A Quantitative Evaluation of Word Sketches

Adam Kilgarriff, Vojtěch Kovář, Simon Krek, Irena Srdanovic, Carole Tiberius
Lexical Computing Ltd., Brighton, UK; Masaryk University, Brno, Czech Republic;
Trojina, Ljubljana, Slovenia; Tokyo Institute of Technology, Tokyo, Japan;
Instituut voor Nederlandse Lexicologie, Leiden, Netherlands

A word sketch is an automatic corpus-derived summary of a word's grammatical and collocational behaviour. Word sketches were first prepared in 1999 for the compilation of the Macmillan English Dictionary for Advanced Learners *(Rundell 2002). They have since been integrated into the Sketch Engine corpus query tool (Kilgarriff et al. 2004), prepared for fifteen languages, and used on a large scale for lexicography by a number of publishers. We are frequently told how impressive they are and how little they miss - but we would like a more rigorous assessment.*
We describe a formal, quantitative evaluation of word sketches, from a user perspective, for four languages (Dutch, English, Slovene, Japanese), with the critical question being 'is the collocation suitable for inclusion in a published collocation dictionary'. For each language, we inspected twenty collocates for each of forty-two headwords. In each case two thirds or more of the collocations were of publishable quality.

1. Introduction

A word sketch is an automatic corpus-derived summary of a word's grammatical and collocational behaviour. Word sketches were first prepared in 1999 for the compilation of the *Macmillan English Dictionary for Advanced Learners* (Rundell 2002). They have since been integrated into the Sketch Engine corpus query tool (Kilgarriff et al. 2004), prepared for fifteen languages, and used on a large scale for lexicography by a number of publishers. A simplified word sketch for the English noun *flour* is:

flour *noun*

OBJECT_OF sift sieve grind mix add raise produce put
ADJ_MODIFIER self-raising wholemeal seasoned plain white organic fine strong
NOUN_MODIFIER wheat soya tbsp maize corn rice bread cup
MODIFIES tortilla milling mill mixture
AND/OR salt butter sugar flour cook rice bread cereal egg wheat grain powder
PP_INTO bowl

How good are they? We are frequently told how impressive they are and how little they miss - but we would like a more rigorous assessment. In this paper we present a formal evaluation for Dutch, English, Japanese and Slovene.

2. Developer and user evaluation

The nature of an evaluation depends on whose interests it is serving. Two different interested parties are system developers and potential users. Developers see the point of evaluation as 'making the system better': they need to be able to work out, if they change one method or module or resource, does it improve performance? But for potential users – here, dictionary publishers – the point is to work out whether the whole system can help their enterprise. The evaluation needs to tell them if the system is good enough to help their task: here, making dictionaries.

Whereas developer's evaluation picks apart the distinct components and treats each separately, a user views the system as a whole, to be evaluated overall: they may or may not have a practical option of changing components. The critical question for the customer is:

how well does the whole system performs once the developers have set up all components as well as possible?

To prepare word sketches, we need:

- A corpus
- NLP tools: tokeniser, lemmatiser and part-of-speech tagger
- A sketch grammar for the language
- Statistics to select salient collocations

A developer evaluation will assess each of these separately, whereas a user evaluation will take the best available of each, and assess the outcome. In this paper we take a 'user evaluation' perspective. (See Ivanova et al. 2008 for a developer evaluation.)

A word sketch aims to present a full and complete account of a word's grammatical and collocational behaviour. As a reference point for what this might mean, we propose the Oxford Collocations Dictionary (OCD, 2007). The OCD was compiled by lexicographers studying corpus evidence but without using word sketches. It is a high-quality product which 'shows all the words that are commonly used in combination with each headword: nouns, verbs, adjectives, adverbs, and prepositions as well as common phrases.'[1] Here is the entry for *flour* (with example sentences stripped out, as they do not form part of this evaluation).

flour *noun*

ADJ	**strong | plain, self-raising white, wholemeal | stone-ground | unbleached | rice, rye, wheat, etc.**
QUANT	**bag, packet, sack**
VERB + FLOUR	**use | add, blend, fold in, mix (in), rub sth in/into, stir (in) | sieve, sift**
FLOUR + NOUN	**mill**

OCD serves as a model for what we wish to produce automatically. Our goal for what word sketches aim to do is provide a grammatically-organised list of collocates which would form a suitable entry for a collocations dictionary such as OCD.

For English, there exists, in OCD, a modern collocation dictionary which serves as a model. For other languages, we have no guarantee that any such resource exists. Also we note that:

- OCD has its own anomalies and omissions.
- OCD has a 'phrases' category where items are often included on grounds of non-transparency of meaning: this is beyond our remit
- Grammatical category names are not well-matched, and the categories themselves show some differences
- Both word sketches and OCD entries include some collocations of more than two words, which raise a number of issues where OCD and word sketch policies are not aligned
- OCD distinguishes synonymous collocates (separated by comma) with non-synonymous ones (separated by vertical bar). While the Sketch Engine can produce word sketches organised in this way, this is outside the scope of this exercise.

We use OCD as a model for what we are aiming to do but do not use it at a practical level.

[1] Promotional material on OUP website, http://www.oup.com/elt/catalogue/isbn/0-19-431243-7?cc=global, 9 Jan 2008.

3. Precision and Recall

Across the information sciences, when considering evaluation we must distinguish precision and recall. Precision is the percentage of the answers given, that are correct. Recall is the percentage of all correct answers, that are found. If my word sketch for *flour* contains only *sift* and *sieve*, it has 100% precision, since all the collocates given are correct, but low recall, since there are many other collocates it does not give. As a response gets bigger, precision usually falls off (since some incorrect answers creep in) but recall improves (as more of the correct answers are included). Changing the size of the answer is a matter of adjusting the 'precision/recall tradeoff'.

Typically, recall is harder to measure than precision: we can measure precision simply by examining the responses we have, but to assess recall we have to consider all the other answers which are correct but which were not given. They are frequently not readily available.

We calculate precision as follows. For a sample of dictionary entries, for each collocate in the word sketch we ask a human expert: would the collocation have been suitable for including at this entry in a dictionary like OCD?

To assess recall we need the expert to examine enough potential collocates for each word, so that we are confident all actual collocates are among them. We are currently addressing this issue.

4. 3+-word-collocations, grammar, word sketch size

The question arose: if the system lists a word which only collocates with the headword within a three-or-more-word unit, as *put* is a collocate[2] for *cat* only in the context of *out* ('put the cat out'), is the collocate good or bad? Our decision was to treat it as good, as it is enough to signal to a lexicographer that there is a collocation to be included in a collocation dictionary, even if the system has not found all of it. But it was not a decision that human evaluators were comfortable with.

We need to determine which grammatical relations to include in the evaluation. Different relations raise different issues for different languages. Verb-object pairs, for example, are substantially harder to find in German than in English. The and-or relation, which lexicographers find useful for spotting distinct meanings, is not in the repertoire of relations covered by OCD or other typical taxonomies of collocations. There is merit to the notion of assessing each grammatical relation separately, though this makes the exercise larger and does not so easily support comparisons across languages or an overview of the success of the system. Here, we look globally across all relations which are handled both in word sketches and in collocations dictionaries.

The word sketch identifies both collocates and their grammatical relation to the headword and sometimes the collocate will be valid, but the grammatical relation incorrect. The expert should be able to note 'valid but misclassified.'

[2] We use *collocate* to refer to the word that joins with the headword to form a *collocation.* For any headword, a list of its collocates is a list of the words that it combines with to give its collocations.

Another question is 'how big should the word sketch be?' As a practical matter, we must keep word sketches quite small, as we are expecting our human experts to make a judgement on each collocate, and their time is limited. (The larger the word sketch, the more we are in a position to assess recall as well as precision.)

We would like to define 'word sketch size' in a way that is comparable across the different languages, and this requires that it is set in a simple way. We shall simply say: the word sketch will contain the twenty best collocates, according to the salience statistic (subject to the constraint that no more than two thirds relate to any single grammatical relation).[3]

5. Sampling

We wish to evaluate word sketches for 'words in general' but what words are these? Word sketches are designed for the core of the vocabulary: not the very rare words, or the grammatical words, but the common nouns, verbs and adjectives that make up 99% of the headword list in a standard dictionary, in a ration of roughly 2:1:1.[4] (Adverbs are a far smaller category, usually accounting for less than 1% of dictionary headword lists.) OCD has collocations for 9000 headwords, but that seems a modest number. Intermediate-level learners' dictionaries typically have around 30,000 headwords.

We take a sample from the 30,000 commonest nouns, verbs and adjectives in the corpus, with the sample structured as in Table 1. Within these constraints, the sampling was random. Table 1 also shows the words selected for English.

	Noun	Verb	Adj	Totals
Common (top 2999)	6	4	4	14
	space solution opinion mass corporation leader	serve incorporate mix desire	high detailed open academic	
Mid (3000-9999)	6	4	4	14
	cattle repayment fundraising elder biologist sanitation	grieve classify ascertain implant	adjacent eldest prolific ill	
Low (10,000-30,000)	6	4	4	14
	predicament adulterer bake bombshell candy shellfish	slap outgrow plow traipse	neoclassical votive adulterous expandable	
Totals	18	12	12	42

Table 1. Lexical sample structure, also showing actual words used for English.

6. The corpora and the NLP tools

The quality of the word sketch depends on the quality and size of the corpus, the tokeniser (specially for languages which do not insert spaces between words, like Japanese), lemmatiser, POS-tagger, grammar and statistic. The evaluation implicitly evaluates all components.

[3] Twenty is a reasonable number of collocates to present in a collocation dictionary for a high-frequency word, but a high number for medium and low-frequency words. It might have been better to vary the number of collocates with the frequency band.

[4] For Japanese there were substantially fewer adjectives and verbs, which was due to the way in which the Japanese tagset ChaSen includes under the noun tag adjectives in *–na*, being formed from nouns (for example 'genki', *vigour, health* & 'genki-na', *healthy*, *well*) and *suru* verbs, being formed from nouns (for example, 'kekkon', *marriage* & 'kekkon suru', *to get married*.

The same statistic, based on the Dice coefficient, was used throughout.[5] The grammars were written by the authors of this paper and colleagues. The corpus and NLP tools used for each language were:

Dutch: The ANW corpus, a 102-million word corpus of contemporary Dutch which has been under development at the Instituut voor Nederlandse Lexicologie (INL) for a number of years, tagged and lemmatised by in-house tools.

English: UKWaC, a large web corpus (Ferraresi 2008), tagged by TreeTagger.[6]

Japanese: JpWaC, a 400-million word web corpus (Srdanovic, Erjavec and Kilgarriff 2008), tokenised and POS-tagged using the ChaSen toolset.[7]

Slovene: FidaPLUS (Arhar et al. 2007, Arhar 2008), a 620-million word reference corpus with texts of different genres spanning from 1990 to 2006, POS-tagged and lemmatised with a toolset developed by the Amebis software company.

7. Evaluation practicalities

We prepared a customised version of the Sketch Engine in which word sketches contained only the twenty highest-scoring collocates for each word, and in which each collocate was associated with a menu with the following items:

- Good
- Good but wrong grammatical relation or POS-tagging error
- Maybe (not striking collocate)
- Maybe (specialised vocab)
- Bad

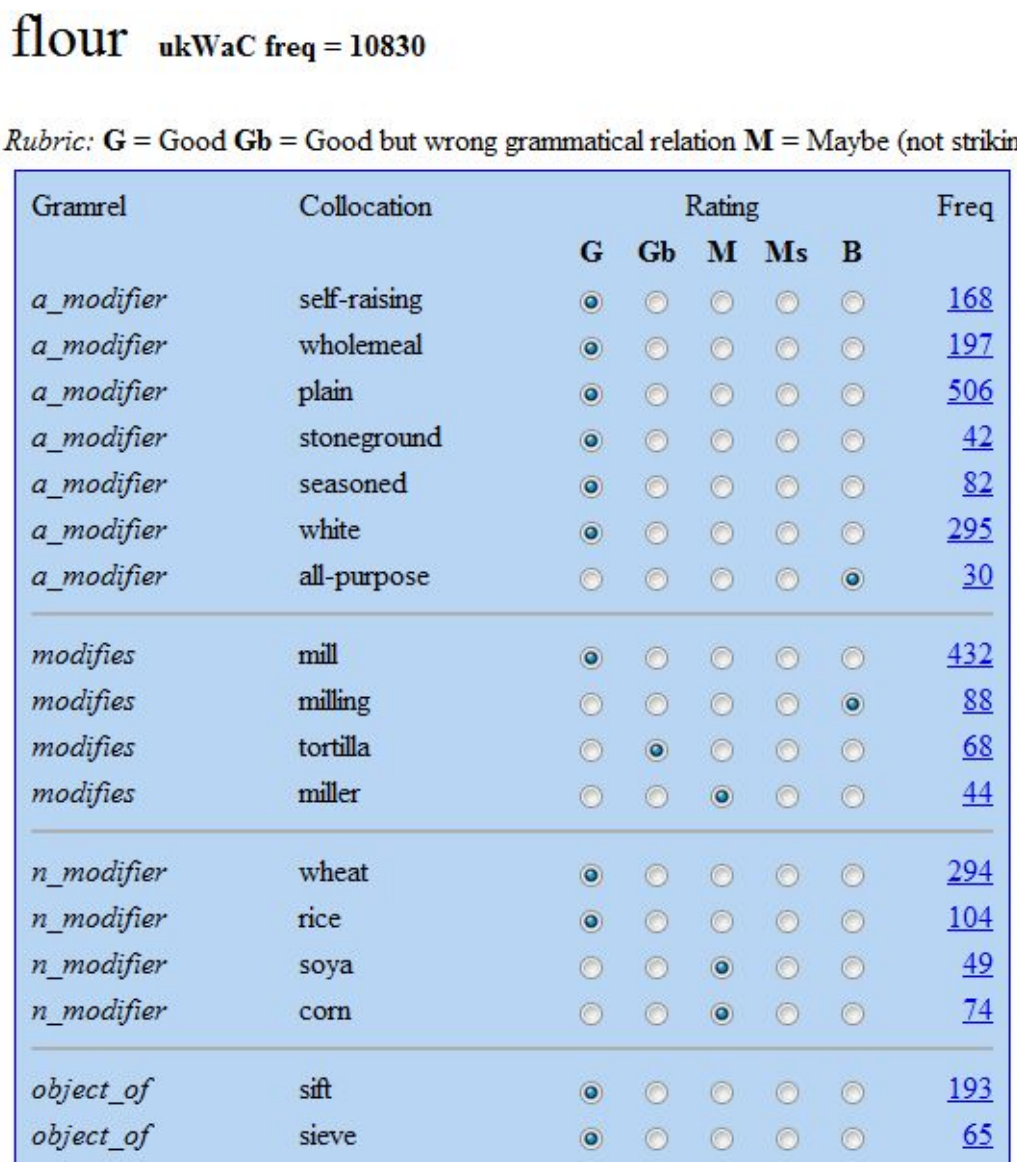
For Dutch and English two, and for Japanese and Slovene three, linguists and lexicographers assessed each collocate. In order to rule out 'unclear' data, we distinguish those instances where all evaluators agree from those where they disagree, and base our results only on the agreement cases. We noted that agreement on the boolean decision, ‘good or not good’ was substantially higher than agreement on finer-grained categories, so we merged ‘Good’ and ‘Good-but’ as ‘good’ and all other categories as ‘bad’.

A screenshot of the interface is shown in Figure 1. Evaluators selected the relevant item from the menu. Choices were stored in a database.

[5] See *Statistics used in the Sketch Engine* on the Sketch Engine website, http://www.sketchengine.co.uk.

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

[7] http://chasen.naist.jp.



flour **ukWaC freq = 10830**

Rubric: **G** = Good **Gb** = Good but wrong grammatical relation **M** = Maybe (not striking collo

Gramrel	Collocation	Rating					Freq
		G	Gb	M	Ms	B	
a_modifier	self-raising	◉	○	○	○	○	168
a_modifier	wholemeal	◉	○	○	○	○	197
a_modifier	plain	◉	○	○	○	○	506
a_modifier	stoneground	◉	○	○	○	○	42
a_modifier	seasoned	◉	○	○	○	○	82
a_modifier	white	◉	○	○	○	○	295
a_modifier	all-purpose	○	○	○	○	◉	30
modifies	mill	◉	○	○	○	○	432
modifies	milling	○	○	○	○	◉	88
modifies	tortilla	○	◉	○	○	○	68
modifies	miller	○	○	◉	○	○	44
n_modifier	wheat	◉	○	○	○	○	294
n_modifier	rice	◉	○	○	○	○	104
n_modifier	soya	○	○	◉	○	○	49
n_modifier	corn	○	○	◉	○	○	74
object_of	sift	◉	○	○	○	○	193
object_of	sieve	◉	○	○	○	○	65

Figure 1. Screenshot of word sketch, evaluators' interface.

8. Results

In the Table 2 we present, for each language, the total number of collocations assessed by all assessors (which is in each case slightly less than the maximum possible of 20 collocates for each of 42 headwords, owing to a range of minor anomalies and omissions), and the number for which all evaluators agreed, and for these, the number that were good and the number that were bad (where 'bad' includes 'maybe').

Language	Total colls assessed	Evaluators all agreed on	**Good**	**Bad**	**% good**
Dutch	782	501	332	169	66.3
English	794	519	367	152	70.7
Japanese	747	294 (690)	278 (600)	16 (90)	94.6 (87.0)
Slovene	800	550	391	159	71.1

Table 2. Evaluation results by language.

For all languages, two thirds or more of the collocations on which the assessors agreed were of publishable quality.[8]

9. Discussion

Sources of bad collocates were POS-tagging and lemmatisation errors, duplication in the corpora, and corpus 'junk' (including, for English, some tracts of text generated by computer for purposes of advertising poker; our efforts to find and remove all such material have not yet been entirely successful). Multi-word items were a recurring concern, as it did not seem natural to the evaluators to mark *cat* as good at *put* when the word sketch gave no indication that *out* was also needed: this was the evaluators' most often-voiced concern. Some grammatical relations performed better than others (and, for Slovene, the exercise has already led to changes in the sketch grammar).

10. Conclusion and further work

We have undertaken a formal evaluation of word sketches, from a user perspective, for four languages, with the critical question being ‘is the collocation suitable for inclusion in a published collocation dictionary’. For each language, we inspected twenty collocates for each of forty-two headwords. In each case two thirds or more of the collocations were of publishable quality.

We are currently pursuing a developer-oriented variant of the evaluation paradigm (which will support an assessment of recall as well as precision). This will allow us to comparatively evaluate corpora, POS-taggers and other NLP tools, sketch grammars and salience statistics, since, if all else remains the same, then we can say that, for collocation-dictionary-extraction purposes, the corpus or tagger or grammar or statistic that gives rise to the higher-scoring sketches is the better one.

[8] For Japanese there was three-way agreement among lexicographers for less than half of the data, so we also give figures for two-out-of-three agreement, in brackets.

References

Arhar, Špela; Gorjanc, Vojko; Krek, Simon. (2007). 'FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools'. In: *Proceedings Corpus Linguistics*. Birmingham, UK.

Arhar, Špela. (2006). 'FidaPLUS : the upgrade of the Slovene reference corpus'. In: *Wörterbuch und Übersetzung / 4. Internationales Kolloquium zur Lexikographie und Wörterbuchforschung, Universität Maribor.* Hildesheim/Zürich/New York: Georg Olms, 2008.

Ferraresi, A.; Zanchetta, E.; Bernardini, S.; Baroni, M. (2008). 'Introducing and evaluating ukWaC, a very large web-derived corpus of English'. In: Pro*ceedings 4th WAC workshop LREC*. Marrakech.

Ivanova, K.; Heid, U.; Schulte im Walde, S.; Kilgarriff, A.; Pomikalek, J. (2008). 'Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case'. In *Proceedings LREC*. Marrakech.

Kilgarriff, A.; Rychly, P.; Smrz, P.; Tugwell, D. (2004). 'The Sketch Engine'. *Proceedings Euralex.* Lorient.

Rundell, M. (ed.; 2002). *Macmillan English Dictionary for Advanced Learners*. Macmillan.

Srdanovic, I.; Erjavec, T.; Kilgarriff, A. (2008). 'A web corpus and word sketches for Japanese'. In *Japanese Journal of NLP* 15 (2).