Corpus-based Extraction of Collocations for Near-Synonym Discrimination

Mariusz Kamiński
University of Applied Sciences, Nysa
e-mail: mariusz.kaminski@pwsz.nysa.pl

Abstract

This paper shows a method of extracting collocations from a corpus for the purpose of discrimination of near-synonyms. Information on synonym discrimination, which is often part of usage notes appended to entries in general dictionaries, specifies points of differences between words of similar meaning. Using computer technology, one can automate the process of collocation retrieval and reduce the amount of time and effort needed to prepare such information. The paper emphasises the importance of examining collocational behaviour when establishing meaning differences, but it also underlines the necessity of a critical analysis of existing records of word meaning, in particular dictionaries. An evidence from a parsed corpus is used in order to find collocations for a set of synonymous adjectives (*false, counterfeit, artificial, bogus, forged*) standing either in attributive or predicative positions. The corpus findings are compared against definitions in *MacMillan Dictionary* online.
Keywords: collocation; synonymy; discrimination; extraction; corpus; parser

1 Introduction: Synonym Discrimination in Dictionaries

The issue of adequate treatment of synonymy in English dictionaries was raised in the mid-nineteenth century by Richard Chenevix Trench, a member of the Philological Society. In his famous criticism *On Some Deficiencies in our English Dictionaries* (1857), Trench deplores shortcomings of English dictionaries, among which he mentions the insufficient distinction between synonymous words, and advocates a careful examination of literary passages (Trench 1860: 37). Although Trench's approach was difficult to follow in its entirety in every dictionary entry, major American dictionaries of that period, such as Worcester (1860) and Webster (1864) seem to have placed considerably more attention to synonym discrimination than any other general dictionary before them.

Before the advent of computer age, the method of excerpting synonym information was based on the lexicographer's intuition or /and manual collection of evidence from printed texts, including other dictionaries. Because the preparation of dictionaries providing synonym discrimination was considered a real challenge for publishers and lexicographers, dictionary makers often turned to earlier works as a starting point for their projects. For example, Hayakawa's *Modern Guide to Synonyms*, published in 1968 drew on an outdated Fernald's *English Synonyms and Antonyms*, published in 1914 (Landau 2001: 137). Part of the problem with compiling such works lies in the difficulty of retrieving information that would explicitly mark points of differences between words of similar meanings. It is only recently that the details of patterns of word usage and collocations can be observed clearly thanks to corpora and computer software. In particular, information on collocations, i.e. "lexical co-occurrences of words" (Sinclair, 1991: 170), has recently come to the fore as an important marker of meaning differences. Words sharing the same collocates are often regarded as potential candidates for synonyms.

2 Corpus-Based Collocation Retrieval

Modern computerised databases and tools help lexicographers work much easier than in the past. Lexicographers are now in a better position to describe word meanings, as they have access to a large amount of textual data displayed typically as KWIC concordances. However, although this

relatively new form of data presentation allows lexicographers to observe different patterns of word use on screen, the problem arises when they are confronted with the information overload, where numerous concordance lines are practically impossible to read or analyse (cf. Kilgarriff and Kosem: 2012). This is also a challenge for lexicographers attempting to distinguish between words of similar meaning, as meaning differences are hidden behind various patterns of usage. Near-synonyms are words that show a high degree of semantic overlap but have certain differentiating characteristics (Cruse 1991: 266). The differences can be observed in one or more respects, typically in variety, register, emotive content, degree of specificity, or collocation (Moon 2013: 261). However, it seems that not all of these types of information can be accessed directly from a corpus. Identifying a variety or register of the text in which a given word occurs is relatively easy for lexicographers, provided that the corpus has textual annotation specifying the source of the document, which is accessible by the computer program. As is the case in certain online lexicography software,[1] such information is retrieved automatically and displayed together with a concordance output. However, corpora do not seem to provide lexicographers with direct access to meaning components or the degree of semantic specificity. Rather information on meaning can be retrieved indirectly through collocations, thanks to computer programs employing a more sophisticated retrieval algorithm.

There are at least two ways of automatic retrieving of collocations from a corpus. One way depends on proximity of two words (Kilgarriff & Kosem, 2012), and the other on grammatical relationships. The former rests on the assumption that two words are collocates of one another if they co-occur with a frequency far greater than chance (Atkins & Rundell 2008: 369). In this approach, lists of collocates are extracted using a measure of the strength of association between two words. A commonly used measure is Mutual Information (MI), which is based on the probability of occurrence of the combination of the node and the collocate compared to the probabilities of the two words separately (Church and Hanks 1989). There are also alternative statistics, including t-score, z-score, log-likelihood ratio, the Dice coefficient, and logDice (Rychly 2008). Computer systems have been developed to retrieve candidates for collocates by calculating the above scores for words occurring in the surrounding context of a given node word. However, a drawback of this approach is that the output includes quite a number of words that are not necessarily collocates, imposing extra work on lexicographers. The lists of collocate candidates generated by the systems have to be filtered out of rare words (systems based on MI) or function words (based on log-likelihood) (Kilgarriff & Kosem 2012). The lists contain word-forms rather than lemmas, so forms such as “go”, “went”, and “gone” figure in separate positions in the lists, requiring lexicographers to draw inferences about the word usage. Another disadvantage is that the systems gather together collocates irrespective of the functional relations to the node word, which means that extra effort is needed to identify the collocates that stand in a specific relations to the node word. Furthermore, collocation extraction requires one to set up a span within which the collocates are sought, which may be problematic for lexicographers, who have no idea of how far from the node word they should look for (Kilgarriff & Tugwell 2002: 127).

An alternative solution to retrieving collocations from a corpus consists in extracting only those collocates that stand in a specific grammatical relation to the node word. In order for such a tool to work, a corpus has to be syntactically parsed so that a grammatical structure of sentences is identified. Using syntactic parsers, one can retrieve separate lists of subjects, objects or other functional categories that a given word relates to. This approach has been used by Kilgarriff et al (2004) in Sketch Engine, a corpus query tool which produces, among other types of output, summaries of a word’s patterns of usage, known as word sketches. Explicit indication of functional categories of collocates seems to be essential to discriminating between words of similar meaning. Since near-synonyms are likely to vary in collocational

[1] For example, the interface of SketchEngine and BYU interface of COCA provide concordances with the information on text source.

preferences, their collocational patterns should be taken into account when identifying meaning differences.

If a corpus is unparsed but POS tagged, an alternative solution is to identify grammatical relations on the basis of tags and regular expressions. Such an approach has been implemented as Corpus Query Language (CQL) in Sketch Engine, which also offers searching facilities based parsed corpora. However, CQL requires some familiarity with the code, as it is the user who defines grammatical relations by taking into account the possible lexical environment of the words being searched. For example, in order to define "verb-object" relation, one should take into account the fact that the noun may be preceded by a determiner, a numeral, and one or more adjectives, adverbs or nouns (for details see Kilgarriff et al 2004). The knowledge of this language allows the user to make sophisticated searches.

The following part of the paper demonstrates the utility of the parsed corpus and the computer scripts written in R in retrieving collocational information for the purpose of synonym discrimination. A corpus evidence is used in order to find collocations for the following near-synonyms standing either in attributive or predicative positions: *false, fake, counterfeit, artificial, bogus, forged*. We are especially interested in the collocations that are shared by these adjectives in order to see whether there is a common context of their usage. For illustrative purposes, this study is deliberately restricted to a comparison of adjectives in syntactic relations specified above.

3 Method

The study was conducted on the data excerpted from the BNC corpus,[2] which is a balanced sample of general English. The retrieval of the data and their subsequent processing had the following stages:

1) generating concordance lines for each occurrence of the words under analysis,
2) parsing the concordance lines with the Stanford Parser (Chen & Manning 2014),
3) retrieval of the adjectives standing in either attributive or predicative positions together with their collocating head nouns,
4) generating a frequency-ordered matrix of collocation frequencies,
5) visualising the data.

The data from BNC were processed using scripts written in R.[3] Concordances for each adjective with a span of -8 to +8 were generated and cleaned up of all tags. In order to retrieve adjectives in either attributive or predicative positions, the concordances were parsed with the Stanford Parser, a program that produces grammatical dependencies between phrases. The parser outputs more than 50 typed dependencies, but for the purpose of this analysis, we were interested only in two dependency relations that involve adjectives: one in which the adjective is found in an attributive position with respect to the head; and the other in which it is in a predicative position with respect to the subject. In the Stanford Parser output, the above relationships are indicated as "amod" and "nsubj", respectively. For example, as shown below, the sentences "*Various methods were tried, including artificial ventilation"* and "*The style is artificial"* were parsed in such a way that the relationship between *artificial* and *ventilation* is indicated by "amod", and that between *style* and *artificial* by "nsubj":

amod(methods-2, Various-1)
nsubjpass(tried-4, methods-2)
auxpass(tried-4, were-3)
root(ROOT-0, tried-4)
case(ventilation-8, including-6)

[2] BNC XML Edition.
[3] R is an open source programming language.

amod(ventilation-8, artificial-7)
nmod:including(tried-4, ventilation-8)

det(style-2, The-1)
nsubj(artificial-4, style-2)
cop(artificial-4, is-3)
root(ROOT-0, artificial-4)

Grammatical words that functioned as subjects in BNC, such as "one", "it", "they", "that", etc. were excluded from the list of collocates. The remaining 1915 collocates were extracted and ordered according to the total frequency of their occurrence in the context of the adjectives. Based on this vector, a contingency table (see Table 1) was produced, with 1915 rows occupied by collocates, and 6 columns by adjectives. An advantage of constructing such a table is that we can easily see which collocates are shared by which adjectives. This information may be useful in establishing meaning differences. A subset of the data from the table were plotted in Fig. 1 and 2. However, since collocational information is not sufficient for identification of meaning differences (at least in the case of the adjectives studied), the semantic analysis was complemented by the examination of the definitions of the adjectives provided in the *Macmillan Dictionary* online. The corpus evidence was compared against the dictionary definitions.

4 Results and Discussion

A selection of collocates for the adjectives under study is shown in Table 1. A subset of the first 100 most frequent collocates are visualised in Fig. 1 and 2, where the size of a square reflects the frequency of a collocational pair. The square size is in a direct relation to the raw frequency of the collocation: the larger the square size, the more frequent the collocation is. For reasons of space, the current discussion is limited to the most frequent collocates, but it should be remembered that the adjectives have also their collocates outside this range, and they are not shown in the figures.

Collocate ~~Node~~	false	fake	counterfeit	artificial	bogus	forged
intelligence	0	0	0	90	0	0
teeth	75	0	0	5	0	0
pretences	79	0	0	0	0	0
impression	62	0	0	0	0	0
sense	57	0	0	3	0	0
statement	58	0	0	0	0	0
starts	55	0	0	0	0	0
name	47	1	0	0	0	0
start	47	0	0	0	0	0
alarms	45	0	0	0	0	0
light	1	2	0	41	0	0
belief	41	0	0	0	0	0
consciousness	39	0	0	0	0	0
information	38	1	0	0	0	0
alarm	36	1	0	0	0	0
economy	37	0	0	0	0	0
claims	21	0	0	0	12	0
insemination	0	0	0	33	0	0
description	32	0	0	0	0	0
imprisonment	31	0	0	0	0	0

Table 1: Part of the contingency table of adjectives (columns) and their collocates (rows).

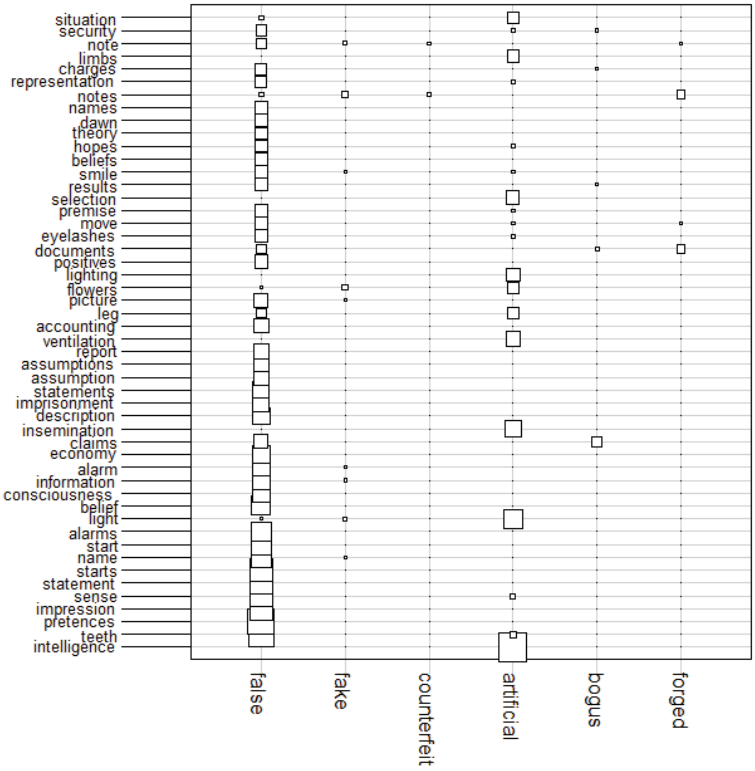

Figure 1: Visualisation of the most frequent collocates ranked 1 to 50.[4]

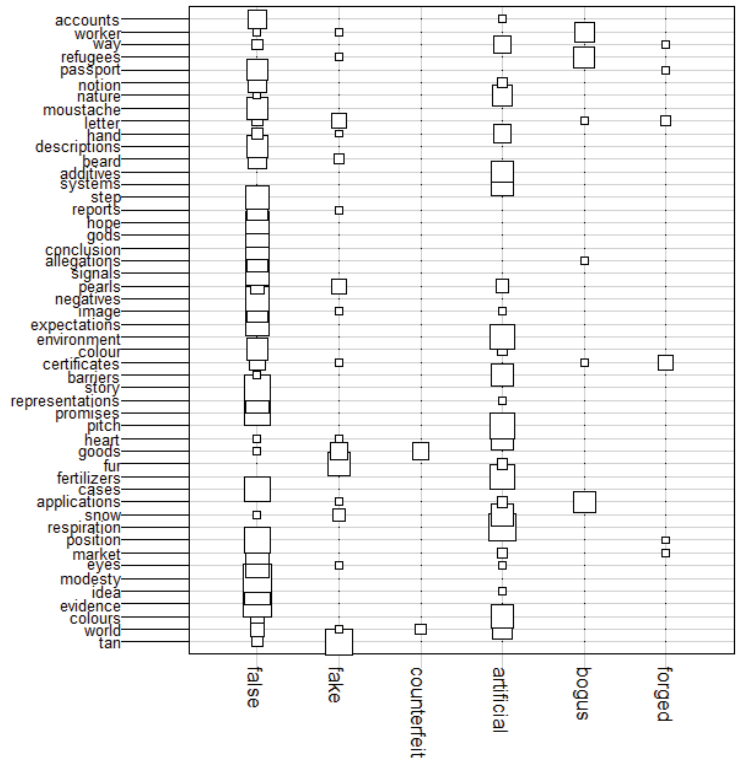

Figure 2: Visualisation of the most frequent collocates ranked 51-100.

With the collocations extracted, one can attempt a more detailed analysis of meaning differences. From Fig. 1 and 2, which show only a part of the complete collocational range, it can be seen that the adjective *false* has the widest range of application. The adjective enters into numerous

[4] The plot was created using an R function developed by Daniel Chessel.

collocations with other words (e.g. *teeth, start, economy, imprisonment*, etc.), many of which seem to be unique to this adjective, for example: *false pretences, evidence, modesty*, *expectations, promises, story, impression*, *statement, starts, economy*, etc. On the other hand, there are contexts in which *false* is hardly used, while *artificial* is preferred instead, for example in the context of *intelligence*, *light*, *lighting, insemination*, *ventilation*, *selection, limbs, pitch, respiration*, *fertilizers / fertilisers, additives*. These nouns apparently denote things that are not natural. In turn, there is some collocational overlap between *false* and *artificial*, as the adjectives share collocates such as *leg*, *pearls*, *world*, and *teeth*. With the exception of *counterfeit*, which collocates with *currency*, the remaining adjectives under study do not seem to establish distinct collocations, as they co-occur with the words that are also shared by the other adjectives; for example *flowers* collocates with *fake* as well as with *artificial* and *false*. Likewise, *claims* co-occurs with *bogus* as well as *false*. A word of caution is necessary, as the raw frequencies do not allow us to compare the strength of collocations.

Because of considerable collocational overlap of the adjectives under study, it seems that the analysis of collocations provides a partial answer to the explanation of meaning differences. Thus, in what follows, we will refer to dictionary definitions:[5]

false 1 not true ... *a false statement* 2 based on a mistake or on wrong information ... *a false impression/belief/hope* 3 made to look like something real *false eyelashes* a. not real and intended to trick people *a false passport* 4 not showing what you really feel or intend ... *a false smile*

fake made to look like something valuable or important, often in a way that is meant to trick people: *fake jewellery or fur*

counterfeit made to look exactly like something valuable or important and used illegally to trick people: used especially for describing illegally produced money: *counterfeit currency/traveller's cheques*

artificial made to have the same features or do the same job as something else that exists naturally: *artificial cream/sweeteners/flavourings*

bogus (*informal*) false and used for tricking people or pretending to be somebody you are not: *bogus auto parts* ♦ *a bogus repairman*

forged made to look exactly like something valuable or important and used illegally to trick people: *a forged signature/passport/painting*

(*MacMillan Dictionary* online)

The first definition of *false* ("not true"), which is general and simple, reflects a wide range of application of the word and its rich polysemy. The corpus findings regarding the preference of *artificial* for collocates denoting non-natural things is highlighted in the definition above, which mentions the word *naturally.* As can be seen in the definitions of *fake*, *counterfeit*, *bogus,* and *forged*, the idea of tricking or deceiving people is common to them. All these adjectives, except for *bogus*, are applied to things which are "made to look exactly like something valuable and important", though in the corpus we have counter-evidence for *bogus* modifying *certificates* and *applications*. The definition of *bogus* suggests that the adjective can modify human as well as non-human objects. Although the former type of objects cannot be seen in Fig. 1 or 2 (as *bogus applications* / *allegations* / *claims* seem to dominate the collocational range), the corpus evidence provides collocations with *refugees* and *worker* in the rank range between 100 and 150.

Summing up, collocation retrieval is only part of the lexicographers job when identifying meaning differences between near-synonyms. This is a time-consuming task, which can be made much faster thanks to a parsed corpus and corpus processing tools. However, identification of clear boundaries between them requires an in-depth analysis of not only the collocational preferences but also a wider context of word use.

[5] These definitions, except for *false*, accompany the entry for *false*, under the heading "Synonyms: false. Other ways of saying false".

Finally, it should be pointed out that the process of generating collocations presented in this paper can be automated, provided that the corpus is parsed in advance. A parsed corpus enables one to retrieve dependency relations which are necessary for this method of collocation extraction. What is more, the visualisation of the data does not have to be limited to tabular form, though this one is certainly necessary for lexicographers, as it provides the whole range of word collocations. However, the data can also be presented in any graphic way that is convenient to lexicographers, for example by means of a visual network of a subset of the data, showing links between the adjectives and collocates. Such links can be further distinguished with regard to collocational strength, for example by highlighting the lines representing the higher strength. In such a case, it is necessary to convert raw frequencies into the scores of collocational strength.

5 References

Atkins, B. T. Sue, and M. Rundell (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Chessel, D. *Plot of Contingency Tables*, retrieved from http://rpackages.ianhowson.com/rforge/ade4/man/table.cont.html [10/01/2016]

Chen, D. and C. D. Manning. (2014). *A Fast and Accurate Dependency Parser using Neural Networks*. *Proceedings of EMNLP 2014*

Church, K. and P. Hanks. (1989). Word association norms, mutual information and lexicography. In *ACL Proceedings, 27th Annual Meeting,* Vancouver, Canada. Pages 76-83.

Cruse, D. A. (1991). *Lexical semantics*. Cambridge: Cambridge University Press.

Davies, M. (2008-) *The Corpus of Contemporary American English: 520 million words, 1990-present*. (COCA). Available online at http://corpus.byu.edu/coca/

Fernald, J. Ch. (1914). *English Synonyms and Antonyms with Notes on the Correct Use of Prepositions.* New York: Funk & Wagnalls.

Hayakawa S. I. (1968). *The Funk & Wagnalls Modern Guide to Synonyms and Related Words; Lists of Antonyms, Copious Cross-References, a Complete and Legible Index.* New York: Funk & Wagnalls.

Kilgarriff, A. and D. Tugwell. (2002). *Sketching words.* In Marie-Hélène Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins.* Grenoble: EURALEX, pp. 125-137.

Kilgarriff, A., P. Rychly, P. Smrz, D. Tugwell. (2004). *The Sketch Engine*. Proc EURALEX 2004, Lorient, France; pp. 105–116.

Kilgarriff, A. and I. Kosem. (2012). Corpus tools for lexicographers. In Sylviane Granger and Magali Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press. (pp. 31-55)

Landau, S. (2001). *Dictionaries. The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Macmillan Dictionary. (2009-2015). Macmillan Publishers Limited. The entry for *false*. Retrieved from http://www.macmillandictionary.com/ [10/06/2015]

Moon, R. (2013). Braving Synonymy: from Data to Dictionary. *International Journal of Lexicography*. Vol. 26 No. 3. pp. 260-278.

R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria.

Rychly, P. (2008). A lexicographer-friendly association score. In P. Sojka and A. Horák (eds.) *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing*, *RASLAN* 2008, 6-9. Brno: Masaryk University.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Trench, R. C. (1857 [1860]). *On Some Deficiencies in our English Dictionaries*. London.

Webster, N. (1864). *An American Dictionary of the English Language*. Royal quarto edition. Revised by Chauncey A. Goodrich and Noah Porter. Springfield: G.&C. Merriam.
Worcester, J. E. (1860). *A Dictionary of the English Language*. London: Sampson Low.