

The Use of Corpora in Pedagogical Lexicography

Stephen Bullon

Longman Dictionaries
Pearson Education
stephen.bullon@pearson.com

Abstract

Over the last twenty or more years, pedagogical dictionaries have pioneered the use of corpora in lexicography. Corpus is now so central to the lexicographic process that it would be inconceivable for anyone in the 21st century to publish a new monolingual learner's dictionary without having used a corpus in the compilation process.

This paper will look at three stages of producing a corpus-based dictionary. Firstly, the corpus itself: what goes into it? How are the constituent parts selected? Secondly, the analytical process: how do lexicographers unearth the semantic, syntactic, and pragmatic information that is buried in these millions and millions of words? What sort of software do they need in order to be able to manipulate such large quantities of data? And thirdly: how do lexicographers convert this information into user-friendly entries?

1. Introduction

Even before the pioneering corpora compiled at Brown University and at Lancaster/Oslo/Bergen in the 1960s, linguists had been gathering natural language for the purposes of analysis and description. The Survey of English Usage, begun in 1959 at University College, London, gathered a citation collection of spoken and written language. Initially existing as thousands of annotated slips, this data was used in the compilation of the first edition of the *Longman Dictionary of Contemporary English*, (1978) in which many examples were drawn from the Survey data. By 1985, the Survey had computerized a collection of 1 million words, comprising 200 texts of 5,000 words each. The spoken component of this collection, 100 texts, is known as the London-Lund Corpus.

As a first step, these collections confirmed the potential of computerized corpora to assist in the lexicographic process; but at merely 1 million words each, there were also drawbacks. One million words was simply an insufficient amount of data to form the basis of a description of any but the most frequent words. However, as computers became ever more powerful and as optical scanners became more efficient, and as more published text became available in electronic format, dictionary houses were able to embark on compiling larger corpora.

2. The early corpora

In 1987, two learner's dictionaries based on corpora were published. The second edition of the *Longman Dictionary of Contemporary English*, and the first edition of the *Collins Cobuild English Language Dictionary*.

The Longman dictionary used a citation Corpus of 25 million words, supplemented by later additions of two million words drawn from British and American newspapers and half a million words of citations covering recent additions to the language.

The Cobuild dictionary was based on the Birmingham Collection of English Text - initially 7.3 million words, later expanded to 18 million words. This corpus was available to the lexicographic team in the form of printed and microfiche concordances.

2.1. Scaling up

By the late 1980s, it was clear that corpora needed to be larger still, and in 1991 a consortium began the creation of the British National Corpus (BNC). The consortium included Addison Wesley Longman (now Pearson Longman), Oxford University Press, Larousse Kingfisher Chambers; Oxford University Computing Services, and Lancaster University's Centre for Computer Research on the English Language.

At the same time, HarperCollins (Cobuild) began work on their own corpus, which became known as the Bank of English (BoE).

The aim for both corpora was to include a wide variety of text types in order to assure that there would be a sufficiently representative sample. To enable the inclusion of the greatest possible number of sources, the BNC set an upper limit of 45,000 words from any given text, while the BoE insisted on including entire texts.

These two corpora drew on a broad range of sources. The BNC in particular was extremely rigorous in its selection of texts, using a wide range of selection criteria. The final BNC includes over 89 million words of written English, over 6 million words of context-governed spoken English, and over 4 million words of demographically gathered spoken English. (Demographically gathered spoken data mirrors the demographics of the target group in terms of age, gender, region, educational or social background.) The

written materials came from over 3,000 different texts. It is not within the scope of this paper to list exhaustively the contents of any corpus, but see www.natcorp.ox.ac.uk/what/balance.html for details of the balance and constitution of the British National Corpus.

Since the early 1990s, corpora have continued to expand. The BofE now stands at 524 million words, while the Longman Corpus Network stands at over 400 million words.

The Longman Corpus Network has grown thanks to the addition of further British material, and also as a result of an initiative to gather large amounts of American texts, including a demographically composed corpus of 5 million words of Spoken American English. This spoken corpus was compiled with the help of the University of Santa Barbara, and involved 120 volunteers from across the US spending two weeks with a Walkman constantly on, recording every conversation throughout the two weeks.

Longman Spoken American Corpus		
120 volunteers		
5 million words		
Male		50%
Female		50%
age		
	18 – 24	20%
	25 – 34	20%
	35 – 44	20%
	45 – 59	20%
	60+	20%
White		
		75%
Black		
		13%
Hispanic		
		8%
Asian		
		4%
Degree or higher degree		
		33%
College education		
		33%
High School education		
		33%

Table 1: *Demographic breakdown of volunteers in the Longman Spoken American Corpus*

2.2. Other corpora

As well as a general corpus, learner's dictionaries can also benefit from the use of a learners' corpus. This is a collection of writing by students, classified according to mother tongue, type of task, level of student, and country of study. At Pearson Longman, we use a corpus of 12 million words of learners' English, drawn from over 30,000 texts at all levels, representing learners from 70 different languages and countries.

Various types of information can be extracted from a learners' corpus, and this information can influence the

way in which a dictionary entry is put together. An obvious example of this is when the learners' corpus shows a repeated error, made by students from several language backgrounds, and where the repeated error forms a significant proportion of the combined "correct" and incorrect formulations. The lexicographer who discovers this error will make sure that the dictionary makes prominent the "correct" formulation (be it a collocation, a syntactic pattern, a prepositional choice, or whatever) and might also write a more detailed note giving extra help to the student. (See 4.6)

3. Corpus and learner's dictionaries

It is important to remember that corpora are of more value to compilers of learner's dictionaries than of native speaker dictionaries. Learner's dictionaries aim to describe the central core of the language, and to provide help to non-native speakers who want to write in English (encoding, as opposed to decoding). So unlike native speaker dictionaries, learner's dictionaries give detailed information about phraseology, about collocation, about transitivity and valency and about the various grammatical patterns associated with each word. It is the more frequent words of the language which display the most variety in their syntactic and collocational behaviour, and the detail of this behaviour is revealed by the corpus. Because of the long tail on the word frequency graph, there is relatively little information about a large proportion of the lexicon. However, these less frequent items tend not to be included in a learner's dictionary, and so the shortcomings of corpus on that account are of little significance to pedagogical lexicographers. It is the native speaker dictionaries which attempt to account for the boundaries of the language as well as the core, and even 100 million words (BNC in 1994) or 250 million words (BofE in 1994), provides insufficient information about the infrequent frequent elements of the lexicon to provide a complete basis for a native speaker dictionary.

Table 2 shows a run of headwords selected at random from a leading English native speaker dictionary. This particular dictionary has 1872 pages of A-Z text in three columns.

Headword (and inflections)	Frequency in BNC
fuel rod (fuel rods)	38
fug (fugs)	25
fuggy (fuggier fuggiest)	3
fugacious	0
fugaciously	0
fugaciousness	0
fugacity	1
fugal	11
fugally	1
fugato	6
fugio (fugios)	0

fugitive (fugitives)	215
fugitively	0
fugitiveness	0
fugitometer (fugitometers)	0
fugleman (fuglemen)	1
fugue	95
fuguelike	0
fuguist (fuguists)	0

Table 2: BNC frequency of 19 headwords and run-ons in the middle column of page 656 of Collins English Dictionary 6th edition (2003).

There are 12 entry words, plus seven run-ons, plus nine inflected forms. Of these 28 forms, 12 have no representation on the BNC. The other 16 mustered 395 occurrences between them.

It would be inconceivable to find a column in a learner's dictionary in which nearly half the entered words did not occur in the BNC. This underlines the difference between a learner's dictionary, which describes the central core, and a native speaker dictionary, which needs to account for the boundaries of the language (the theoretically possible but highly infrequent **fugaciousness**; or the rather specialized **fugitometer** which even on Google musters only 23 for the singular and just one for the plural, (as at 21 December 2005)).

If we take the run of 14 headwords starting from **fuel rod** in the *Longman Dictionary of Contemporary English* (4th edition 2003, hereafter LDOCE4), we find that the lowest frequency word has 20 citations (and that one is the compound headword **full beam**).

So frequency plays a part in influencing the selection of what goes into a learner's dictionary and what stays out. But care needs to be taken with frequencies. Raw frequency alone is not enough, as other factors need to be taken into account, such as the learning environment. In a dictionary aimed at language learners, it seems reasonable to include the word **declension** despite the fact that it scores a very low frequency count (just eight occurrences on BNC). Other words with an equally low frequency, such as **siliciclastic**, **lipaemia**, or **ephemeris** are too specialized or technical or obscure to merit inclusion in a general-purpose learner's dictionary.

In the other direction, we sometimes exclude words which appear to be reasonably frequent, but which come from too restricted a context. For instance, the word **boilie** has 38 occurrences on the BNC and might be deemed frequent enough to merit inclusion in a learner's dictionary. But closer examination reveals that it occurs in only 4 different sources, namely

19 times in the *Anglers' Mail*
 11 times in *Coarse Fisherman*
 7 times in *Advanced Coarse Fishing*
 and once in *Practical Fishkeeping*

This is clearly a specialized and restricted use, and no learner's dictionary has an entry for this word. Similarly, the corpus has many other specialized words such as **epigastric** thanks to a few issues of the *Journal of Gastroenterology and Hepatology*.

4. Analyzing the corpus

As we saw earlier, some of the first lexicographic confrontations with large-scale corpora were in the form of printed concordances. In the case of frequent words, even from a 7 million word corpus, these could amount to scores of pages of close-typed text. And because of the fixed nature of the output, it was not possible to manipulate the data in any way once it arrived on the lexicographer's desk. So concordances were normally right-sorted before being printed. That is to say, they were ordered according to the first word alphabetically to the right of the node word. This highlighted typical collocations of attributive adjectives, typical objects of transitive verbs, typical following prepositions after nouns, adjectives or verbs, and so on. Lexicographers used a variety of techniques, including marking lines with coloured pens to keep track of different meanings, collocations, and syntactic patterns. But this paper-bound approach was short-lived, pending the more widespread availability of appropriate software, and affordable computers with bigger, faster disks.

4.1. Real-time, interactive access

The computerized interface between the lexicographer and the corpus is crucial. Several criteria need to be met, including speed of access, flexibility, tractability, and comprehensiveness.

In order to compile an entry, the lexicographer needs to be able to call up the citations for the lemma being described, selecting only those forms that are appropriate for the particular part of speech under scrutiny. To this end, the corpus needs to be tagged for part of speech and the query language needs to be able to handle requests which combine both lexical and syntactic parameters.

So, to take an example from the interface currently in use at Pearson Longman, the request

try@=v~

will display all those occurrences of the forms **try tries trying tried** that are tagged as verbs.

and the request

try@=n~

will display those occurrences of the forms **try tries** that are tagged as nouns.

It is also important to be able to look for multi-word items, for example when compiling entries for phrasal verbs, or when trying to find instances of phrases that include optional or variable words. The query language needs to be flexible enough to be able to do more than simply search for a string of characters. If you are looking for the phrase **turn the tables**, you know in

advance that the word order is variable. So you might have to do two queries:

```
turn@ the tables
tables =? =? turn@
```

The second query here allows for 0, 1, or 2 words to intervene between **tables** and **turned**. But in fact, we can consolidate this query into one:

```
<turn@ tables = ==>
```

specifies that the words **turn** (or any of its inflections) and **tables** must appear in any order within a five-word span, thus revealing *Now, to a degree, the tables had been turned* as well as *Fate had a nasty way of turning the tables against her*.

Returning to the earlier example, if you want to compile an entry for the verb **try**, you will be faced with 83,950 occurrences when looking at a combined British and American component of the Longman Corpus Network (LCN). At an average 15 words per display line, this is 1,259,250 words. This sounds like a lot, and it is (Tolstoy's *War and Peace* is a mere 564,000 words). So the role of software is vital in processing the data and presenting sensible, salient information to the lexicographer in manageable chunks.

One thing we can choose to do is select just a subset of all the occurrences. By specifying that we want to look at, say, 400 lines, we can access a manageable amount of data. And the software makes sure that the lines presented reflect the distribution of the word throughout the texts on the corpus. The actual number of lines that lexicographers look at varies from one lexicographer to another. There is no hard and fast rule, but looking at 50 lines gives a good first impression. Looking at a second 50 words will reveal things not apparent in the first batch, and the process can continue iteratively until the lexicographer is satisfied that all relevant meanings and collocations and syntactic patterns have been revealed.

But while some lexicographers choose to start with as few as 50 lines and work iteratively, many others prefer to look through a larger sample initially. Provided they come to broadly similar conclusions, both approaches are valid.

4.2. Reports

Once a query is entered, the software prepares various reports which can be viewed even before the lines appear on screen.

One of the well-known differences between humans and computers is that computers are good at recall and not so good at precision, while humans are much better at precision than they are at recall. That is to say, a computer can look through a huge amount of data and suggest a large number of *possibly* relevant results of its search, though it is likely to include some red herrings. Humans, by contrast, struggle to come up with a comprehensive range of suggestions, but are good at knowing which ones are *actually* relevant. So the ideal

dictionary is compiled using a combination of the computer's power of recall and the lexicographer's eye for precision.

The reports prepared by the software can focus the lexicographer's attention on a range of pertinent issues: collocation; variety; level; part of speech; mode, etc.

For instance, if we look for the word **vacation**, the initial screen shows graphically that the word is substantially more frequent in American English than in British English. (See appendix 1.) Here, the cluster of peaks on the right hand side of the chart, where the American data is represented, shows clearly that although the word *does* exist in British English, it is much more frequent in American English.

The reports also show whether a word is more frequent in spoken or written English. For example, the word **actually** is seen to be 7.7 times more frequent in spoken British English than in written British English, and just 4.3 times more frequent in spoken than written American. (See appendix 2.)

4.3. Collocation

Unlike a human, the computer can very rapidly check all the occurrences of a word and come up with a list of co-occurrences in frequency order. This is an interesting exercise, but all too often it produces noise, because it cannot distinguish between the genuinely interesting and the blindingly obvious. In order to get round this problem, most lexicographic houses make use of statistical measures to filter this output and present the co-occurrences in order of statistical significance.

At Pearson Longman, the statistical measures we use include T-score and mutual information (M-I).

As we will see, output from these procedures varies, depending on the nature of the word being examined and the statistical routine being employed, and the lexicographer is well advised to look at the output from all the available sources, rather than rely on a single statistical measure.

For example, when looking at the word **strategy**, T-score has fairly frequent words collocating to the left. A selection of them are: *market, economic, development, business, corporate, overall*. For the most part, they have a relatively high frequency of co-occurrence.

Item	Number of co-occurrences
market strategy	20
economic strategy	162
development strategy	162
business strategy	148
corporate strategy	96
overall strategy	72

Table 3: T-score collocates for **strategy**

M-I provides a different slant:

Item	Number of occurrences
information-retrieval strategy	10
cost-leadership strategy	13
middle-out strategy	5
island-driving strategy	9
left-to-right strategy	11

Table 4: *M-I collocates for strategy*

The actual frequency of the co-occurrences is much lower, and in many cases we have collocations which occur in only one text. These are much more specialized collocations. The LDOCE4 entry for the word **strategy** picks out several collocations in bold, and - unsurprisingly in this instance - they are all rather more prominent in T-score than they are in M-I: *economic, business, and overall*.

Taking the word **temperature**, and looking at T-score, we find (see appendix 3) that the main collocates to the left of temperature are *room, water, body, high, low, surface, air*. And to the right of the node word, we see that temperatures *drop, rise, or change*.

Temperature has over 8,000 occurrences on the LCN. The software has done some immensely valuable work in sifting through and identifying significant collocations. The amount of processing involved would take a human days rather than seconds. But as was mentioned earlier, the super-efficient recall of computers means that more collocations are identified than we really need, and this is where the skill and judgment of the lexicographer come to the fore, in assessing the results and selecting those collocates that he or she considers to be most useful for the target audience and showing them in natural examples drawn from the corpus. The result is shown below, where collocations for the first meaning of **temperature** are highlighted in the block of examples.

[+of] *The temperature of the water was just right for swimming.* | *Water boils at a temperature of 100°C.* | *The seeds should be stored at low temperatures.* | *a gradual rise in ocean temperatures* | *It took me a few days to become accustomed to the change in temperature.* | *In summer, the temperature can rise to 120 degrees Fahrenheit.* | *The temperature in New York dropped to minus 10° last night.* | *The refrigerator keeps your food at a constant temperature.* | *Red wine should be served at room temperature.* | *Exercise raises your body temperature.* | *The sun beat down and temperatures soared into the 30s.*

Figure 1: *Examples at sense 1 of temperature in LDOCE4*

4.4. Frequency of multi-word items

Most people, when asked, will say that dictionaries define "words". This is of course true, though it is not the whole story. Words have no meaning outside a context, and some words have a meaning only when in combination with specific other words. Consider the case of **kith**, which is virtually unheard of outside the expression **kith and kin**. And of course there are well-known phrases and sayings which have to be accounted for, such as **too many cooks spoil the broth**, or **like a dog with two tails**. These are easily identifiable as phrases or sayings, and are relatively infrequent compared to the majority of occurrences of **cook** or **dog**.

But there are other, far more frequent combinations which are a major feature of the language. For instance, in the entry for the noun **place**, dictionaries have to account for the multi-word item **take place**. On the LCN, we find that the noun **place** has a frequency of 67,610, and the multi-word item **take place** (including all inflected forms of **take** as a verb) has a frequency of 13,224. That represents around 20%, which is to say that one in 5 instances of the noun **place** occurs in the expression **take place**. One of our principles is to order senses according to their frequency, and in this case, LDOCE holds the expression **take place** as the third sense at the noun.

Another case where multi-word items take precedence is in the entry for the word **lookout**. The *American Heritage Dictionary of the English Language*, (4th edition 2000) offers as the first definition "The act of observing or keeping watch"; the 1913 *Webster's Revised Unabridged* offers "a careful looking or watching for any object or event"; while the *Compact Oxford English Dictionary* (2005) offers "a place from which to keep watch or view landscape". In all these instances, the definitions are for the single word **lookout**. The first two dictionaries do not refer to any significant phraseology, yet the corpus shows that the word is overwhelmingly used in the expressions *be on the lookout (for)* or *keep a lookout (for)*. For this reason, we include them first in the entry, before coming on to the more concrete but much less frequent senses of a person or a place. (In the case of the Oxford dictionary, the phrase is listed separately at the end of the entry, after four other definitions.) Once again, this underlines the differences between learner's dictionaries and native speaker dictionaries.

4.5. syntax

Grammatical patterns are also revealed by the corpus. Collocation displays show typical prepositions, preceding intensifiers, objects of transitive verbs, and so on. And simple sorting of the concordance can reveal these patterns too. By analysing and evaluating the importance of these patterns we can ensure dictionary entries account for all those patterns that are significant, while not drawing attention to patterns that are

idiosyncratic or rare. Although these grammatical patterns may seem obvious when they appear in a dictionary entry, it is in fact quite difficult to sit at a desk and retrieve them all purely using introspection. As we all use language, we feel we "know" the facts about the language and are seldom surprised when we see them written down. But without the corpus we would undoubtedly find it harder (if not impossible) to include all the significant patterns, and would possibly include unlikely patterns through lack of any more reliable means of verification than intuition.

And when it comes to some of the thorny issues of English syntax, we can at least shed some light on them if not actually resolve them.

For example, is there a difference between **different from**, **different to**, and **different than**? Teachers in the UK prefer **different from**. Does the corpus support them?

different	per million			
	British		American	
	written	spoken	written	spoken
from	38	17	29	20
to	5	13	0.15	1.2
than	0.5	1.7	4.5	14

Table 5: Relative frequencies in the LCN for **different** followed by **from**, **to**, or **than**

The corpus will not tell us what is "correct", (what is correct here?) but it does show what the regional preferences are. The following preposition *from* is the preferred option in both varieties, but as a second preference, British opts for *to* (especially in spoken English) while American opts for *than*. Obviously, there are a number of cases where *than* works as a conjunction rather than a preposition, as in *Her father seemed different than he did in New York*. But there are many other instances where *than* is unambiguously used as a preposition, as in *We're no different than any other corporation in America*. This is a well-established pattern in American English, though not so entrenched in the British variety. In a dictionary which provides coverage of both British and American English, it is useful to provide a commentary on such matters, and LDOCE4 has a note:

⚠ In spoken British English, **different from** and **different to** are both common, but teachers prefer **different from**. **Different than** is also used in American English and occasionally in British English. Do not say **different of**.

Figure 2: note in the entry for **different** in LDOCE4

4.6. Using the learners' corpus

The final comment in the note at **different** was motivated by the number of occurrences of **different of**

that we find in the learners' corpus. Looking further at the learners' corpus, we come across lines such as: *Almost each family of the world celebrates Christmas holidays*. On its own, this is unremarkable. But further investigation reveals more problems with the use of **each**. *...a card game which is played by nearly each German...* | *Each of them are kind and friendly*. | *The reason was that I thought each school was not very good*. And there are more, many, many more like this. What is interesting here is that this is not restricted to speakers of one language, or even to speakers of one language family. The four lines quoted here are from, respectively, an upper-intermediate Czech student; an advanced German student; a Proficiency level Chinese student; and an advanced Japanese student. The problem is a difficult one, and not necessarily susceptible to treatment in an orthodox dictionary entry. So the solution is to write a discursive note explaining some of the difficulties and helping clarify the differences in meaning between words such as **each** and **every**. (See figure 3.)

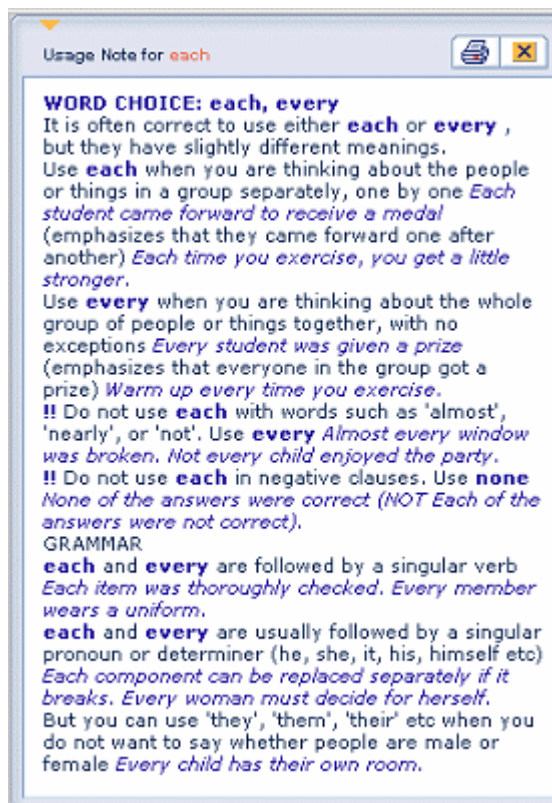


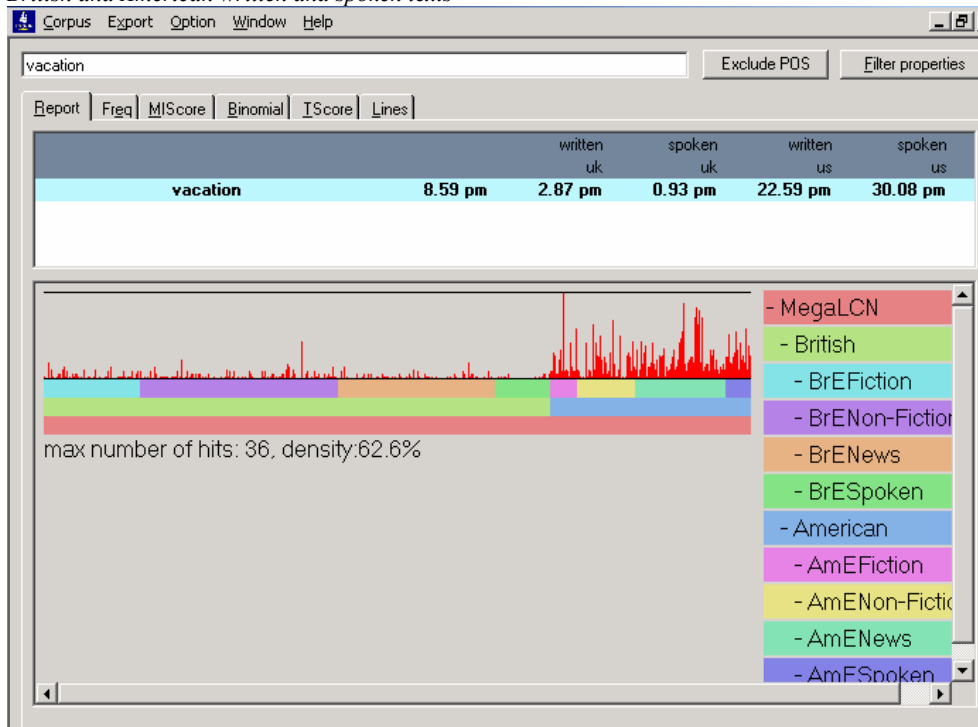
Figure 3: Word choice note for **each** and **every** from the CD of LDOCE4

5. Conclusion

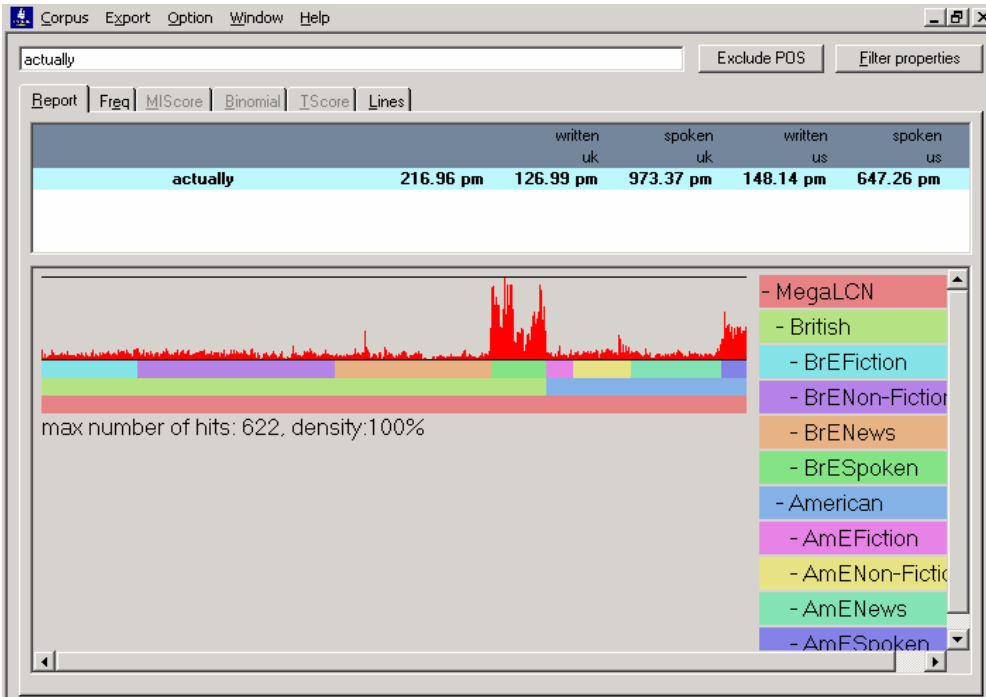
We have, over the last 30 years, made great strides in our ability to handle and analyse large quantities of linguistic data. This discussion has focused largely on the writer's experiences at Longman Dictionaries and does not attempt to speak for any other dictionary publisher. Indeed, many other publishers adopt different approaches, have differently constituted corpora, and use very different software. But we all have the same goal – to capture and present the basic facts of the language in a form that is accessible to language learners, and that reflects the language as it used today.

Corpus has now reached a point where it is not only the basis of the lexicographic analysis, but also an integral part of the dictionary itself. The Cobuild dictionary CD includes 5 million words of raw corpus material, while the CD of LDOCE4 contains a further 1 million examples drawn from the Longman Corpus Network. The corpus is no longer exclusively in the hands of specialist researchers and lexicographers, it has now reached the principal people for whom pedagogical lexicography has been developed – learners and teachers.

Appendix 1: *corpus screen showing distribution of the form **vacation** across British and American written and spoken texts*



Appendix 2: corpus screen showing written and spoken frequencies for *actually* in British and American texts



Appendix 3: T-score collocates for *temperature*

The screenshot shows a software interface for corpus analysis. At the top, there is a search bar containing the word 'temperature'. Below the search bar, there are buttons for 'Exclude POS' and 'Filter properties'. A menu bar includes 'Report', 'Freq', 'MIScore', 'Binomial', 'IScore', and 'Lines'. The main display area is a table of collocates.

-4	-3	-2	-1	0	1	2	3	4
temperatur	at	at	the	temperatur	of	the	degree	degree
be	hour	the	room	.	pressure	water	water	temperatur
water	to	a	water	and	humidity	range	temperatur	.
warm	change	when	body	be	which	low	low	.
let	increase	change	high	.	below	temperatur	(and
30	depend	low	a	:	about	freezing	C	be
hot	minute	as	low	drop	drop	room	below	of
high	light	if	surface	rise	high	C	be	a
at	cool	rais	air	change	between	high	and	at
change	temperatur	high	its	range	around	below	celsius	to
keep	maintain	with	average	at	rise	pressure	melt	in
cold	above	pressure	constant	control	above	above	high	room
gas	keep	raise	maximum	difference	a	than	freezing	water
hour	rise	extreme	melt	increase	rais	air	at	high
pressure	serve	take	same	fall	rise	reaction	surface	with
2	stand	rise	transition	rise	C	earth	range	as
condition	below	drop	ambient	(salinity	minute	1	low
degree	raise	reduce	critical	for	low	zero	deg	that
to	water	increase	global	gauge	(fall	zero	body
depend	min	average	in	dependen	reach	(level	for
maintain	rais	control	internal	gradient	so	humidity)	(
room	or	maintain	oven	variation	each	hot	K	have
light	condition	where	correct	should	but)	critical	change
volume	pressure	lower	core	in	moisture	drop	point	:
day	drop	mean	minimum	inside	allow	warm	can	or