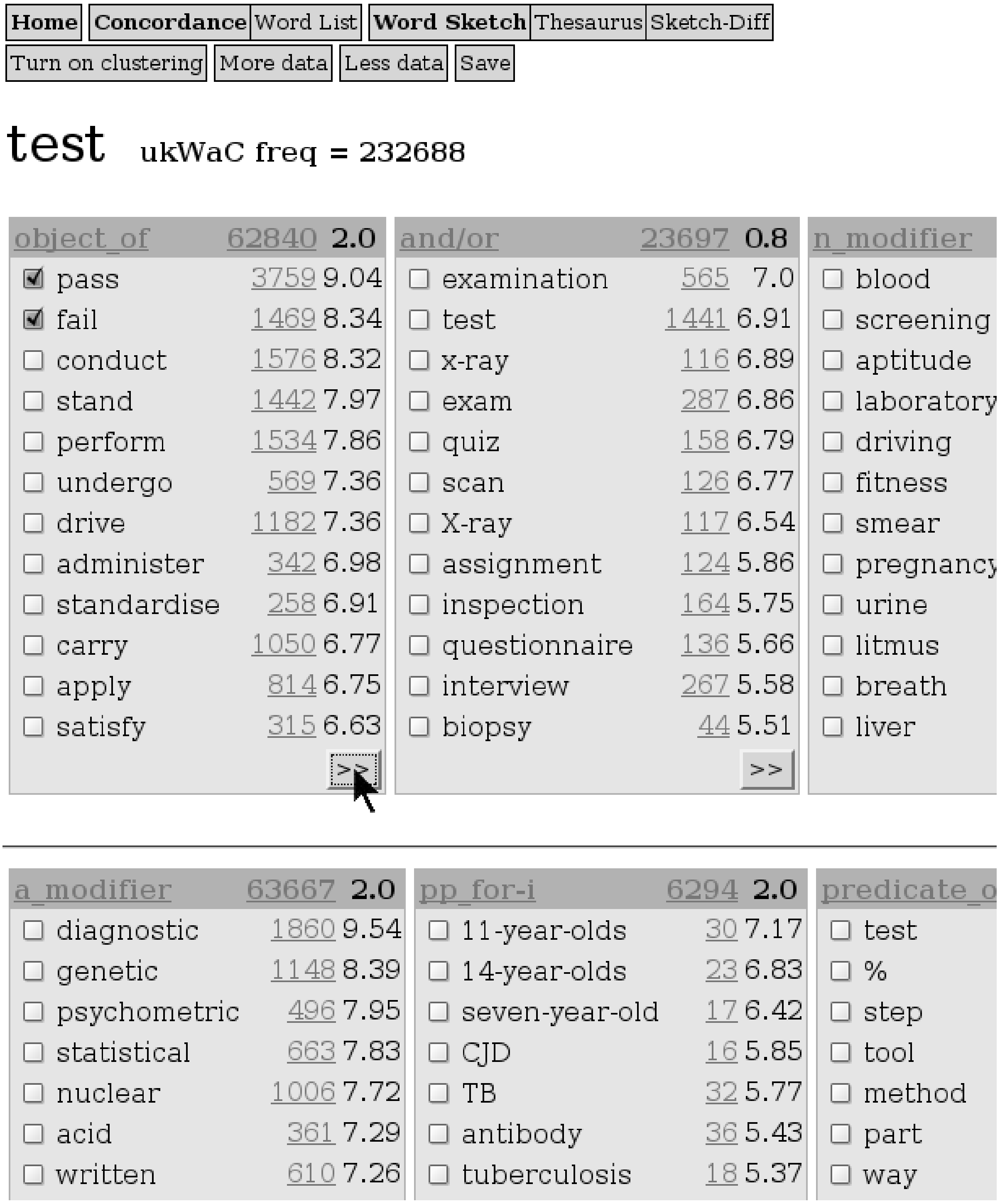
Word Sketch Entry Form

Corpus: ukWaC
Lemma: test
Part of speech: noun
Advanced options ⊟
Subcorpus: create new
Sort grammatical relations: ☑
Minimum frequency: auto
Minimum salience: 0.0
Maximum number of items in a grammatical relation: 25
Sort collocations according to: ● Salience ○ Raw frequency
Tickbox Lexicography template: vanilla Examples per collocate: 6
Cluster collocations □
Minimum similarity between cluster items: 0.15

*Figure 1. Word Sketch Entry Form*

Each target dictionary has its own TBL application, based on the DTD (for XML-based systems) or field names used in the dictionary. The system then copies the collocations and examples, embedded in an XML structure as required by the user's dictionary-editing system and target dictionary, onto the clipboard. (The XML fragment for the example above, with a 'vanilla' DTD, is shown in Figure 4.) The lexicographer can then paste the structure into the dictionary editing system. For this we use the operating system's clipboard functions. While this is not, in computer science terms, the most elegant technique, we have found it to be the most convenient, and widely-applicable method for transferring data between programs on the same computer. It has the great advantage that all users know and understand it.

Thus, TBL models and streamlines the process of getting corpus data out of the corpus system and into the dictionary editing system.



test ukWaC freq = 232688

| object_of | 62840 | 2.0 |
|---|---|---|
| ☑ pass | 3759 | 9.04 |
| ☑ fail | 1469 | 8.34 |
| □ conduct | 1576 | 8.32 |
| □ stand | 1442 | 7.97 |
| □ perform | 1534 | 7.86 |
| □ undergo | 569 | 7.36 |
| □ drive | 1182 | 7.36 |
| □ administer | 342 | 6.98 |
| □ standardise | 258 | 6.91 |
| □ carry | 1050 | 6.77 |
| □ apply | 814 | 6.75 |
| □ satisfy | 315 | 6.63 |

| and/or | 23697 | 0.8 |
|---|---|---|
| □ examination | 565 | 7.0 |
| □ test | 1441 | 6.91 |
| □ x-ray | 116 | 6.89 |
| □ exam | 287 | 6.86 |
| □ quiz | 158 | 6.79 |
| □ scan | 126 | 6.77 |
| □ X-ray | 117 | 6.54 |
| □ assignment | 124 | 5.86 |
| □ inspection | 164 | 5.75 |
| □ questionnaire | 136 | 5.66 |
| □ interview | 267 | 5.58 |
| □ biopsy | 44 | 5.51 |

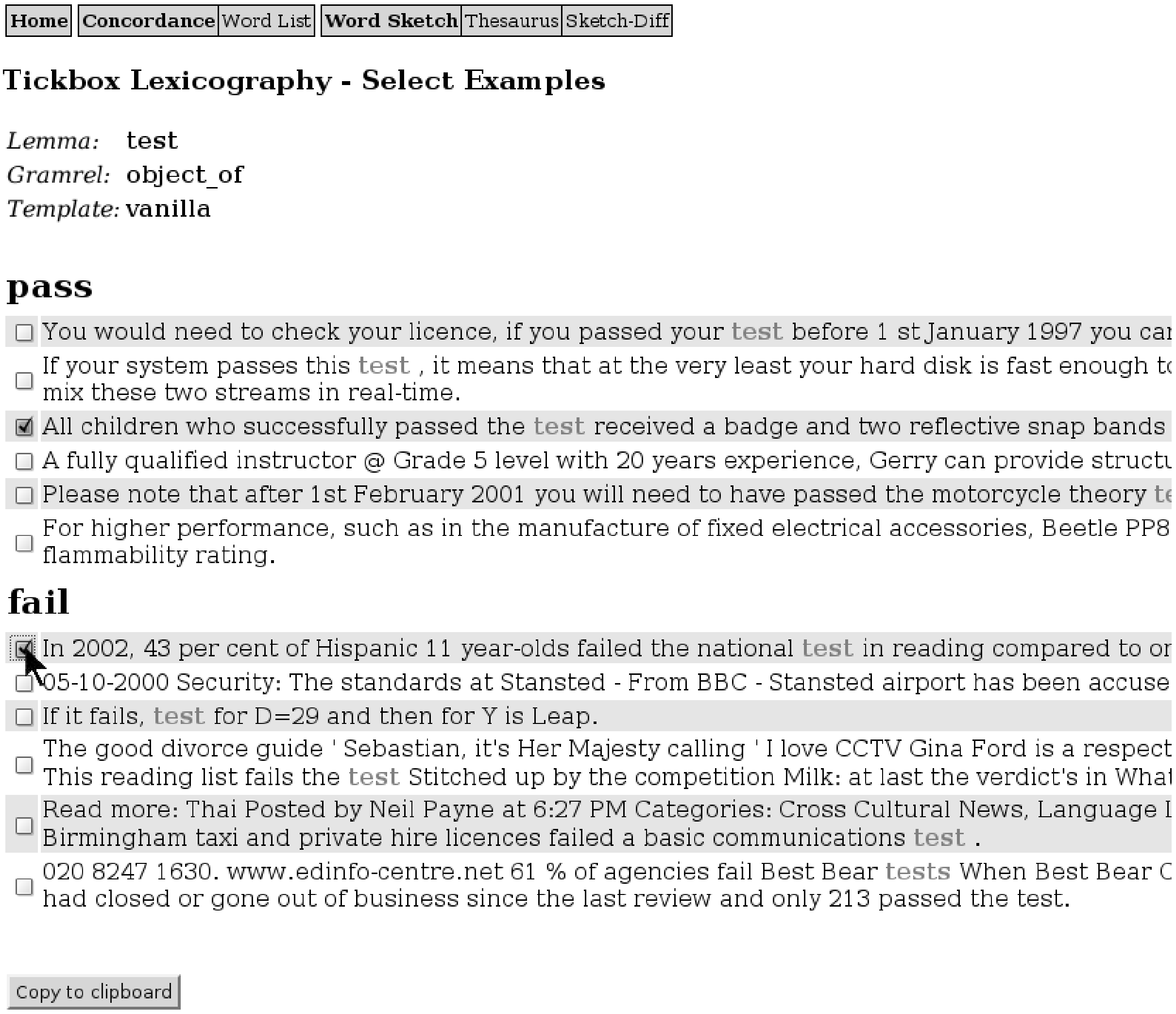
| n_modifier |
|---|
| □ blood |
| □ screening |
| □ aptitude |
| □ laboratory |
| □ driving |
| □ fitness |
| □ smear |
| □ pregnancy |
| □ urine |
| □ litmus |
| □ breath |
| □ liver |

| a_modifier | 63667 | 2.0 |
|---|---|---|
| □ diagnostic | 1860 | 9.54 |
| □ genetic | 1148 | 8.39 |
| □ psychometric | 496 | 7.95 |
| □ statistical | 663 | 7.83 |
| □ nuclear | 1006 | 7.72 |
| □ acid | 361 | 7.29 |
| □ written | 610 | 7.26 |

| pp_for-i | 6294 | 2.0 |
|---|---|---|
| □ 11-year-olds | 30 | 7.17 |
| □ 14-year-olds | 23 | 6.83 |
| □ seven-year-old | 17 | 6.42 |
| □ CJD | 16 | 5.85 |
| □ TB | 32 | 5.77 |
| □ antibody | 36 | 5.43 |
| □ tuberculosis | 18 | 5.37 |

| predicate_o |
|---|
| □ test |
| □ % |
| □ step |
| □ tool |
| □ method |
| □ part |
| □ way |

*Figure 2. Word sketch for the English noun* ‘test’, *data from the UKWaC corpus*



Tickbox Lexicography - Select Examples

*Lemma:* test
*Gramrel:* object_of
*Template:* vanilla

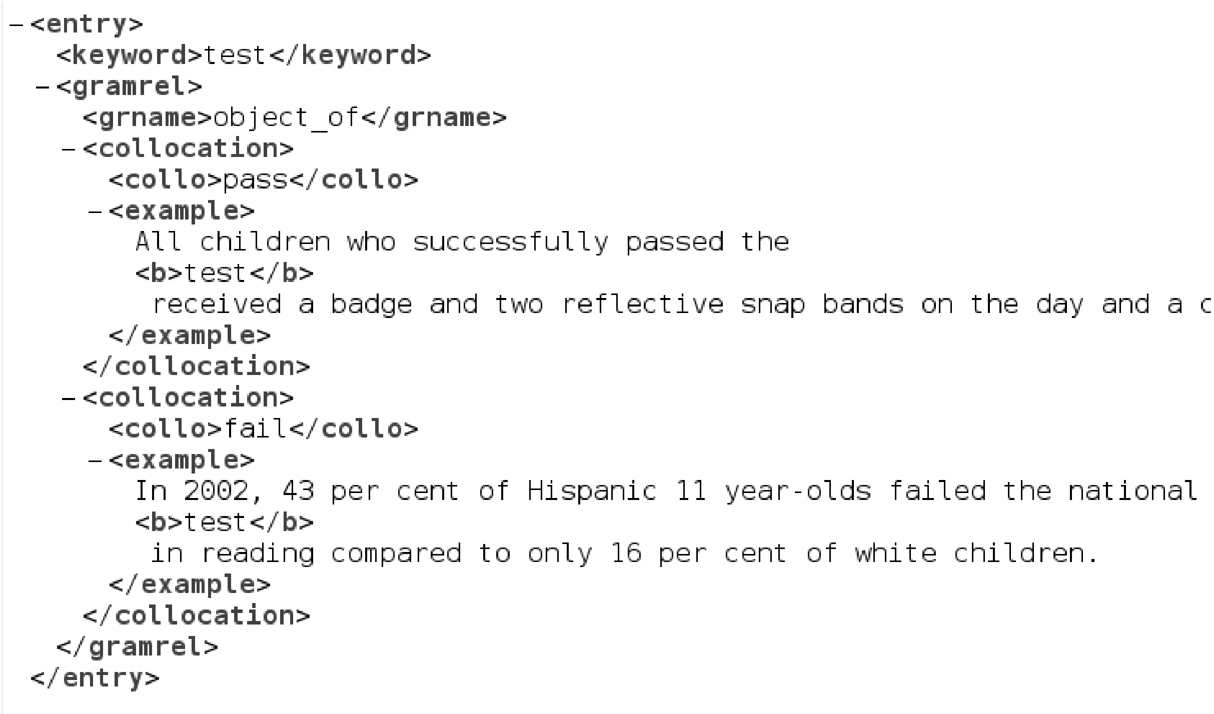
pass

- □ You would need to check your licence, if you passed your test before 1 st January 1997 you car
- □ If your system passes this test , it means that at the very least your hard disk is fast enough to mix these two streams in real-time.
- ☑ All children who successfully passed the test received a badge and two reflective snap bands
- □ A fully qualified instructor @ Grade 5 level with 20 years experience, Gerry can provide structu
- □ Please note that after 1st February 2001 you will need to have passed the motorcycle theory te
- □ For higher performance, such as in the manufacture of fixed electrical accessories, Beetle PP8 flammability rating.

fail

- ☑ In 2002, 43 per cent of Hispanic 11 year-olds failed the national test in reading compared to or
- □ 05-10-2000 Security: The standards at Stansted - From BBC - Stansted airport has been accuse
- □ If it fails, test for D=29 and then for Y is Leap.
- □ The good divorce guide ' Sebastian, it's Her Majesty calling ' I love CCTV Gina Ford is a respect This reading list fails the test Stitched up by the competition Milk: at last the verdict's in What
- □ Read more: Thai Posted by Neil Payne at 6:27 PM Categories: Cross Cultural News, Language I Birmingham taxi and private hire licences failed a basic communications test .
- □ 020 8247 1630. www.edinfo-centre.net 61 % of agencies fail Best Bear tests When Best Bear C had closed or gone out of business since the last review and only 213 passed the test.

*Figure 3. GDEX examples for each selected collocate.*



```
-<entry>
   <keyword>test</keyword>
  -<gramrel>
     <grname>object_of</grname>
    -<collocation>
       <collo>pass</collo>
      -<example>
         All children who successfully passed the
         <b>test</b>
          received a badge and two reflective snap bands on the day and a c
       </example>
     </collocation>
    -<collocation>
       <collo>fail</collo>
      -<example>
         In 2002, 43 per cent of Hispanic 11 year-olds failed the national
         <b>test</b>
          in reading compared to only 16 per cent of white children.
       </example>
     </collocation>
   </gramrel>
 </entry>
```

*Figure 4. An XML entry draft, as copied to the clipboard*

## 4. Projects

At time of writing there are two large-scale dictionary projects where TBL is in daily use: the *Macmillan Collocations Dictionary* and the *Algemeen Nederlands Woordenboek.*

### *4.1 Macmillan Collocations Dictionary*

At Macmillan Publishing, the *Macmillan Collocations Dictionary* (MCD) is in preparation. MCD starts from MEDAL (2007), and provides a full account of the collocations of the core senses of around 4,000 common and highly ‘collocational’ words (Kilgarriff 2006). As in word sketches (and in other collocations dictionaries such as Oxford’s (OCD 2002, 2009), collocations are organised according to the grammatical relations. Some collocations are illustrated with examples in the paper book; all have examples available by mouse-click in online and other electronic versions.

To set up TBL for MCD, we first developed customised word sketches in which the grammatical relations were those to be used in MCD. This required work on the underlying part-of-speech tagging and grammatical-relation-finding software. (The parsing, to identify grammatical relations, uses regular expressions over part-of-speech tags and is built in to the Sketch Engine: see Kilgarriff *et al.* 2004.) GDEX was also customised, with the incorporation of a long list of ‘stop’ words, to minimise the chances that GDEX would select examples containing offensive material.

In the first trials, lexicographers selected all the example sentences (typically six per collocate) that were to be used in the electronic version of MCD, but this proved too slow. We changed to a strategy where only the examples which are to appear in the book are selected by lexicographers. For all others, GDEX will be trusted to deliver good examples. (The manually-selected items will be edited as necessary by lexicographers, whereas the others will be full and unedited corpus sentences.) These sentences will be selected in a batch process after the main phase of the lexicography is complete, as this will reduce the volume of data to be handled by the clipboard and the dictionary editing system, and will allow us to use a new version of GDEX. We anticipate using the experience gathered during the project to fine-tune GDEX according to Macmillan’s preferences and observations, and this can only take place at or near the end of the project.

MEDAL 2007 already contains 1000 ‘collocation boxes’ for word senses of common words, with collocations classified according to grammatical relations, and further collocations in bold in regular entries. It was desirable to carry them across into MCD, in a way which integrated with MCD lexicography. To this end we:

- analysed MEDAL to find all collocations, either in collocation boxes or shown in bold within regular entries
- identified the grammatical relation they stood in to the headword
- checked to see if they were already in the word sketch:
    - if they were (as they usually were), colour them red (in the word sketch) and pre-tick the tickbox, as they will almost always be wanted in MCD
    - if they were not, add them in (in red), with links to their corpus instances and pre-ticked tickboxes.

The dictionary editing software used for MCD accepts XML pasted from the clipboard. This means that, once the lexicographer has

- called up the customised word sketch for the headword,
- selected the grammatical relation,
- selected collocates,
- selected examples for the paper dictionary,

… they click a ‘copy to clipboard’ button, and then paste the material (using standard CTRL-V) into the dictionary entry.

### *4.2 Algemeen Nederlands Woordenboek (ANW)*

At the Institute for Dutch Lexicology (INL), the *Algemeen Nederlands Woordenboek* (General Dutch Dictionary, ANW) is a large Dutch dictionary project running from

2001 til 2018. The project has been using the Sketch Engine for corpus access since 2007 (for background and a full account, see Tiberius and Kilgarriff 2009).

Within the ANW dictionary project, example sentences are gathered together with bibliographic information for each citation.

TBL offers an architecture for efficiently collecting information from the corpus and packaging it for insertion into the dictionary database. While the system had been designed with linguistic information in mind, it was readily adjusted for the ANW dictionary project to gather and insert bibliographic information as well. We developed a TBL installation with a specific template for INL where, when the lexicographer ticks the example, not only the example but also the title, author, publisher and date of its source are assembled in an XML fragment and placed on the clipboard. (We also made it possible to select multiple examples at a time, as this fitted the way that INL lexicographers worked.) The dictionary editing software (Niestadt 2009) was customised to interpret these XML structures so, when the user pastes the example from the clipboard into the editor, the different components of the reference are placed in the appropriate database fields with a single mouse-click.

## 5. Conclusion

We have shown how TBL works in the general case, and how it has been used in two large projects. We believe TBL has great potential for both streamlining corpus lexicography and making it more accountable to the corpus.

## References

ATKINS, B. T. S. (1994) A corpus-based dictionary. In *Oxford-Hachette English-French Dictionary* (Introductory section). Oxford: Oxford University Press: xix – xxxii.

ATKINS, B. T. S. and RUNDELL, M. (2008) *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

KILGARRIFF, A. (2006). Collocationality (and how to measure it). In *Euralex Proceedings*. Torino, Italy.

KILGARRIFF, A., RYCHLÝ, P., SMRŽ, P. and TUGWELL, D. (2004) The Sketch Engine. In *Euralex Proceedings*. Lorient, France, July: 105-116.

KILGARRIFF, A., HUSÁK, M., MCADAM, K., RUNDELL, M. and RYCHLÝ, P. (2008) GDEX: Automatically finding good dictionary examples in a corpus. In *Euralex Proceedings*, Barcelona,

MEDAL (2007) *Macmillan English Dictionary for Advanced Learners*. Second edition. Edited by M. Rundell. Oxford.

NIESTADT, J. (2009). De ANW-artikeleditor: software als strategie. In E. Beijk et al. (eds), *Fons Verborum: Feestbundel Fons Moerdijk*. Amsterdam: Gopher: 215-222.

OCD (2009 [2002]) *Oxford Collocations Dictionary*. Oxford.

TIBERIUS, C. and KILGARRIFF, A. (2009). The Sketch Engine for Dutch with the ANW corpus. In E. Beijk et al. (eds), *Fons Verborum: Feestbundel Fons Moerdijk*. Amsterdam: Gopher BV. : 237-255.