Finding Multiwords of More Than Two Words[1]

Adam Kilgarriff, Pavel Rychlý, Vojtěch Kovář & Vít Baisa

Keywords: *collocations, multiword expressions, multiwords, corpus lexicography, word sketches.*

Abstract

The prospects for automatically identifying two-word multiwords in corpora have been explored in depth, and there are now well-established methods in widespread use. (We use 'multiwords' to include collocations, colligations, idioms and set phrases etc.) But many multiwords are of more than two words and research for items of three and more words has been less successful.

We present three complementary strategies, all implemented and available in the Sketch Engine. The first, 'multiword sketches', starts from the word sketch for a word and lets a user click on a collocate to see the third words that go with the node and collocate. In the word sketch for *take*, one collocate is *care*. We can click on that to find *ensure*, *avoid*: *take care to ensure*, *take care to avoid*.

The second, 'commonest match', will find these full expressions, including the *to*. We look at all the examples of a collocation (represented as a pair/triple of lemmas plus grammatical relation(s)) and find the commonest forms and order of the lemmas, plus any other words typically found in that same collocation. For *baby* and *bathwater* we find *throw the baby out with the bathwater*.

The third, 'multi level tokenization', allows intelligent handling of items like *in front of*, which are, arguably, best treated as a single token, so lets us find its collocates: *mirror*, *camera*, *crowd*.

While the methods have been tested and exemplified with English, we believe they will work well for many languages.

1. Introduction

Since (Church and Hanks, 1989) the prospects for automatically identifying two-word multiwords in corpora have been explored in depth, and there are now well-established automatic methods, implemented, evaluated and widely used in practical lexicography. But many multiwords[1] are of more than two words and research for items of three and more words has been less successful. While numerous researchers have sought to extend statistical methods for two-word multiwords to three and more (see Section 5 below) none have been widely adopted.

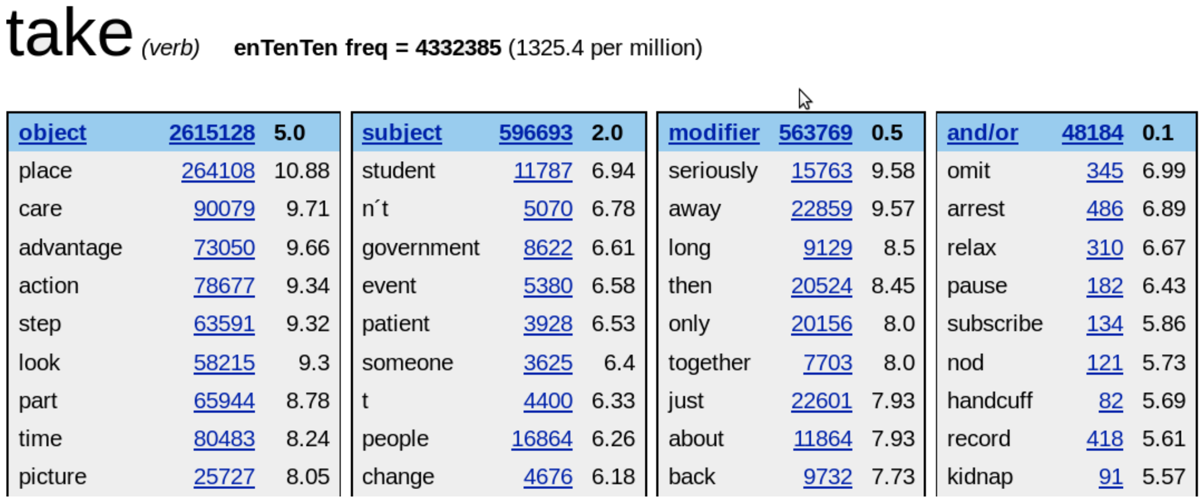
In this paper we present three complementary strategies to tackle the issue, all implemented and available. We call the methods:

- multiword sketches,
- commonest match and
- multi level tokenization.

All examples given are for English. We think methods will apply equally well to all languages though this is not yet tested.

2. Multiword Sketches

A word sketch is a one-page summary of a word's grammatical and collocational behavior (Kilgarriff et al., 2004). Typically they show the single lemmas that collocate with the nodeword. Multiword sketches allow the user to click on a word in a word sketch, to see the third collocates which go with the nodeword and collocate clicked on. For example, the user might be writing an entry for *take*; looking at its word sketch (Figure 1), they note *advantage* and see that this is very common. They should cover it in the dictionary. Next they ask "what are typical contexts and collocations for *take advantage*?" Multiword sketch functionality allows them to click on *advantage* in the word sketch for *take* to see the word sketch for *take advantage* as in Figure 2. The 'and/or' column promptly shows two flavours of *taking advantage*: one with negative semantic prosody (*abuse, exploit, manipulate, hurt*) and the other positive (*appreciate, recognize, discover*). The people most inclined to take advantage of things are students. We trust this is the positive flavour!



take (verb) **enTenTen freq = 4332385** (1325.4 per million)

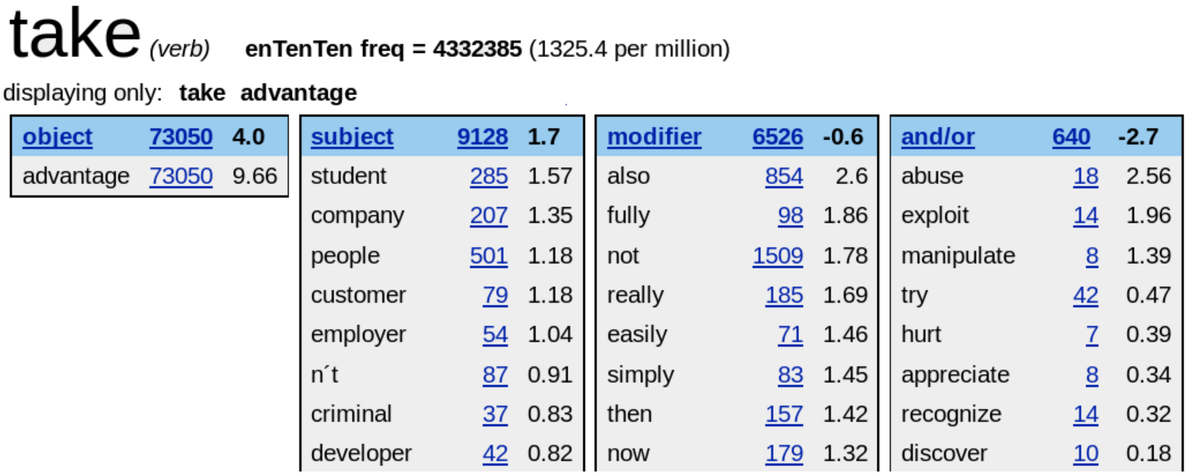
object	2615128	5.0
place	264108	10.88
care	90079	9.71
advantage	73050	9.66
action	78677	9.34
step	63591	9.32
look	58215	9.3
part	65944	8.78
time	80483	8.24
picture	25727	8.05

subject	596693	2.0
student	11787	6.94
n´t	5070	6.78
government	8622	6.61
event	5380	6.58
patient	3928	6.53
someone	3625	6.4
t	4400	6.33
people	16864	6.26
change	4676	6.18

modifier	563769	0.5
seriously	15763	9.58
away	22859	9.57
long	9129	8.5
then	20524	8.45
only	20156	8.0
together	7703	8.0
just	22601	7.93
about	11864	7.93
back	9732	7.73

and/or	48184	0.1
omit	345	6.99
arrest	486	6.89
relax	310	6.67
pause	182	6.43
subscribe	134	5.86
nod	121	5.73
handcuff	82	5.69
record	418	5.61
kidnap	91	5.57

Figure 1. Word sketch for *take*



take (verb) **enTenTen freq = 4332385** (1325.4 per million)

displaying only: **take advantage**

object	73050	4.0
advantage	73050	9.66

subject	9128	1.7
student	285	1.57
company	207	1.35
people	501	1.18
customer	79	1.18
employer	54	1.04
n´t	87	0.91
criminal	37	0.83
developer	42	0.82

modifier	6526	-0.6
also	854	2.6
fully	98	1.86
not	1509	1.78
really	185	1.69
easily	71	1.46
simply	83	1.45
then	157	1.42
now	179	1.32

and/or	640	-2.7
abuse	18	2.56
exploit	14	1.96
manipulate	8	1.39
try	42	0.47
hurt	7	0.39
appreciate	8	0.34
recognize	14	0.32
discover	10	0.18

Figure 2. Multiword sketch for *take advantage*

Regular word sketches organize collocates by the grammatical relation that the collocate stands in, in relation to the nodeword. Where three words are involved, the third word might be in the multiword sketch owing to its relation to the nodeword (*take*), or to its collocate (*advantage*), or both. We have explored several display option:

- Divide the display into two parts, for words related to *take* and for words related to *advantage*.
- Keep the usual format, but some columns will contain words in a relation to *take*, others with words in a relation to *advantage*, and others again, a mixture (this is the format illustrated).
- Dispense with grammatical relations as a way of structuring the sketch and give a list of collocates, with or without grammatical relation labels.

2.1. *More than Three Words*

The approach is iterative, so the user can click on a third-word collocate to find four-word collocates, and so on. We can click on *student* in the sketch for *take advantage* to give a sketch for *student take advantage*.

Note that it takes a large corpus and very common two-word and three-word expressions for this to give useful information; Figures 1 and 2 use the 3-billion-word enTenTen corpus. Word sketches are usually only interesting if based on several hundred data instances, and, unless we move into multi-billion word corpora, few three-word collocations have that many occurrences.

3. Commonest Match

Once we have found two lemmas which frequently go together, as in Figure 1, we can look in the data to see if there is a common string within which the two lemmas co-occur. To do this, we first see which, if any, inflected forms for the lemmas dominate, and then, whether we can 'grow' the multiword by finding words that commonly go between the words (if they do not usually occur next to each other), before the leftmost collocate, and after the rightmost collocate. 'Commonest match' output is illustrated in Figure 3 and the algorithm is given in Figure 4.

Figure 3 is an improvement on a standard word sketch as it immediately shows:

- two set phrases - *as the crow flies, off to a flying start*
- *sortie* occurs as object within the noun phrase *operational sorties* (a military expression), which is generally in the past tense
- flying *saucers* and *insects* are salient. The previous level of analysis, in which *saucer* was analysed as object of *fly*, and *insect* as subject, left far more work for the lexicographer to do, including unpacking parsing errors
- *sparks* go with the base form of the verb
- objects *flag* and *kite*, and subjects *plane, bird* and *pilot* are regular collocates, occurring in a range of expressions and with a range of forms of the verb.

The need for this function became evident in the course of evaluation. We needed the linguist or lexicographer doing the evaluation to be able to tell, at speed, whether a candidate collocation as proposed by the software was 'good', that is, whether they would include it in

Collocate	Freq	Salience	Commonest match	%
saucer	3001	9.92	flying saucers	52.3
flag	1176	8.79	—	
crow	279	8.46	as the crow flies	89.2
kite	367	8.33	—	
sortie	283	8.17	flew operational sorties	47.3
spark	256	8.02	sparks fly	40.6
aircraft	799	7.84	aircraft flying	40.8
plane	527	7.57	—	
airline	297	7.39	airlines fly	30.0
helicopter	214	7.34	helicopter flying	29.9
start	980	7.24	off to a flying start	64.8
bird	917	7.08	—	
insect	245	6.93	flying insects	82.0
pilot	350	6.68	—	

Figure 3. Commonest match output for subject and object collocates of the verb *fly*. 'Percentage' is the percentage of the hits (column 2) which the commonest match accounts for.

a published collocations dictionary. Much of the time, it was a straightforward judgment and judges agreed with one another. But one common kind of case where judges disagreed involved expressions comprising more than two content words where it was not clear, from just seeing the lemmas, what the expression was. For example, seeing *world* at *final*, one judge assessed the item as bad, whereas another first checked the concordance lines, saw *world cup finals*, and judged it good. The evidence from the evaluation gives definition to an often-noted shortcoming of word sketches that they have offered only abstract relations between lemmas. Sometimes that is all that there is to be said about an item, but sometimes it is not a transparent way to present the linguistic unit.

It also addresses a long-running dispute within corpus linguistics: lemmas, or inflected forms? Many prefer lemmas, since it allows more data to be pooled to make generalizations, and if lemmas are not used we are likely to see *invade, invades, invading* and *invaded* in the word sketch for *army*. But others (including many in the 'Birmingham school') object that this moves away from the data and misses critical facts. We are hopeful that the algorithm provides a resolution, presenting constituents of the multiword as lemmas where they occur within the multiword in a range of inflected forms, but as inflected forms, if the multiword generally uses that form.

Commonest match can be applied to multiword sketches as well as 'basic' sketches.

4. Multi-level Tokenization

The purpose of multi-level tokenization is to provide a different view on the corpus with regard to tokenization. Consider e.g. the expression *in front of.* Sometimes we want to treat it as three words.but at others, as a single unit, e. g. a preposition. Multi-level tokenization allows us both options in a single corpus.

In multi-level tokenization, level 0 is the finest-grained level. Then user-defined queries determine which words are to be joined or deleted on the higher level. Examples of the queries are shown in Figure 5.

```
Input: two lemmas forming a collocation candidate,
       and N hits for the two words

Init:  initialize the match as, for each hit, the string that starts
       with the beginning of the first of the two lemmas and ends
       with the end of the second.

For each hit, gather the contexts comprising the match,
       the preceding three words (the left context) and
       the following three words (the right context)

Count the instances of each match.
Do any of them occur more than N/4 times?
If no, return empty string.
If yes:
    Call this ’commonest match’
    n = Frequency of ’commonest match’
    Look at the first words in its right and left contexts
    Do any of them occur more than n/4 times?
    If no, return commonest match.
    If yes:
        Take the commonest and add it to the commonest-match
        Update n to the frequency of the new commonest match
        Look at the first words in the new right and left contexts
        Do any of them occur more than n/4 times?
        If yes, iterate
        If no, return commonest extended match
```

Figure 4. Description of the commonest match algorithm.

The queries are evaluated at corpus compilation time and a multi-level token index is created, allowing the lexicographer to use any of the defined levels of tokenization. Tokens from different levels can then be used in word sketches as well as in other functions, such as multi-word sketches and commonest match.

A word sketch that uses multi-level tokenization is shown in Figure 6.

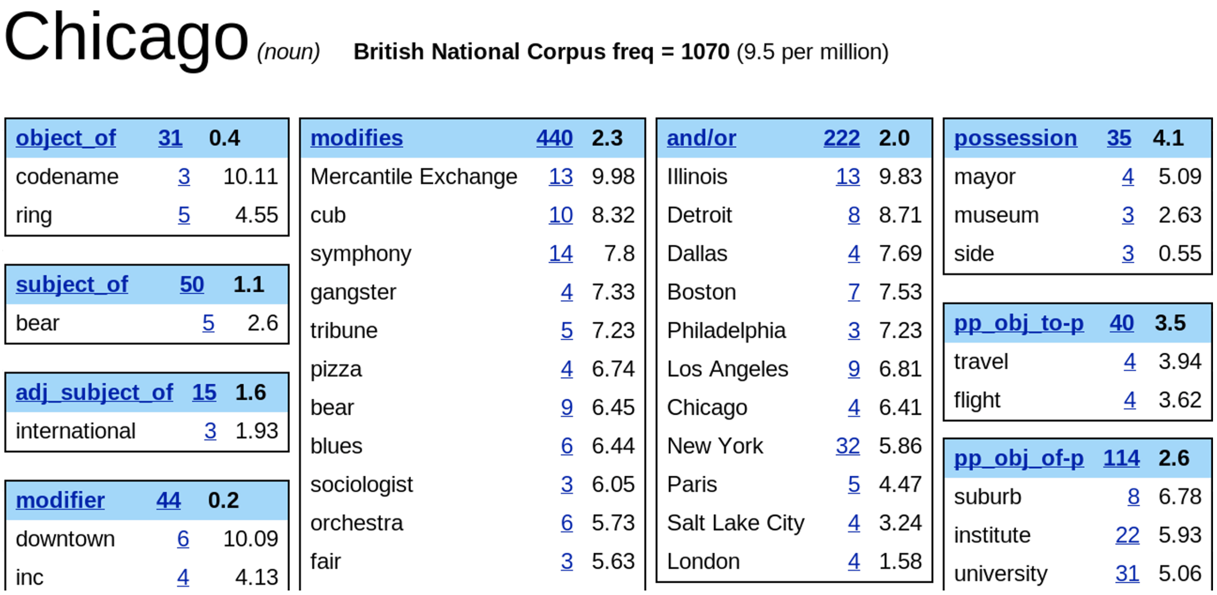
5. Related work

Following Church and Hanks’s early work, numerous other statistics for two-word collocations were proposed. A first systematic evaluation was by (Evert and Krenn, 2001). There have since been several more evaluations, including (Wermter and Hahn, 2006), which show that sorting by plain frequency performs well, and adding grammatical knowledge helps more than changing the statistic.

Work aiming to extend two-term statistics to three and more terms often does not incorporate grammar (Petrovic et al., 2010). (Dias, 2003) presents a complex system that incorporates grammatical variation and statistics but which has not, to the best of our knowledge, been tested on large corpora. ’Lexical gravity’ (Daudaravičius and Marcinkevičienė, 2004) offers a method of identifying the start and end points of multiwords that we shall be examining shortly.

```
concat(.,.,2)    [word="New"] [word="York"]
concat(.,.,-1)   [word="[A-Z].*" & tag="NP.*"]{2,}
concat(.,.,-1)   <unit/>
```

Figure 5. Example of multi-level tokenization definitions. The first query joins 'New York' into one token. The second joins any sequence of two or more capitalized proper nouns into one token. The third assumes that the markup has been added into the input file, with any sequence of tokens to be treated as a single token at the higher level enclosed in a <unit> element. `concat(.,.,2)` means the word form and lemma for the new token are the concatenations of word forms and lemmas of the old, and the tag is taken from the second token. `concat(.,.,-1)` means the tag for the new item is the tag from the last of the sequence of old items.



Chicago *(noun)* **British National Corpus freq = 1070** (9.5 per million)

object_of	31	0.4
codename	3	10.11
ring	5	4.55

subject_of	50	1.1
bear	5	2.6

adj_subject_of	15	1.6
international	3	1.93

modifier	44	0.2
downtown	6	10.09
inc	4	4.13

modifies	440	2.3
Mercantile Exchange	13	9.98
cub	10	8.32
symphony	14	7.8
gangster	4	7.33
tribune	5	7.23
pizza	4	6.74
bear	9	6.45
blues	6	6.44
sociologist	3	6.05
orchestra	6	5.73
fair	3	5.63

and/or	222	2.0
Illinois	13	9.83
Detroit	8	8.71
Dallas	4	7.69
Boston	7	7.53
Philadelphia	3	7.23
Los Angeles	9	6.81
Chicago	4	6.41
New York	32	5.86
Paris	5	4.47
Salt Lake City	4	3.24
London	4	1.58

possession	35	4.1
mayor	4	5.09
museum	3	2.63
side	3	0.55

pp_obj_to-p	40	3.5
travel	4	3.94
flight	4	3.62

pp_obj_of-p	114	2.6
suburb	8	6.78
institute	22	5.93
university	31	5.06

Figure 6. Example of the word sketch for *Chicago* with use of multi-level tokenization

6. Conclusion

We have presented three approaches for automatically detecting multiwords of three or more words. All of the proposed solutions are currently implemented and have been shown to work well with very large corpora. They can be combined with each other, forming together a powerful tool for discovering and exploring multiwords.

Notes

[1] This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013, by the Ministry of the Interior of Czech Republic within the project VF20102014003, by the Czech Science Foundation under the project P401/10/0792 and by the European union within the PRESEMT project ID 248307.

[2] We use the term broadly to include all manner of multi-word expressions: chunks, prefabs, collocations, colligations, idioms and set phrases.

[3] Tokens include words and punctuation. We say simply 'words' where this does not introduce ambiguity.

References

Church K. W., Hanks P. 1989 'Word association norms, mutual information, and lexicography.' In Hirschberg J. (ed.), *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, ACL '89, Stroudsburg, PA, USA, 76–83.

Daudaravičius V., Marcinkevičienė R. 2004. 'Gravity counts for the boundaries of collocations.' In Mahlberg M. (ed.), *International Journal of Corpus Linguistics*, 9.2: 321–348.

Dias G. 2003. 'Multiword unit hybrid extraction.' In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment – Volume 18*, MWE '03, Stroudsburg, PA, USA. Association for Computational Linguistics, 41–48.

Evert S., Krenn B. 2001 'Methods for the qualitative evaluation of lexical association measures.' In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, July 2001. Association for Computational Linguistics, 188–195.

Kilgarriff A., Rychlý P., Smrž P., Tugwell D. 2004 'The Sketch Engine.' In *Proceedings of European Association for Lexicography*, 105–116.

Petrovic S., Snajder J., Basic B. D. 2010. 'Extending lexical association measures for collocation extraction.' *Computer Speech & Language*, 24.2: 383–394.

Wermter J., Hahn U. 2006. 'You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction.' In Calzolari N., Cardie C., Isabelle P. (eds.), *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, July 2006, 785–792.