Corpus delicti 1

Chris Payne confesses the error of his ways.

Corpus delicti is defined by the *Oxford English Dictionary* as 'the facts and circumstances constituting a crime'. The crime committed in this case was my own – of not incorporating corpus-informed language into my classes.

What is a corpus?

A corpus is a carefully laid out collection of real examples of spoken and written language stored on a computer. Because the language found in a corpus has actually been used, it consists of descriptive rather than prescriptive language. The information that corpora contain is typically presented in the form of word frequency lists and concordances. Concordances display the key word in context in example sentences.

Corpora are used to create and inform multifarious teaching resources. These include: dictionaries, reference grammars, grammar practice activities, exam practice tests and an array of materials for teaching vocabulary and lexical sets, collocations, phrasal verbs and idioms. Some coursebook writers also use corpora by consulting word frequency lists.

Why should we use corpora?

Authenticity

Corpora are a valuable resource of authentic language for all teachers. Although we tend to trust our intuitions about grammar and vocabulary, corpus evidence shows that these intuitions are sometimes flawed, and that words we think of as common are actually infrequent. Teachers can consult a corpus or a corpus-informed dictionary in order to ascertain which words are used most frequently and to keep abreast of language change.

We have been taking authentic material into our classrooms for many years, in the form of books, newspapers, magazines, leaflets, etc. Today, many of us use what is arguably the world's biggest corpus, the internet and its search engines, to find topical or engaging texts for our learners.

Frequency

Thanks to corpora, we now have more information than ever before about the differences between spoken and written English. A corpus allows us to observe important variations in the frequency of many words and structures between these two ways of communicating.

Context

As well as informing us about the frequency with which grammar and lexis occur, corpora can give us an insight into the preferred context in which words occur – some words, such as *cause*, might be used mainly in a negative context. This is sometimes referred to as *semantic prosody*.

Collocation

Corpora also show us the most common collocates and colligations of words. The box on page 17 shows the first few concordance lines for the word *crime* from a spoken corpus of British English. It is immediately clear that the collocation *crime prevention* is a frequent one.

Prioritisation

Corpus evidence is extremely useful for teaching vocabulary. Vocabulary learning creates an enormous memory load for our students, and it becomes an Augean task unless we have a sound organising principle. The *Collins Cobuild Corpus* shows that a core vocabulary of 2,500 words accounts for about 80 percent of the words in spoken and written texts. With the help of a corpus, we can identify these words and teach them as a priority to elementary levels.

Recycling

Words need to be revisited several times and in different contexts to increase the chance of them being truly acquired.

To your discussion on erm possible Nazi war	**crime**	trials coming up. Yes. My
Coming up fairly soon of course is the National	**crime**	Prevention Week and I think we ought as
Sentences. Let's have sentences which fit the	**crime.**	Because there are murders and murders aren't
But the theft element you know this rising in	**crime**	in breaking into shops Yeah. Erm
er for example has been working to prevent	**crime**	or if your group leader at school in the
Bangkok. The crackdown on switchblade	**crime**	in Glasgow. Who'll win and who'll lose
Just want stay in the game? When petty	**crime**	I just want to come back I want to come
Mm If they'd promised to reduce	**crime**	Mm and they don't deliver
Which are a large reason for the rise in	**crime**	in the first place Okay. So you have
and hospitable and generous. Is	**crime**	quite serious there and what about the drugs
private sector people er either	**crime**	prevention which there are quite a few
I mean how much do they know about the kind of	**crime**	prevention work
Only a significant role and I think sort of	**crime**	prevention as a
Of agencies which can have an influence on	**crime**	prevention as possible erm largely
Re likely to have any impact on the instance of	**crime**	the fear of crime that you can to then
Of were having to go back what is	**crime**	prevention. It is particularly

Stephen Krashen recommends extensive reading as an aid to vocabulary acquisition and retention. This is undoubtedly good advice, but the use of a concordance can be even more effective because learners are presented with a word in multiple contexts which can be read in very little time. It would take even the most omnivorous reader far longer to encounter as many examples and contexts with extensive reading.

Communication

If we aim and claim to teach communicatively, as most of us do these days, then our learners ought to be exposed to language that is used in real communication outside the classroom. We can liken learning a language to learning to drive. Sooner or later, a learner driver will need to leave the relative safety of the local industrial estate and drive in real traffic. Likewise, our learners will be in a better position to cope, when the need for communication arises outside the classroom, if we can offer them a diet of actually-used language in our lessons. We cannot always rely on a coursebook to give them the natural-sounding English they need. When the onus is on the teacher to supply more authentic language, a corpus can be a useful tool.

Simplification

It is natural to simplify language. After all, we simplify our English when we are speaking to children and non-native speakers of English outside the classroom. It should be axiomatic that some language needs to be adapted and redesigned for the specific purpose of learning English. Clearly, learners can benefit considerably from language content concocted specifically for teaching. Also, in the unpredictable environment of the classroom, we often have to think on our feet and use our own 'bespoke' examples of language.

Our learners will be in a better position to cope outside the classroom, if we can offer them a diet of actually-used language in our lessons

However, despite there being justification for a certain amount of simplified content, we should reflect on how much of it we use. It is not desirable to expose learners to an excess of contrived content. Students who encounter simplified language too often could end up learning English that is not just simplified, but simply restricted or, even worse, distorted.

If our teaching situation permits us to use some corpus-informed content, this will ensure that what our students learn is truly representative of the target language.

What can we learn from a corpus?

Corpus evidence can further our own and our students' language awareness. Of course, some data will confirm what we already know, such as the fact that question tags (*isn't it? aren't they?* etc) are almost exclusively found in spoken English. But most corpus findings will enable us to make more informed choices about what grammar and lexis to prioritise and teach, and when to teach it.

Let us look at some examples of frequency and semantic information we can obtain about a word. Space allows me to cite just a few examples, but some of the following findings may be of interest.

Frequency information

- The future continuous is 300 times more frequent than the future perfect.
- The zero conditional is the most frequently occurring pattern out of the different types of conditionals.
- Seven prepositions are in the top 20 most frequent words. Here they are in order of frequency: *to*, *of*, *in*, *for*, *on*, *with* and *at*.

Corpus delicti 1

- Of the top 50 words, 49 are grammar words, ie articles, prepositions, pronouns, conjunctions, modal and auxiliary verbs.
- Chunks containing a word may account for many of its occurrences. This is true of *hand*, where over half of all its occurrences are with chunks, *on the other hand* being by far the most common.

Semantic information

- Sixty percent of the use of *like* is prepositional and means 'to resemble something', eg *Between 1944 and 1946, Italy was like a Third World country.*
- Less than half the uses of *in* refer to place or time, but are found in adverbials and fixed phrases like *in fact*.
- The word *see* is much more common in spoken corpora with the meaning 'understand' (eg *I see* or *I see what you mean*) than it is with the meaning 'perceive with the eyes'.
- *Must* is first taught for referring to obligation. Corpora confirm for us that its function for expressing speculation or deduction, as in *You must be hungry*, is also a very frequent grammar pattern. The perfect form *must have been* is extremely common in spoken English. Perhaps its place in syllabuses should be reassessed.
- In a mixed corpus of American English, *would* is the 46th most frequent word. Dave Willis claims that, in spite of conventional EFL wisdom, *would* denoting 'used to' is remarkably common.

How should we use a corpus?

There are different kinds of corpora, both large and small, available for us to consult. Among them are general corpora of spoken and written American, British or other varieties of English. There are also specialised corpora, including academic and business English, and teacher, learner and non-native-speaker corpora. As teachers, we should remember that native-speaker corpora tell us a lot about the way native speakers use language, but nothing about the way languages are *learnt*. Thus, it's a good idea to look at a learner corpus, which lets us see the problems *learners* might experience. Then we can compare learner and native-speaker corpora to see why errors occur.

We need to make judicious use of corpora, which entails critically interpreting corpus findings and selecting language wisely for teaching. This is important because we want to avoid having to modify or alter corpus information, for this would defeat the object of choosing it as authentic material in the first place.

> *Native-speaker corpora tell us a lot about the way native speakers use language, but nothing about the way languages are learnt*

Caution is also required when consulting frequency information. The fact that a particular example of language use is attested as frequent does not automatically mean it is suitable for teaching purposes. Some language contained in corpora is inappropriate for the classroom, irrespective of whether the classroom is L1 or L2. Other language is best taught for reception only, a point raised by Peter Wells in Issue 65 of ET*p*, when referring to slang.

Nor should we use frequency evidence alone without considering other criteria, such as the learnability of the language and whether it is relevant to our learners' needs and interests. The words *Tuesday* and *Wednesday* are relatively low in frequency compared with the other days of the week, but they form part of the same lexical set and we would not contemplate leaving them off a beginners' syllabus. I pointed out earlier that *see* meaning 'understand' and *would* meaning 'used to' are common occurrences. Yet this does not mean that these senses of the words should be taught before or to the exclusion of their other meanings.

★ ★ ★

As a linguistic resource, corpora are especially useful for promoting noticing, and there is a strong case for using them for language learning.

My crime was that I had failed to make use of the invaluable work carried out by corpus linguists like John Sinclair, Michael McCarthy and others. We are doing our learners a disservice if we do not exploit the significance of the patterns of grammar and lexis revealed by modern corpora. Teaching of the four skills can also benefit by looking at how communication works in speech and writing. A corpus needn't be considered as an esoteric research tool or as the preserve of applied linguists. By using one we can add another string to our pedagogic bow.

Having first confessed, in the next issue of ET*p* I would like to address the use of corpora and will suggest some practical activities. ETp

There are many corpus-based resources available online, and some of them are free. You can download examples of non-native-speaker talk for free from the *Michigan Corpus of Academic Spoken English*. Other useful corpora are *The British National Corpus* and the *Collins Cobuild Corpus*.

O'Keeffe, A, McCarthy, M and Carter, R *From Corpus to Classroom* CUP 2007

Tomlinson, B *Materials Development in Language Teaching* CUP 1998

Tribble, C and Jones, G *Concordances in the Classroom* Athelstan Publications 1997

Willis, D *The Lexical Syllabus* Collins 1990

Chris Payne is the owner of Paddington School of English and has been teaching in Spain since 1993. He has published several articles on ELT and is particularly interested in a greater focus on lexis in language learning.

paddington@terra.es