

Augmenting online dictionary entries with corpus data for Search Engine Optimisation

Holger Hvelplund,¹ Adam Kilgarriff,² Vincent Lannoy,¹ Patrick White³

¹IDM, Paris, France

²Lexical Computing Ltd., Brighton, England

³Oxford University Press

E-mail: hvelplund@idm.fr, adam@lexmasterclass.com, lannoy@idm.fr, patrick.white@oup.com

Abstract

Search Engine Optimisation is a challenge for dictionary publishers. As soon as a dictionary appears online, one part of its success will be measured by its web traffic. Central to its volume of web traffic is where it appears on search engine results pages when a user searches for a word. There are many strategies for improving search engine rankings: the one explored here is automatically augmenting dictionary entries with corpus-derived collocates and related words, as identified by the Sketch Engine's word sketches and distributional thesaurus. We took the online version of the Oxford Advanced Learner's Dictionary and augmented a set of entries, to find whether they then saw an increase in web traffic. They did.

Keywords: corpus, collocation, SEO, Search Engine Optimisation, online dictionary

1. Introduction

A challenge faced by online dictionaries with no parallels in paper dictionaries is Search Engine Optimisation (SEO): coming top (or somewhere near top) of search engine listings when a user googles (e.g., searches in a search engine) for a word. SEO is a new art form of great importance to any enterprise using the web. For an online dictionary to reach a large audience, it must do its SEO well.

Lannoy (2010) demonstrates how a resource such as WordNet can support SEO by contributing relevant, hyperlinked text to online dictionary entries. This paper develops that work in two ways: first, by using collocations and related words discovered through a state-of-the-art corpus query system to augment entries, and secondly, through an experiment on the online version of a leading, branded dictionary, where we test the hypothesis that the additions really do bring more traffic to the website.

The dictionary in question is the Oxford Advanced Learner's Dictionary. (<http://oald8.oxfordlearnersdictionaries.com>).

2. Corpus data

The corpus methods used were 'word sketches' and a distributional thesaurus as generated (for a large number of languages, though in this case, English) within the Sketch Engine corpus query tool (Kilgarriff et al 2004, <http://www.sketchengine.co.uk>). Word sketches are one-page summaries of a word's grammatical and collocational behaviour. They have been used by lexicographers since 1998. A distributional thesaurus shows, for the target word, the words that share most collocates with it, in the sense that *tea* and *coffee* both 'share' the collocate *drink* (in the grammatical relation "object of").

For each word, the dictionary entry can be augmented with the collocates¹ from the word's word sketch, and the 'related words' from its thesaurus entry.

This information is valuable both to the dictionary user, since it tells them more about the usage of the word, and for SEO.

3. Benefits for SEO

All else being equal, pages with more text and more links are preferred by search engines, in the sense that search engine robots have more material to crawl. However, the text and links must be relevant: the search engines go to great lengths to counteract the efforts of spammers to put spam pages at the top of search results and have sophisticated algorithms for identifying junk text and junk links. As the collocates and related words are specific to the headword, and are relevant for the user, we believe they are, and will remain, acceptable to the search engines.

Each collocate and related word can be made into a link to its entry in the dictionary. This is useful to the user, as they can then click to see the entry for that word, and also beneficial for SEO. The links, to other pages on the dictionary's website, will be site-internal: site-internal links have lower weighting, within the search engines' ranking algorithms, than incoming links from external sources, but they do still carry weight.

4. Infrastructure

OALD online is managed by IDM, in DPS4. IDM created a local installation of the Sketch Engine and set up IDM

¹ In our terminology, a *collocation* comprises the node word and the *collocate*, standing in a specific grammatical relation to each other. Thus the words from the word sketch which are added to the node word's entry are its *collocates*.

DPS Processing script to use the Sketch Engine API to gather collocates and related words from the Sketch Engine. To allow flexible re-use in one or several dictionaries the script saves auto-generated content entries in a DPS project. The DPS process responsible for delivery of content for the online dictionary adapts and merges the new data into the manually produced and editorially checked OALD entries.

For fine-tuning and adapting the auto-generated content to editorial requirements, the method described here has proven to be flexible and extensible.

5. Experiment

To run the experiment, it was necessary to answer the following questions.

1. which entries would we augment?
2. which collocates and related words would we add, and how many of them?
3. how would we present the new information to the user?
4. How would we measure results of the experiment?

Throughout, it was essential to pay heed to the OUP brand: OUP is authoritative, and does not make mistakes or present nonsensical material.

5.1 Which headwords?

The headwords we used for the experiment were a random sample of 231 low-frequency words, presented below.

abalone abjure abstruse adroit aerobatics aggrandizement
agoraphobia ague amanuensis ammonite antonym apostate
apprise arachnid arrears askance askew auburn aura
autoimmune avocation azure backgammon ballpoint
barbell bargaining barista bashful beanie berserk besotted
bespoke beta betrothal bidet bigamy bitumen bling blinker
bonkers bonsai booger brainiac brainwave burlesque
calumny cardamom cashew centigrade centipede
cephalopod ceramic chamois charged chicanery chiropodist
chirpy chivalrous cliffhanger clunk colander concatenation
consonance contextualize cordially countable covetous
credulous curtsy decision-making denotation diphthong
dirge disestablish doldrums doodle dork douche downtime
dumpling dystopia edification effrontery egress emoticon
enamoured esophagus extrovert fascia feces fellatio
fricative frostbite futon gerund get-together geyser glutton
google gruel guava hale highbrow hold-up homonym
homophone hovercraft hypotenuse iconoclast igloo
incensed inchoate incorrigible infatuated ingenuous ingress

interjection intransitive introvert iterate jingoism khaki kin
lackadaisical laminate languor lassitude legit leitmotif
levity lexis liquorice located loquacious lychee lye mankind
marsupial masseuse media meerkat merry-go-round
mezzanine mnemonic mocha muffler mugging mutton
myrrh naught neigh newbie niqab obdurate obeisance
obliging obsequious occult okra omnivore ostentation
panoply parallelogram paramour paroxysm peeve peevish
perdition perfidy pestle phishing plasma platinum pre-empt
prevaricate proboscis prosody crude psychotic puerile
pugnacious quietude quintessence recon retrograde ruckus
satiare satiety scissors scotch segue sepulchre smartphone
snazzy snitch snorkel snowdrift sorority spendthrift stapler
stole sty sudoku sunglasses suntan supercilious sycophant
synecdoche taciturn tarmac tautology thither thyroid tidings
tights trendsetter triage troubleshoot truant turmeric typhoid
uncountable unflappable verbose vexation wallflower
well-being wizened wrestling wrought xylophone

5.2 Which collocates and related words?

The items to add were the highest-scoring collocates from the word sketch and the highest-scoring related words from the distributional thesaurus. The score, for both collocates and related words, was the standard measure in use in the Sketch Engine.² Ensuring the quality of these items involved a number of iterations and checks.

Initially we used the UKWaC corpus (Baroni et al 2012), comprising 1.3 billion words. However, for many of the low-frequency headwords in our sample there was not enough data: a collocate based on less than five hits is not trustworthy, and many of the words did not have collocates meeting that threshold. So we switched to enTenTen12 (Jakubicek et al 2013), with 11.2 billion words.

In the Sketch Engine, each collocation has three parts: the headword, the collocate, and the grammatical relation holding between them (eg, *object*, *modifier*). After some discussion we decided to include the grammatical relation as well as the collocate in the augmented entry. We also removed duplicates where the same collocate occurred with more than one grammatical relation. (These cases were sometimes linguistically valid, for example *brush*, at headword *hair*, can be both the verb that the headword is object of ("she brushed her hair") and a modified noun ("the hair brush"); however, the duplicates were often the outcome of part-of-speech tagging errors, and in any case, the duplication would not be helpful for the dictionary user.)

² The measure for collocates is logdice, based on the Dice coefficient. Measures are defined in the Sketch Engine documentation at <http://trac.sketchengine.co.uk/wiki>

It was important not to overload the user with too much information. We set a limit of 20 collocates in a given grammatical relation and 20 related words. We did not present related words if there was only one to present.

All words presented had to be entries in OALD themselves. All added words were then links to the word's OALD entry.

To add a collocate, the frequency of the collocation had to be at least five. This was set after some discussion of the precision-recall trade-off: a higher threshold would give fewer lexicographically dubious collocates, but would mean there were fewer entries which were augmented, so reducing the scale of the experiment.

In the experiment, collocates and related words were all checked by an OUP lexicographer. The work took 8 to 10 hours for the initial 250 entries. (For 19, there were no collocates or related words that passed all filters, leaving 231 where entries were augmented.) Of 3367 links automatically added, 98 (3%) were removed.

While this procedure would make it expensive to augment all entries, for an experiment it was of great value as it exposed a number of areas of difficulty. One of these was web spam, a significant problem in enTenten12 (Kilgarriff

and Suchomel 2013). The exercise has focussed efforts on developing very large corpora without, or with very little, web spam. Another was failure to identify, and set aside, proper names which were also lexical words.

We have a number of further ideas for improving the automatic filtering. We hope to gain access to a corpus which is smaller, but spam-free and processed with different tools. We would then only include collocates if the collocation occurred at least once in the second corpus, and related words if they were above a threshold there.

5.3 Presentation

The presentation of the augmented dictionary entry is shown below, for a concrete noun (*myrrh*), a verb (*iterate*), an adjective (*peevish*) and an abstract noun (*languor*). These entries also show entries with many, and few, added words.

The data was ready and the experimental run started on July 4th 2013. Usage statistics were gathered using Google Analytics. At time of writing, the experiment is still underway and the results presented are provisional. Also the augmented entries accounted for just 0.5% of OALD web traffic, so data sizes at this point are modest.

Fig 1: Augmented entry for *myrrh*

myrrh NOUN
 mɜː(r) BrE ; mɜːr NAme
 [UNCOUNTABLE]



a sticky substance with a sweet smell that comes from trees and is used to make perfume and incense

Beta: Collocates	▪ frankincense	▪ musk
MODIFIER	▪ aloe	PREPOSITIONAL OBJECT OF
▪ powdered	▪ patchouli	▪ tincture
MODIFIES	▪ sandalwood	PREPOSITIONAL OBJECT WITH
▪ unguent	▪ balsam	▪ perfume
AND/OR	▪ incense	

Beta: Related Entries	▪ eucalyptus	▪ geranium
▪ frankincense	▪ cardamom	▪ rosemary
▪ patchouli	▪ hyssop	▪ thyme
▪ sandalwood	▪ marjoram	▪ saffron
▪ bergamot	▪ nutmeg	▪ coriander
▪ peppermint	▪ musk	▪ anise
▪ chamomile	▪ jasmine	▪ lavender

Fig 2: Augmented entry for *iterate*

iterate VERB

'ɪtəreɪt  BrE ; 'ɪtəreɪt  NAmE



[INTRANSITIVE]

to repeat a **mathematical** or **computing** process or set of instructions again and again, each time applying it to the result of the previous stage

Beta: Collocates	▪ loop	▪ innovate
SUBJECT	AND/OR	▪ reiterate

Fig 3: Augmented entry for *peevish*

peevish ADJECTIVE



'pi:viʃ  BrE ; 'pi:viʃ  NAmE

easily annoyed by unimportant things; bad-tempered

► **SYNONYM** IRRITABLE

▪ *Sebastian was a sickly, peevish child.*

► **peevishly**

'pi:viʃli  BrE ; 'pi:viʃli  NAmE

ADVERB



▪ *'It's your own fault,' she said peevishly.*

Beta: Collocates	▪ fretful
AND/OR	▪ irritable

Beta: Related Entries	▪ morose	▪ sullen
▪ sulky	▪ grouchy	▪ petulant
▪ churlish	▪ resentful	▪ fretful
▪ uncommunicative	▪ testy	▪ despondent
▪ dissatisfied	▪ taciturn	▪ uncooperative
▪ querulous	▪ quarrelsome	▪ irascible
▪ glum	▪ touchy	▪ crabby

Fig 4: Augmented entry for *languor*

languor NOUN



'læŋgə(r)  BrE ; 'læŋgə  NAmE

[UNCOUNTABLE, SINGULAR] (LITERARY)

the pleasant state of feeling lazy and without energy

▪ *A delicious languor was stealing over him.*



► **languorous**

'læŋgərəs  BrE ; 'læŋgərəs  NAmE

ADJECTIVE

▪ *a languorous pace of life*

► **languorously**

 BrE ;  NAmE

ADVERB

Beta: Related Entries	▪ lassitude	▪ debility
------------------------------	-------------	------------

5.4 Results for users

With the experiment only running for two months at time of writing, on a small sample of entries, it is early to have gathered feedback from users and this paper emphasises SEO benefits. However we have received three unsolicited reviews, from Poland:

I have opened the dictionary today and saw the additions for the first time. I think it is a great idea and very useful! Both Collocates and Related Entries can help my students and myself in learning and teaching English. They are very intuitive and easy to use. I do hope you will develop this BETA version and we will be able to use more of it soon. Congratulations on great improvement!

From Italy:

I've just come across the beta version panel and I think it is a great idea. I do like it and I wish I could find it as much as possible

And from Spain:

I really appreciate the usefulness of the "Relative Entries" addition. I think they are a good complement that helps very much in learning vocabulary. With them it is a pleasure to relate words that in another way are difficult to find for a foreign student. I would like that, little by little, you could increase the number of entries.

5.5 Results for SEO

To establish whether the augmentations have made a difference, we have to compare web traffic for the same entries, before and after the augmentations. Moreover, since web behaviour displays annual cyclical behaviour, it is best to compare data for the same dates in different years. Web traffic is here measured using two variables: pageviews (the number of times a page was viewed), visits (where a single visit may involve a number of pageviews, as the user navigates to and fro).³ In Table 1 we present figures for the 231 test entries for the same time periods (4 July - 3 Sept) in 2012 (before augmentation) and 2013 (after).

	2012	2013	% change
Pageviews index	100	177	77%
Visits index	100	196	96%

Table 1: Test entries web traffic 2012 and 2013.

OALD web traffic has been increasing overall between 2012 and 2013, and this must be allowed for in determining if the augmentations have made a difference. The figures for OALD overall are presented in Table 2.

³ These constructs are defined in detail in Google Analytics documentation, where the relation between the indexes in the table and the actual numbers is also presented.

	2012	2013	% change
Pageviews index	100	142	42%
Visits index	100	166	66%

Table 2: All entries web traffic 2012 and 2013.

Thus pageviews increased by 77% less 42% so 35% more for the test entries, than for OALD overall; visits increased by 30% more.

To establish whether the change in pageviews was significant, we established, for each of the 231 words in the sample, whether the 2013 figure was more than 42% higher than the 2012 figure. In 141 cases it was. On the null hypothesis that the augmentation had had no impact, this number would have had a mean of $231/2=115.5$, and a standard deviation of 7.6. The observed figure of 141 is 25.5, or 3.36 standard deviations, from the mean. We apply a two-tailed test and conclude with 99.9% confidence that the null hypothesis is false. Augmentations increase web traffic.

The change can also be observed in a graph. For the ten entries having most pageviews in 2013, Fig 5 shows search traffic for the months from January to July 2013. The red line shows the point where the augmentations were made. Four trend lines are shown in the graph:

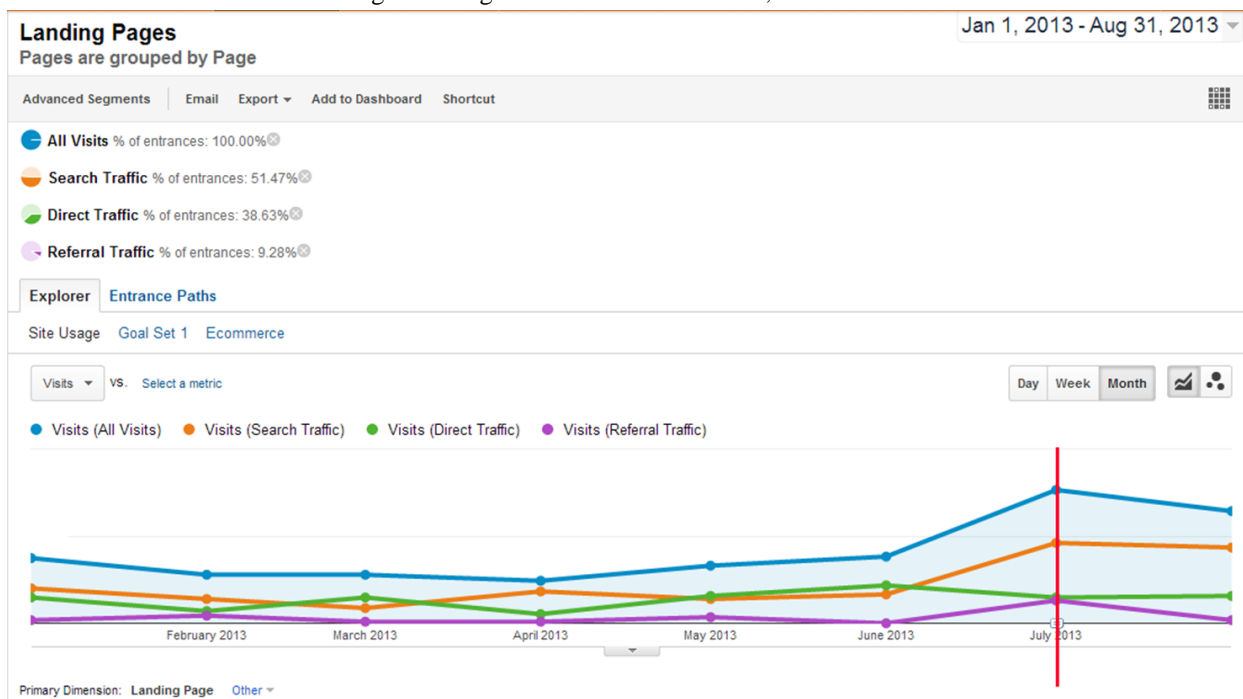
- The blue line shows all visits to the ten entries.
- The orange line shows visits from search engines to the ten entries.
- The green line shows all visits from direct traffic (that is, not from search engines) to the ten entries.
- The purple line shows referral traffic. Referral traffic is used to describe visitors who come from direct links on other websites rather than directly or from search engines.

6. Corpus size

As noted above, for the sample of words selected, there was often not enough data in 1.3b words. However these samples were of fairly infrequent words. A one billion word corpus will be adequate for, very approximately, the 20,000 commonest words of a language.

Another perspective is that, for the world's major languages, where there is ample data on the web, we are in a position to prepare these very large corpora. Lexical Computing Ltd. has recently built corpora of over 5 billion words for Arabic, English, French, Japanese, Portuguese Russian and Spanish.

Figure 5: Pageviews for ten test entries, Jan-Jul 2013



7. Conclusion

Dictionary publishers in the age of the web need their dictionary to fare well in search engine rankings. They need to engage with Search Engine Optimisation. While there are many ways to do it, one that fits well with a corpus philosophy, and which improves entries for human uses as well as for SEO, is to add collocates and related words (all hyperlinked to their own entries) to the entry. We ran an experiment to test the hypothesis that this method would increase web traffic. The experiment, for English, used the online version of the Oxford Advanced Learner's Dictionary and augmented entries automatically with collocates and related words found using the Sketch Engine in the 11.3 billion words of the enTenTen12 corpus. The experiment was run for a sample of 231 entries. Web traffic for these entries increased by 77% over the same period in the previous year, as against 42% for OALD in general.

Automatically augmenting dictionary entries with corpus-derived collocates and related words is an

effective way of boosting web traffic with useful and relevant information to human users.

References

- Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3): 209-226
- Jakubíček, M., A. Kilgarriff, V. Kovář, P. Rychlý, V. Suchomel (2013). The TenTen Corpus Family. *Proc. Int. Conf. on Corpus Linguistics*, Lancaster, UK.
- Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell D. (2004). [The Sketch Engine](#) *Proc. Euralex*. Lorient, France.
- Kilgarriff, A. & Suchomel, V. (2013). Web Spam. *Proc. 8th Web as Corpus Workshop (WAC-8)*, Lancaster, UK.
- Lannoy, V. (2010) [The IDM Free Online Platform for Dictionary Publishers](#). *Proc. Euralex*, Leeuwarden, Netherlands.