

**Turning numbers into thoughts: making sense of language corpora
(1999) *The Language Teacher*, 25-27.**

Michael McCarthy
Professor of Applied Linguistics
University of Nottingham, UK

Abstract

Corpora are growing bigger, and there is little doubt that they will continue to be important in language teaching. But we should consider the ideological issues they engender, especially spoken corpora. Among key issues are the status of external evidence, the authority of native speakers, and the approach expert users might take in interpreting corpus evidence.

Technology and observing language

In 1966, when I first became an English language teacher, computers were in their infancy and were only accessible to a privileged few scientists. Linguists certainly had not yet been able to utilise their potential for language study, even though in that very year, both Halliday (1966) and Sinclair (1966) published forward-looking papers which presaged much of the later research into vocabulary that became possible once linguists got their hands on decent, easy-to-use computers. At that time, observations of language were either made from an anthropological viewpoint (that is to say 'out in the field', with linguists living amongst and observing linguistic communities, recording their languages with extensive field-notes and rather clumsy portable tape-recorders), or simply from the linguist's intuition. Nowadays, computing power is cheap and easy to use, digital tape-recorders and text-scanners make data collection very straightforward, and publishers are keen to invest in corpus projects which they believe will yield new and more powerful information about language usage which can be used in language study materials. The landscape, therefore, has shifted irrevocably. Few would anymore doubt the value of large corpora as a basis for the construction of dictionaries, and it is not at all science fantasy to envisage corpora of billions of words in the very near future, which might inform coursebooks, CD- and DVD-ROM packages, and be available at the drop of a hat on the Internet. Yet we should perhaps stop and ponder awhile on the implications of our present-day abilities, both at the theoretical and practical levels, for our now awesome power at once opens new positive vistas and throws up some potentially thorny issues.

The evidence: internal or external?

We are increasingly told by corpus linguists (myself among them, see McCarthy 1998), that our intuitions are not always as reliable as we might like to think they are when it comes to deciding what we really do say and write rather than what we think we say and write. Or rather, in my own case, I would take the line that we are perhaps better at intuiting written forms than spoken ones. This is because we can usually reflect when we write, and we can certainly stop reading and reflect on any piece of writing with relative ease. Speech is different: the vast majority of spoken words we produce drift off into the air never to be heard again, and speech is most typically face-to-face, or at least produced in real-time, with little opportunity to reflect. It is

my contention, therefore, that when informants are asked to judge the grammaticality of sentences, they ‘translate’ them into written texts and judge them against written norms. Many sentences deemed ungrammatical in writing pass completely unnoticed and unproblematically in even the most educated speech. But the main point about intuition is that it is internal; the evidence comes from within the mind of the linguist or teacher. There is no need to have recourse to the ‘world out there’, especially in the case of the native speaker, for he/she is endowed with ‘competence’, that invisible underbelly that ‘knows’ its native language, even if the visible manifestation, ‘performance’ is often wanting. What flows from a trust in intuition is not only faith in the power of internal evidence, but, almost necessarily, that native speakers know better than others, that the educated native speaker represents the highest authority in any dispute over usage. Where native speakers are not to hand, grammars and dictionaries based on the written evidence of the canon of great writers in any culture can be called to give evidence.

Such a view of the world held water until linguists began to get access to large spoken corpora. I am one such privileged linguist. With my colleague, Ronald Carter at the University of Nottingham, UK, I co-direct the CANCODE spoken corpus project, sponsored by Cambridge University Press. The corpus consists of five million words of conversations recorded in everyday situations in the islands of Britain and Ireland. Cambridge University Press, with whom the copyright resides, have also generously given us access to their massive, 100-million word Cambridge International (written and spoken) corpus (CIC), so that we can make comparisons between speech and writing.

Among the many striking things about a spoken corpus is the extremely wide range of tolerance and relaxed use of forms that would often be considered problematic in writing (see Carter and McCarthy 1995; 1997). Writers orient towards standard norms; speakers accommodate to one another and to the moment. The spoken corpus is a vast treasure, but it is one which throws up real challenges. What is more, the evidence is all external: it is simply ‘text’, but that does not mean it is ‘objective’ or free from cultural and ideological problems.

Unlike the written evidence, the spoken corpus is not based on a canon of the writings of the great and the good in any culture; it is simply ordinary people talking in ordinary ways in ordinary situations, and significant ideological issues are raised by using a spoken corpus of native speakers, such as CANCODE. If the range of speakers is demographically representative, then predictably, widely different levels of competence (whether linguistic or communicative) will be apparent among the speakers in the corpus, just as is the case with writing. The spoken corpus will include many speakers who strike us as able, clear, communicative and expressive; it will also include those who stumble, who make a bad job of getting their meanings over, who display eccentric usage, and so on. Many of the native speakers in a corpus will be less proficient than many non-native speakers known to us. The automatic claim of the native speaker to represent the ideal target for the learner is therefore held up to question. Seen from a communicative point of view (and in many cases also from the point of view of grammatical accuracy vis-a-vis standard grammars), in the real world there are expert and inexpert native speakers, and expert and inexpert non-native speakers.

The ideological shift required is one that takes us from the notion of the native speaker to the notion of the *expert and informed user*, in the knowledge that both may be rather difficult to define within our present socio-cultural frameworks. Identifying criteria for expert use of or expert knowledge in a language like English in different

cultural contexts is an urgent one, and one which will be necessary if we are to develop a notion of standard that is not tied to old-world, written norms and simply perceived as another manifestation of linguistic imperialism. The alternative is probably unattainable: to assemble a database that is truly representative of all the thousands of types of spoken English that occur in thousands of contexts around the world, 24 hours of every day.

Humanising the numbers game

So far, I have asserted that corpora, especially spoken ones, are powerful external evidence of how speech communities and cultures communicate, and we need to shift our ideological perspectives to value them fully (a shift from reliance on intuition and from the elevation of the native speaker as the source of authority). But how should expert users of a language such as English, amongst whom I include the (native- and non-native-speaker) readers of this journal, in a practical sense, approach corpus evidence when it is available?

As with so much research, a balance of the quantitative and the qualitative is obviously desirable. ‘Quantitative’ here refers to the allure of numbers and statistics, which computer software can generate with great ease (see Figure 1). ‘Qualitative’ in this case means humanistic interpretation, plausible explanations of the data, seeing through the numbers to the culture that produced them, and modelling the data for language teaching in a way that is *relevant* for our purposes. Another balance necessary to strike is that between language teaching that is *corpus-driven* and that which is *corpus-informed*. A corpus-driven approach is absolutely faithful to the evidence of the corpus; a corpus-informed approach takes insight from the corpus, but filters that insight through common-sense language teaching practices. For instance, we might take a real vocabulary collocation from the corpus, such as ‘have lunch/dinner’, note that it is much more frequent than ‘have a car’ or ‘have two sisters’, but nonetheless prefer a usable, short, invented context for it in our vocabulary teaching at elementary level, rather than simply ‘throwing in’ a real, and perhaps difficult-to-contextualise utterance, unedited straight from the corpus (see McCarthy and O’Dell 1999).

Let us finally, briefly, consider how an expert language user might make sense of a small bunch of numbers. Figure 1 shows the ‘top twenty’ word-forms from one-million-word spoken and written samples of CANCODE/CIC corpus.

Figure 1. Top twenty word-forms, spoken and written

CANCODE spoken

| Rank | Word | Number of occurrences |
|------|------|-----------------------|
| 1 | THE | 34,951 |
| 2 | I | 30,480 |
| 3 | AND | 28,023 |
| 4 | YOU | 27,306 |
| 5 | TO | 23,152 |
| 6 | A | 20,386 |
| 7 | IT | 18,317 |
| 8 | THAT | 17,896 |
| 9 | OF | 16,768 |
| 10 | YEAH | 13,653 |
| 11 | IN | 12,248 |

| | | |
|----|------|--------|
| 12 | ER | 10,968 |
| 13 | MM | 10,563 |
| 14 | WAS | 9,840 |
| 15 | KNOW | 8,740 |
| 16 | IS | 8,456 |
| 17 | SO | 8,391 |
| 18 | IT'S | 8,004 |
| 19 | THEY | 7,783 |
| 20 | BUT | 7,343 |

CIC written

| Rank | Word | Number of occurrences |
|------|------|-----------------------|
| 1 | THE | 56,153 |
| 2 | OF | 29,101 |
| 3 | AND | 27,194 |
| 4 | TO | 27,000 |
| 5 | A | 23,427 |
| 6 | IN | 18,464 |
| 7 | I | 11,869 |
| 8 | IT | 10,212 |
| 9 | THAT | 9,925 |
| 10 | IS | 9,906 |
| 11 | FOR | 9,808 |
| 12 | WAS | 8,470 |
| 13 | YOU | 8,076 |
| 14 | ON | 7,467 |
| 15 | WITH | 7,170 |
| 16 | AS | 7,086 |
| 17 | BE | 6,275 |
| 18 | HE | 5,975 |
| 19 | AT | 5,414 |
| 20 | HAVE | 5,241 |

Making sense of these numbers means not only accounting systematically for the presence or absence of certain forms (such as explaining the high incidence of *know* in the spoken list in terms of the interactive discourse marker *you know*), but also appreciating the broader implications for *any* spoken variety of any language, of the fact that a spoken corpus focuses mostly round the words *I* and *you* (note how much lower they rank in the written), and has a very high proportion of vocabulary devoted to interactivity (*yeah, so, but* and the non-verbal tokens, *er, mm*). And these are only the first 20. Even from these rather semantically empty-looking words, significant qualitative insight can be gained.

With a common-sense, corpus-informed approach, we can achieve the following:

1. Reliable external evidence of usage that is not prey to the vagaries of intuition.
2. A deeper understanding of differences between speech and writing.
3. Insights into the cultural values that underpin language usage.
4. A resource for expert users, whether native- or non-native speakers, to consult and exploit in ways relevant to needs.
5. A database from which corpus-informed language teaching materials and other resources can be generated.

References

- Carter, R. A. & McCarthy, M. J. (1995) Grammar and the spoken language. *Applied Linguistics*, 16 (2), 141-158.
- Carter, R. A. & McCarthy, M. J. (1997) *Exploring spoken English*. Cambridge: Cambridge University Press.
- Halliday, M. A. K. (1966) Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J. R. Firth*. (pp.148-162). London: Longman.
- McCarthy, M. J. (1998) *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
- McCarthy, M. J. & O'Dell, F. (1999) *English vocabulary in use. Elementary*. Cambridge: Cambridge University Press.
- Sinclair, J. McH. (1966) Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J. R. Firth*. (pp.410-430). London: Longman.