Unit 17 Collocation and pedagogical lexicography
(Case study 1)

17.1 Introduction

We introduced collocation statistics in unit 6.5 and discussed the use of corpora in lexicographic and collocation studies in units 10.2 and 13.2. These units should have provided you with a solid grounding on which to undertake the case study in this unit, which will explore how to use BNCWeb to augment the collocation information available in learner's dictionaries.

Most EFL (English as a Foreign Language) learner's dictionaries published in the UK at present claim to be based on corpus data. Yet corpus-based learner dictionaries have a quite short history: it was only in 1987 that the *Collins COBUILD English Dictionary* was published as the first 'fully corpus-based' dictionary. Yet the impact of this corpus-based dictionary was such that most other publishers in the ELT market followed Collins' lead. By 1995, the new editions of major learner's dictionaries such as the *Longman Dictionary of Contemporary English* (LDOCE, 3rd edition), the *Oxford Advanced Learner's Dictionary* (OALD, 5th edition), and a newcomer, the *Cambridge International Dictionary of English* (CIDE, 1st edition) all claimed to be based on corpus evidence in one way or another.

Yet what is a corpus-based learner dictionary? The most common use of corpora in dictionary making is for the selection of entries based on frequency information (cf. unit 10.2). This is not a new approach. Scholars such as Thorndike and Barnhart used frequency information to select the entries for elementary school dictionaries in the 1930s (Thorndike 1935). It is rather surprising, therefore, that it was not until 1995 that such word frequency marking was introduced into EFL learner's dictionaries by UK publishers. This is even more unusual when one considers that in countries like Japan this sort of frequency information had already been introduced in learner's dictionaries in the early 1960s in such a way that each entry was marked with special symbols (e.g. an asterisk, a dagger, etc.) to indicate its relative frequency. Vagaries of history aside, it is clear that corpus-based learner dictionaries now exhibit one important feature – they include quantitative data extracted from a corpus.

Another important feature of corpus-based learner dictionaries, related to frequency information again, is that such dictionaries typically select the vocabulary used from a controlled set when defining the entry for a word. Producing definitions in an L2 that language learners can understand is a problem; language learners may not have a very well developed L2 vocabulary. This makes it necessary and desirable for dictionary makers to limit the vocabulary they use when defining words in a dictionary. This notion, encapsulated in the term 'defining vocabulary', is not new – it was discussed by the vocabulary control movement in the 1930s in the United States (cf. Ogden 1930). However, it was not until the publication of LDOCE (1st edition, 1978) that the words used for defining dictionary entries were actually limited to a set of 2,000 words. Nowadays, most learner dictionary makers prepare a list of defining words, usually ranging from 2,000 to 2,500 words, based on the frequency information extracted from corpora as well as on the lexicographers' experience of defining words.

Another important use of corpus data for lexicography is in the area of example selection. This is true of learner dictionaries also. Traditionally, for unabridged dictionaries such as the *Oxford English Dictionary* and the *Webster's International Dictionary of American English*, examples were often selected from a large collection

of citations held on cards. Nowadays most dictionaries of English use corpora as the source of their examples (see unit 10.2). Hence one might be tempted to say that when learner dictionaries do so they are following a trend that is common to all dictionaries. Yet this is not quite true. In the case of learner's dictionaries, there was a tradition of using examples invented by lexicographers, rather than authentic materials, in dictionary production. This decision was influenced very strongly by the work of lexicographers working on learner dictionaries such as Harold E. Palmer and his successor A. S. Hornby, who worked together to produce the *Idiomatic and Syntactic English Dictionary* (ISED) in 1942, which was later published in the UK as the 1st edition *Oxford Advanced Learner's Dictionary* (OALD). Their reason for resisting authentic examples was simple: they believed that foreign language learners have difficulty understanding authentic materials and therefore have to be presented with simple, rewritten examples in which the use of a given word is highlighted to show its syntactic and semantic properties. It was corpus-based learner dictionary work which challenged this received wisdom: the *COBUILD* project broke with tradition and used authentic data extracted from corpora to produce illustrative examples for a learner dictionary. While there was disagreement among lexicographers concerning the value of authentic examples from corpora (cf. unit 10.8), the 2nd edition of *COBUILD* (1995) continued this policy and shifted to only using corpus examples. *COBUILD* represents an extreme case. Other dictionaries, such as LDOCE or OALD, have adopted some examples from corpora, but they do not strictly follow the policy of 'authentic examples only' and use rewritten examples from corpora whenever they view it as necessary. Nonetheless, the use of authentic examples in learner dictionaries is an area where corpus-based learner dictionaries have innovated.

Though the discussion so far outlines some ways in which corpora have changed learner dictionaries the discussion is illustrative rather than exhaustive. Yet even this short review shows that corpora have had a major impact upon the form and content of learner dictionaries. As well as providing information which can embellish existing lexicographic practice corpora may also make available new data over and above simple frequency data. A good example of this is data related to collocations, which represent, arguably, the greatest contribution that corpora have made to learner focused lexicography. For the last two decades, increasing attention has been paid to information about lexical combinability (cf. Benson 1986) or phraseology (cf. Cowie 1998). Although there have been some publications in this area, including dictionaries such as the *BBI Combinatory Dictionary of English* (1986) and *Kenkyusha's New Dictionary of English Collocations* (ed. by S. Katsumata 1939, 1958, 1995), it was only quite recently that a more serious attempt was made to incorporate collocation information from corpora into a dictionary. Hence in this case study, we will look at the derivation and use of collocation information for learner dictionaries. In doing so we will first show how to extract collocation information from corpora, in this case the BNC using BNCWeb. We will also show that different kinds of collocation statistics are used for different purposes. Following from this we will choose one entry from an EFL learner's dictionary, LDOCE, and examine how corpus data has helped to improve the description of collocation information in its 4th edition in comparison with its 1st edition. Finally, we will explore the possibility of further improving collocation information in learner's dictionaries by examining collocation data. While this study focuses on EFL dictionary making, it should be apparent that the techniques and findings of this case study are also applicable to second language lexicography for other languages.

17.2 Collocation information

Let us first explore how to extract collocation information from the BNC. We assume that you will be able to access the BNC (World Edition) via BNCWeb. In this study we will look at what collocates with *sweet*, specifically looking at what nouns co-occur with *sweet* to see whether there is a pattern in the distribution of *sweet* relative to these nouns. At this point you may want to check your own intuitions before proceeding – which noun is typically premodified by *sweet*? Jot your answers down before proceeding should you wish to do so, then consider your responses after looking at the corpus.

17.2.1 Collocation analysis using BNCWeb

Let us first examine the collocation statistics provided by BNCWeb. We will take *sweet* as an example. To find out the collocation patterns of *sweet* in the BNC using BNCWeb, follow the steps described below:

1. Activate BNCWeb. You will see the default query window of BNCWeb (Fig. 17.1).

2. Type in the search word *sweet* in the search window and click the 'Start Query' button (Fig. 17.2).

3. The results window will appear with some raw data listed (e.g. the number of matches, range, normalized frequencies) (Fig. 17.3).

4. If you click the 'KWIC View' button, you will see the KWIC concordance (Fig. 17.4). You can browse the concordance lines if you want.

5. Now select 'Collocations' from the drop-down menu next to the 'KWIC/Sentence View' button and press the 'Go!' button (Fig. 17.5).

6. A new window will appear which allows you to adjust the 'Collocation Settings'. Here you can simply press 'Submit' to continue (Fig. 17.6).

7. The collocation database will open (Fig. 17.7). This table will display various collocation statistics according to the parameters you set in the upper-half of the window.

8. Since we are interested in the collocation patterns *sweet* followed by a noun, we will define the window span as '+1 to +3' and choose 'any noun' in the 'Filter result by tag' box. Choose 'Rank by frequency' in the 'Statistics' box and press 'Go'. This will enable you to get a list of nouns collocating with *sweet*, ordered by raw frequency, shown in Fig. 17.8.

At this stage, we need to examine the list carefully to check whether the words listed are truly collocates of the node word *sweet*. Also, from a lexicographer's viewpoint, it is important to judge whether the combination of the words (e.g. *sweet smell*) should be dealt with under the main entry *sweet* or under a separate entry (e.g. *sweet tooth*), or simply ignored (e.g. *Sweet Maxwell*, in which case *Sweet* is a person's name).

9. Now we can extract more detailed collocation information. Clicking on the word in the 2nd column ('Word') will display different kinds of statistical measures (e.g. mutual information, log-likelihood, log-log, observed/expected, *z* score and MI3), showing the distribution of the collocate across the individual positions of the chosen window span. Fig. 17.9 shows the details of the collocates of *smell*.

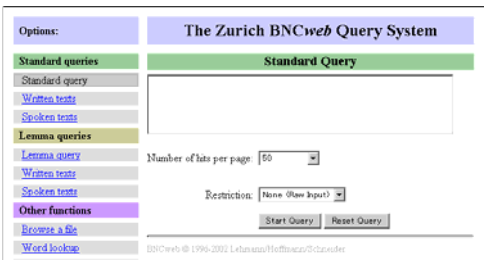

Fig. 17.1 BNCWeb interface

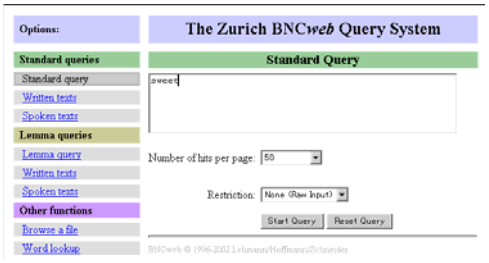

Fig. 17.2 Start query

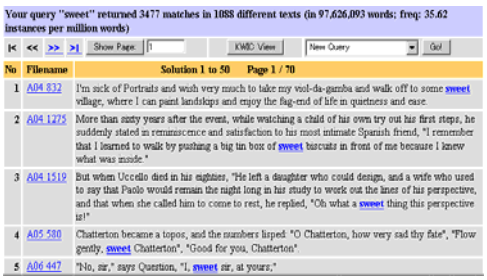

Fig. 17.3 Result window

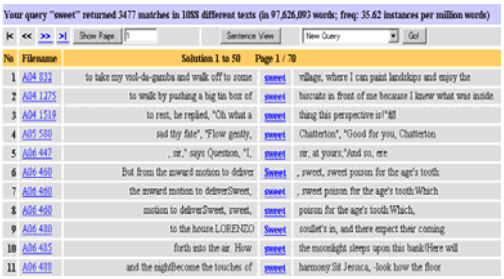

Fig. 17.4 KWIC view

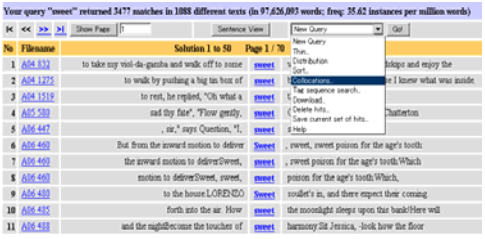

Fig. 17.5 Menu – collocation

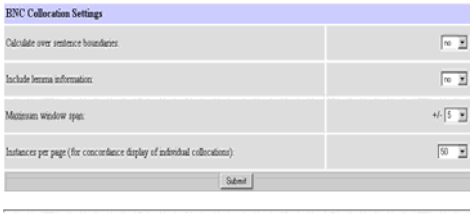

Fig. 17.6 Collocation setting

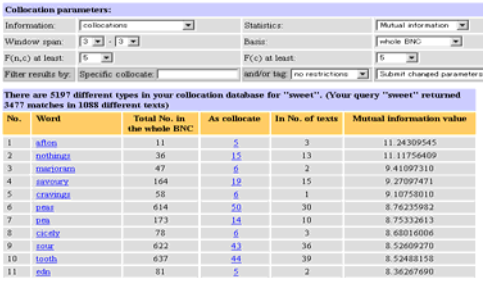

Fig. 17.7 Collocation database

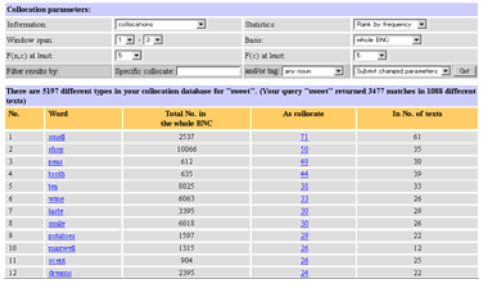

Fig. 17.8 Adjusting parameters

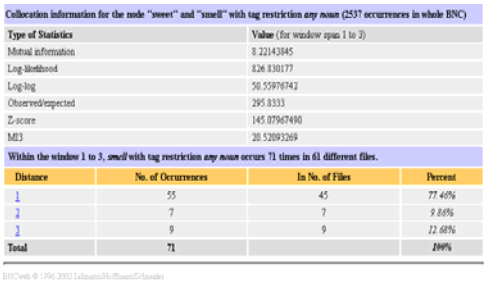

Fig. 17.9 Collocation information

17.2.2 Collocation statistics

Having obtained the various collocation statistics using BNCWeb, it is now appropriate to discuss their characteristics. These statistical measures are commonly used in corpus linguistics (see unit 6.5).

The most basic statistic used for the calculation of collocations is raw frequency. As shown in Fig. 17.8, the word *smell* ranks 1st in the column 'As collocate'. The raw frequency is 71, which means that the word *sweet* co-occurs with the word *smell* 71 times (with *sweet* as a pre-modifier) in the whole BNC. The word ranked 2nd is *shop*, which is pre-modified by *sweet* 50 times. For learner dictionaries, the list is quite useful because we can choose the collocates which tend to occur quite frequently and look familiar even to learners of English. Yet as you can see, when sorted by raw frequency of co-occurrence, frequent words crowd into the top of the collocate list. This holds out the possibility that they may not be collocates as such, rather they may simply be high-frequency words. Raw frequency is a poor guide to collocation. Look, for instance, at the third column 'Total No. in the whole BNC' for the words *smell* and *shop*. You can see immediately the difference in total frequency between the two words (2,537 times for *smell* and 10,066 times for *shop*). The raw frequency is not a reliable measure as the total number of occurrences of the word *shop* in the whole BNC is almost four times greater than that of *smell*. In the case of *smell* and *shop*, while the raw frequency also shows that *sweet smell* is a stronger collocation than *sweet shop*, we have to doubt the reliability of the raw frequency as a measure for collocations as it indicates that the combination *sweet shop* (ranks 2nd) is stronger than *sweet peas* (ranks 3rd) (see Fig. 17.8). In the case of *sweet peas*, *peas* collocates with *sweet* 49 times whilst its total frequency in the whole BNC is only 612. This indicates that *peas* shows a very strong preference to collocate with *sweet*, certainly stronger than *shop*, which occurs in the BNC 10,066 times but collocates with *sweet* only 50 times (see Fig. 17.8). In order to measure the strength of association we need to move away from the raw frequency and use other collocation statistics instead which can capture this relative strength of word combination.

One measure which takes into account the total frequencies of a node word and a collocate in relation to the size of the entire corpus is the 'observed/expected' score. This measure basically shows how far the results differ from what one would expect by chance alone. To derive a list of collocates sorted by the 'observed/expected' score using BNCWeb, select 'Observed/expected' from the pull-down menu for 'statistics' and press 'Go' in Fig. 17.8. The results should look like those given in Fig. 17.10. The list in the figure indicates that *smell* ranks 11th, with an observed/expected score of 298.4599 while *shop* ranks 42nd, with an observed/expected score of 52.7938. This rank order is hardly surprising because, as noted, the raw frequency can also give this result. However, if we consider *pea(s)* and *shop* again, we can see immediately the advantage of the observed/expected measure over the raw frequency. The observed/expected score for *pea* is 868.0599 (ranks 5th; *peas* ranks 6th, with an observed/expected score of 853.8720) whereas the score for *shop* is 52.9738 (ranks 42nd). This shows clearly that the association between *sweet* and *pea(s)* is much stronger than that between *sweet* and *shop*.

A more sophisticated statistical measure than the observed/expected score provided by BNCWeb is the z-score. The z-score is a measure which adjusts for the general frequencies of the words involved in a potential collocation and shows how much more frequent the collocation of a word with the node word is than one would expect from their general frequencies (see unit 6.5). To get a list of collocates sorted by the z-score using BNCWeb, select 'Z-score' from the pull-down menu for 'statistics' and press 'Go' in Fig. 17.8. The results are given in Fig. 17.11. The z-score measure is

widely used and built into corpus tools such as SARA and its new XML-aware variant Xaira. However, as Dunning (1993) observes, this measure assumes that data is normally distributed (see unit 6.3), an assumption which is not true in most cases of statistical text analysis unless either enormous corpora are used, or the analysis is restricted to only very common words (which are typically the ones least likely to be of interest). As a consequence, the z-score measure can substantially overestimate the significance of infrequent words (cf. Dunning 1993). As can be seen from Fig. 17.11, rare words such as *nothings* (with an overall frequency of 36 in the BNC, ranks 1st), *afton* (11, ranks 4th) and *marjoram* (47, ranks 8th) are given on the top 10 collocate list.

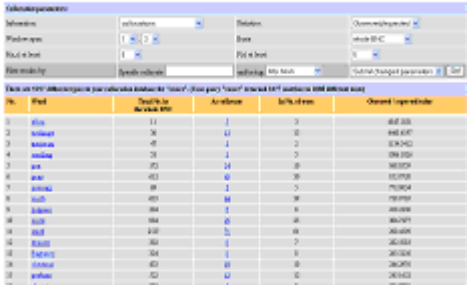

Fig. 17.10 Observed/expected values

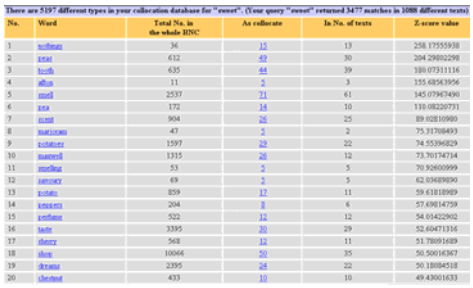

Fig. 17.11 Z scores

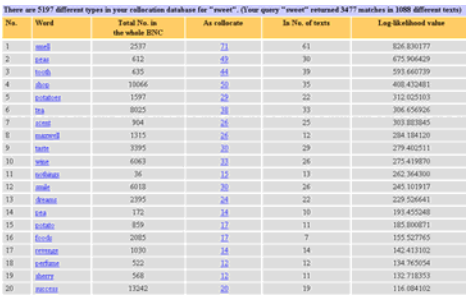

Fig. 17.12 Log-likelihood scores

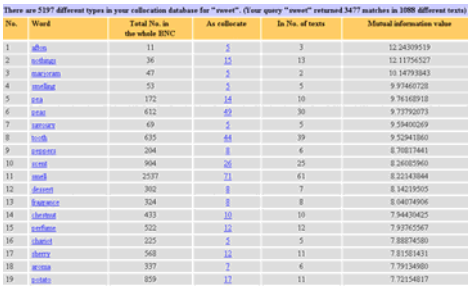

Fig. 17.13 MI scores

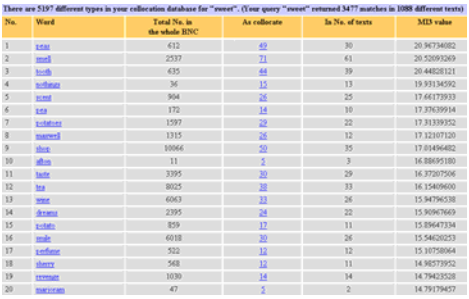

Fig. 17.14 MI3 scores

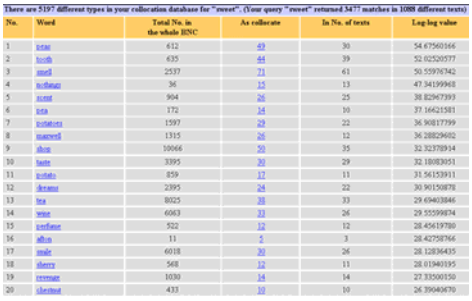

Fig. 17.15 Log-log scores

The solution Dunning proposes for this problem is the log-likelihood (LL) score (see unit 6.4). The LL measure does not assume the normal distribution of data. For text analysis and similar contexts, the use of log-likelihood scores leads to considerably improved statistical results. Using the LL test, textual analysis can be done effectively with much smaller amounts of text than is necessary for statistical measures which assume normal distributions. Furthermore, this measure allows comparisons to be made between the significance of the occurrences of both rare and common features (Dunning 1993: 67). Once again, we are fortunate in that BNCWeb provides this statistic, and hence users do not need to resort to statistics packages like SPSS to calculate the LL score. We can select ‘Log-likelihood’ from the pull-down menu for ‘statistics’ and press ‘Go’ in Fig. 17.8 to get a collocate list sorted by the log-

likelihood score. The results are given in Fig. 17.12. As can be seen, the top 10 collocates based on LL scores include both frequent and infrequent words (but none of the infrequent words in the top 10 list are as rare as *nothings*, *afton* and *marjoram*).
A quite different approach to measuring collocation is mutual information (MI). The MI measure is not as statistically rigorous as the log-likelihood test, but it is certainly widely used as an alternative to the LL and z-scores in corpus linguistics. Readers can refer back to unit 6.5 for a brief description of the MI statistic. To obtain a list of collocates for *sweet* sorted by the MI score, select 'Mutual information' from the pull-down menu for 'statistics' and press 'Go' in Fig. 17.8. The results are shown in Fig. 17.13. As shown in the figure, the top 4 collocates on the list (e.g. *Afton*, *nothings*, *marjoram* and *smelling*) are all rare words which occur less than 100 times (11, 36, 47 and 53 respectively). *Sweet Afton* is a phrase from the lyrics expressing the beauty of River Afton. *Sweet nothings* means 'romantic and loving talk'. *Sweet marjoram* is the name of a plant. For lexicographical purposes, these are interesting and should be treated in a general-purpose dictionary. However, for pedagogical purposes, these expressions are of secondary importance compared with more basic collocations. These examples show that the MI score, like the z-score, gives too much weight to rare words.
There is a way of rebalancing the MI score to address this problem by giving more weight to frequent words and less to infrequent words. The MI3 score was developed for just this purpose. MI3 achieves this effect by 'cubing' observed frequencies (cf. Oakes 1998: 171-172). The cubing of the frequencies gives a much bigger boost to high frequencies than low frequencies, thus achieving the desired effect. To obtain the collocation list sorted by the MI3 score, simply select 'MI3' from the pull-down menu for 'statistics' and press 'Go' in Fig. 17.8. The results are shown in Fig. 17.14. As can be seen, more frequent collocates such as *peas*, *smell*, *tooth* come to the top of the list when MI3 is used. This means that the cubic rebalancing pays off: these collocates are more useful for second language learners at beginning and intermediate levels.
The cubic approach to eliminating any bias in favour of low frequency co-occurrences is not the only remedy to the problem, however. The log-log formula is yet another measure which reduces this undesirable effect of the MI score. The log-log test is basically an extension of the MI formula (see Oakes 1998: 234 for a description). To obtain the collocation list sorted by the log-log score, simply select 'Log-log' from the pull-down menu for 'statistics' and press 'Go' in Fig. 17.8. The results are given in Fig. 17.15. The list looks quite similar to the one based on MI3. Both measures aim to reduce the undesirable effect of MI and produce a collocation list that shows more high-frequency words with a high rank. If you are interested in lexically unique collocations, however, MI-scores might be more useful.
A comparison of the various statistical measures provided by BNCWeb which we have reviewed so far shows that the raw frequency tends to overvalue frequent words whereas the observed/expected, MI and z-scores tend to put too much emphasis on infrequent words. In contrast, the log likelihood, log-log and MI3 tests appear to provide more realistic collocation information.
While the statistical measures reviewed in this section may appear demanding, we are fortunate in that we do not need to compute them manually. As can be seen, they can be computed automatically using corpus exploration tools or statistical packages. It is important, nevertheless, that readers understand the results of these statistical tests.

17.3 Using corpus data for improving a dictionary entry

The previous section provided us with an overview of how we could exploit statistical information when selecting useful collocations. Let us now consider how we can improve the contents of a dictionary entry with corpus data. In doing this, we will compare how the 1st (1978) and 4th (2003) editions of the *Longman Dictionary of Contemporary English* (hereafter referred to as LDOCE1 and LDOCE4 respectively) treat *sweet*.

17.3.1 Focusing on high-frequency words

Fig. 17.16 is the entry for adjectival *sweet* from LDOCE1. If you compare this with the entry from LDOCE4 (see Fig. 17.17), the first striking difference you will find is the amount of space allocated by LDOCE1 and LDOCE4 to the description of this word.

sweet[1] /swi:t/ *adj* [Wa1] **1 a** having a taste like that of sugar: *sweet fruit* **b** containing sugar: *sweet tea* **2** having a pleasant taste and smell; fresh: *sweet water* **3** pleasing to see or hear: *sweet sounds*| *sweet music* **4** gentle or attractive in manner; lovable: *a very sweet person*|*to have a sweet temper* **5 a** having a light pleasant smell, like many garden flowers **b** (of wine) having a taste caused by the presence of sugar; not DRY[1] (9) **6** pleasant: *the sweet smell of success* **7 sweet on** /'· ·/ *infml* in love with —see also SHORT[1] (13) **and sweet** —**~ly** *adv* —**~ness** *n* [U]

sweet[2] *n BrE* **1** [C] a small piece of sweet substance, mainly sugar or chocolate, eaten for pleasure —see also CANDY **2** [C;U] (a dish of) sweet food served at the end of a meal —see also PUDDING, DESSERT **3** [*my*+N] (a word used for addressing a loved one)

Fig. 17.16 The entry *sweet* in LDOCE1 (1978)

sweet[1] S2 W3 /swi:t/ *adj comparative* **sweeter**, *superlative* **sweetest**

1 **TASTE** containing or having a taste like sugar; → **sour**, **bitter**, **dry**: *This tea is too sweet.* | *sweet juicy peaches* | *sweet wine*

2 **CHARACTER** kind, gentle, and friendly: *a sweet smile* | *How **sweet of you** to remember my birthday!* → **SWEET-TEMPERED**

3 **CHILDREN/SMALL THINGS** *especially BrE* looking pretty and attractive; **cute**: *Your little boy looks very sweet in his new coat.*

4 **THOUGHTS/EMOTIONS** making you feel pleased, happy, and satisfied: *Revenge is sweet.* | ***the sweet smell of success*** | ***the sweet taste of victory*** | *Goodnight, Becky.* ***Sweet dreams.***

5 **SMELLS** having a pleasant smell; **fragrant**: *sweet-smelling flowers* | ***the sickly sweet*** (=unpleasantly sweet) *smell of rotting fruit*

6 **SOUNDS** pleasant to listen to; **harsh**: *She has a very sweet singing voice.*

7 **have a sweet tooth** to like things that taste of sugar

8 **WATER/AIR** if you describe water or air as sweet, you mean that it is fresh and clean; **stale**: *She hurried to the door and took great gulps of the sweet air.*

9 **keep sb sweet** *informal* to behave in a pleasant, friendly way towards someone, because you want them to help you later: *I'm trying to keep Mum sweet so that she'll lend me the car.*

10 **in your own sweet way/time** if you do something in your own sweet way, you do it in exactly the way that you want to or when you want to, without considering what other people say or think: *You can't just go on in your own sweet way; we have to do this together.*

11 **a sweet deal** *AmE* a business or financial deal in which you get an advantage, pay a low price etc: *I got a sweet deal on the car.*

Fig. 17.17 The entry *sweet* in LDOCE4 (2003)

In LDOCE1, only 12 lines were used to describe *sweet* whereas LDOCE4 used 48 lines. It might be argued that this is because the coverage of all words became wider in LDOCE4. That is not the case, however. The entry *sweeten*, for example, has 8 lines in LDOCE1 and 9 lines in LDOCE4. There are many other entries which are similar in length in the two editions of the dictionary. The major difference between

the two editions, in our view, lies in the way important words are treated. After corpus data was used in the 3rd edition (1985) of the dictionary, one major change in the editing policy of LDOCE was to focus more on high-frequency words. Note that the entry *sweet* (as an adjective) has the frequency labels [S2] and [W3] in LDOCE4, which indicate that the adjectival use of *sweet* is ranked among the top 2,000 in the spoken corpus data and the top 3,000 in the written corpus data. Primary emphasis was put on these high-frequency words as the lexicographers revised the entries, and as a result more space was allocated to *sweet* as an adjective in the 3rd and 4th editions. Providing quality examples is a further area where corpus data can play an important role in pedagogical lexicography.

17.3.2 Providing examples

Let us now compare the entry of *sweet* in LDOCE1 and LDOCE4 by examining the illustrative examples they provided. Illustrative examples are a crucial piece of lexical information given under a dictionary entry. They provide us with syntactic, semantic and pragmatic information about the headword. Let us first compare the number of examples provided in each dictionary, following the steps described below.

1. Count the number of examples in LDOCE1. Make a distinction between examples in complete sentences and in phrases.

2. Do the same with LDOCE4 and create a table to compare the numbers.

Table 17.1 shows the numbers of illustrative examples given in the two editions. As can be seen, LDOCE4 provides 12 full-sentence examples whereas LDOCE1 provides none of this type. Rather, LDOCE1 only gives 8 short example phrases. Illustrative examples in complete sentences are clearly more useful for language learners as they show the contexts in which headwords are used.

Table 17.1 The number of illustrative examples in LDOCE1 and LDOCE4

Example type	LDOCE1	LDOCE4
Complete sentences	0	12
Phrases	8	8

17.3.3 Providing collocation information

More important than focusing on frequent words and providing examples in context is the collocation information provided by corpus data. Let us now examine, by following the steps below, how corpus data has helped to enrich LDOCE4 with collocation data.

1. List all the examples of *sweet* from LDOCE4, one example per line.

2. Make sure you will put down the definition number for each example.

3. Look at the entry *sweet* in LDOCE1 and pick up examples that are equivalent in meaning to those in LDOCE4. Create a table to contrast how many and what types of examples are available for each definition, as shown in Table 17.2.

We can see immediately from the table that the two editions of the dictionary contrast markedly in the quality of their illustrative examples. In the table, the first column indicates the definition number in LDOCE4. The second column shows the examples from LDOCE4 while the third column gives LDOCE1 examples which have meaning/usage almost equivalent to those in LDOCE4. Clearly the example phrases

in LDOCE1 are usually shorter, showing only the 'adjective + noun' pattern divorced from their contexts. In contrast, the illustrative examples in LDOCE4 are much longer and are given as complete sentences. This way of providing illustrative examples not only makes them sound more authentic in context, it provides the learner with much richer examples also. It is the use of corpus data that has enabled this. At this point, the table is already quite revealing in that it shows that LDOCE4 gives a much more comprehensive account of the uses of *sweet*. If we go on with this experiment following the procedures described below, we will be able to see an even more marked contrast between LDOCE1 and LDOCE4.

4. Go back the collocation list derived from BNCWeb and sort the list by frequency rather than other collocation statistics this time (see Fig. 17.8).

5. Set the window span as +/-3. This adjustment is necessary because we are now interested in the collocation patterns that appear either before or after the node word *sweet* (e.g. *This tea is too sweet* or *sweet tea*). The result is given in Table. 17.3.

Table 17.2 Comparing of illustrative examples in LDOCE4 and LDOCE1

Def	Examples in LDOCE4	Examples in LDOCE1
1	This tea is too sweet.	sweet tea
1	sweet juicy peaches	
1	sweet wine	sweet fruit
2	a sweet smile	a very sweet person
2	How sweet of you to remember my birthday!	to have a sweet temper
3	Your little boy looks very sweet in his new coat.	n/a
4	Revenge is sweet.	
4	the sweet smell of success	the sweet smell of success
4	the sweet taste of victory	
4	Goodnight, Becky. Sweet dreams.	
5	sweet-smelling flowers	n/a
5	the sickly sweet smell of rotting fruit	
6	She has a very sweet singing voice.	sweet sounds; sweet music
8	She hurried to the door and took great gulps of the sweet air.	sweet water
9	I'm trying to keep Mum sweet so that she'll lend me the car.	n/a
10	You can't just go on in your own sweet way; we have to do this together.	n/a
11	I got a sweet deal on the car.	n/a
12	"How much did they pay you for that job?" "Sweet FA!"	n/a
13	a couple whispering sweet nothings to each other	n/a
14	"I got four tickets to the concert." "Sweet!"	n/a

6. Now make a new table to record the nouns co-occurring with *sweet* in each example LDOCE4 and LDOCE1 provide for the entry *sweet* in Table 17.2. For instance, the first example *This tea is too sweet* in LDOCE4 has a combination of *tea* and *sweet*. Simply write down *tea* in the column for LDOCE4. Do the same with the rest of the examples.

7. Insert the columns to show the frequency band of each noun you have recorded using the rank orders given in Table 17.3. If a noun appears on the top 10 list, mark it

with three asterisks (***). If it ranks between top 11 and 30, mark it with two asterisks (**). If it ranks between top 31 and 50, mark it with one asterisk (*).

The results should look like those given in Table 17.4. We can now compare the results between LDOCE1 and LDOCE4 and discuss how corpus data can be used to augment a dictionary with collocation information. In the table, the second column shows whether the noun collocates used in illustrative examples in LDOCE4 actually appear among the top 50 collocates (based on raw frequency) in the BNC. Of 14 nouns, 11 are found on the top 30 list (78.57%), and 6 are on the top 10 list (42.86%). This indicates clearly that lexicographers have chosen these words deliberately, and with confidence, as the examples of *sweet* in LDOCE4. In contrast, LDOCE1 is wanting in this regard with 33.33% (2 words) on the top 10 list and 62.5% (5 words) on the top 50 list (see the last column in Table 17.4). The contrast between LDOCE1 and LDOCE4 shows that collocation information would not have been readily available without corpus data.

Table 17.3 The top 50 noun collocates of *sweet* in a 3:3 window

Rank No.	Word	Rank No.	Word
1	smell	26	love
2	shop	27	foods
3	peas	28	way
4	air	29	heaven
5	tooth	30	nothings
6	taste	31	fruit
7	tea	32	jesus
8	wine	33	mouth
9	smile	34	bill
10	potatoes	35	biscuits
11	scent	36	pea
12	maxwell	37	music
13	voice	38	reason
14	home	39	tastes
15	dreams	40	baby
16	life	41	sherry
17	face	42	sauce
18	Mrs	43	water
19	things	44	lady
20	success	45	perfume
21	revenge	46	spot
22	girl	47	flavour
23	thing	48	wines
24	man	49	boy
25	potato	50	corn

As we do not have access to the corpus originally used by Longman dictionary makers (the Longman Corpus Network), we have used BNCWeb instead. As both the BNC and the Longman corpus are large, balanced and represent the same type of English in roughly the same time frame they should provide similar collocation data. The different corpus that we have used might also explain why some collocates of *sweet* in LDOCE4 (e.g. *peaches* and *deal*) are not on the top 50 list in the second column of Table 17.4.

Table 17.4 Noun collocates in LDOCE1 and LDOCE4

LDOCE4	Frequency band	LDOCE1	Frequency band
air	***	fruit	*
deal	-	music	*
dreams	**	person	-
FA	-	smell	***
nothings	**	sounds	-
peaches	-	tea	***
revenge	**	temper	-
smell	***	water	*
smile	***		
taste	***		
tea	***		
voice	**		
way	**		
wine	***		

Keys: ***=top 10; **=top 11–30; *=top 31–50

17.4 Unit summary and suggestions for further study

This case study explored corpus-based lexicography, primarily intended for language learners. This case study relates to the area of phraseology, i.e. the description of the behaviour of words in relation to the context in which they occur together. The focus of the study was on collocation analysis and the study sought to describe collocation patterns from corpus data and to relate that information to the description of a dictionary entry. The study has also demonstrated how to use BNCWeb for collocation analysis and reviewed the collocation statistics commonly used in corpus linguistics (e.g. the log-likelihood, log-log, MI and z-scores). Other useful functions of BNCWeb such as distribution and cross-tabulation will be introduced in case study 4.

If you have become familiar with collocation analysis using BNCWeb, it would be a useful exercise for you to do a small survey of the validity of collocation information in major collocation dictionaries. For example, you can look at the collocation dictionaries that follow:

- *The BBI Combinatory Dictionary of English Word Combinations*, by Benson, Benson, and Ilson. (1997, 2nd edition)
- *The LTP Dictionary of Selected Collocations*. Edited by Jimmie Hill and Michael Lewis. (1997)
- *Oxford Collocations Dictionary for Students of English*. Edited by Diana Lea. (2002)

A comparative study of these three dictionaries proves to be quite insightful. It is often the case with research in pedagogical lexicography that some empirical evidence should be provided in order to evaluate the quality of dictionaries in an objective way. However, it should be noted that we need to be careful not to overemphasize the value of a particular dictionary over another. Since dictionaries contain different types of information and are designed for specific target users, each dictionary has its own advantages and disadvantages, depending on the type of information provided and the intended use of that information. One should use different dictionaries for different purposes. While it is important to clarify some problems with a dictionary in this type of exploration, we should keep in mind that

any problems we identify may simply reflect a deliberate design decision made in the process of dictionary building which, in context, was quite justifiable.