# Corpus Linguistics and Lexicography: context, selection and interpretation[1]

**Margaret Rogers, Centre for Translation Studies, University of Surrey**

Re-reading some of Peter Newmark's "Translation Now" contributions to *The Linguist*, I am struck once again not only by his encyclopaedic knowledge of art, music and literature—at least, in the sense of what he calls "major works"—but also by his emotional sensitivity, balanced by a critical and committed eye. Such implicit distinctions reflect very well Peter's abhorrence of a value-free world, of cultural relativism (see for instance, Newmark 1998: 9, *et passim*) and of the failure to recognise what he called "truth," especially but not only in translation, which he described as a truth-seeking activity (see for instance, Newmark 2004, *et passim*). It is in this world of emotional absolutes that much of Peter's writing can be understood. However, his engagement with corpus linguistics and lexicography is interestingly ambivalent in many respects.

Peter's love of reference books pervades his writing (see for instance, Newmark 2006/7). As a great fan of dictionaries, he cherished what he saw as their truths. In so doing, he placed his trust in the lexicographer, or, since lexicographers have long worked in teams, more likely in several wordsmiths working together. For instance, he reports trusting what he calls the "dictionary meaning" of an unfamiliar sense of the word 'tweak' ("'He always tweaked the system'") more than he trusts his own "contextual interpretation" (Newmark 2008:28). Pointing out that the meaning of 'tweak' has been so since the 17th century, he cites the definition from an unspecified dictionary as follows: "'to (a) pull, twist, jerk, pinch, tug with short, sharp movements'" (*ibid.*). Yet lexicographers also work with contexts: it has long been lexicographical practice, clearly pre-dating the age of the world wide web (WWW), to work on the basis of citations selected by readers and transferred to "quotation slips" (see for example, Winchester 2003:97-105;113-4 on the making of the *Oxford English Dictionary* (OED)). Whilst the aim of the OED (historical; etymological) differed from that of later, corpus-based incarnations of the lexicographers' art (descriptive of modern usage), an evidence-based approach underlies both tasks. However, two crucial aspects of dealing with any evidence—be it evidence of the historical development of different word senses, or as a record of dominant senses in current usage—need to be considered: selection and interpretation.

The OED relied for its primary data ("the raw materials for the work," Winchester 1998[2]) on readers who volunteered to collect and submit quotations illustrating the use of English words by all writers of all ages and in all senses, each quotation being made on a uniform plan on a half-sheet of notepaper, that they might in due course be arrange and arranged alphabetically and by meanings (ibid.).

Twenty years after this first appeal for readers by the Philological Society, John Murray as editor of what would become the OED issued another appeal in 1879. Although the number of slips submitted in those two decades had been "voluminous" (Winchester 2003: 105), Murray became aware of a paucity of quotations submitted for the "'ordinary words of the language'" (*ibid.*). The first volunteer readers had preferred to focus on what they subjectively regarded as the more interesting words of the language, often meaning those not commonly used, e.g. 'abusion' as opposed to 'abuse,' as reported by Winchester (ibid.). Murray's 1879 appeal aimed to widen the scope of the quotation slips submitted hitherto. There are two levels of selection here: firstly, the range of documents which were to be read (orchestrated by Murray), and secondly, the choice of the reader according to what they felt was "rare, obsolete, old-fashioned, new, peculiar or used in a peculiar way" or in the use of ordinary words which were "'significant'" (Winchester 2003: 109, citing from Murray's Appeal). In other words, the readers were asked to interpret their sources according to Murray's rather vague criteria.

In modern corpus-based lexicography, the selection of documentation, realised in the so-called 'design' of a corpus depending on its intended purpose (see for instance, Ahmad and Rogers 2001:726-30 for a discussion of electronic corpora and corpus linguistics), is nothing other than a selection process. What differs in practice for the modern lexicographer is the possibility to produce through text-processing software, contexts for the totality of words in the corpus ordered alphabetically with frequency counts attached: a rather different exercise from the manual assignment of slips to 1,029 alphabetically ordered pigeon-holes (Winchester 2003: 104). However, the issue of judgment remains: whether the lexicographer is working with ten slips for a headword or 300 lines of a concordance, the task of grouping quotations/contexts according to sense remains an interpretive exercise, albeit made easier by various processing options when working digitally. In principle, the exercise still consists of organising the quotations/examples "that they might in due course be arranged and classified alphabetically and by meanings" (1859 version) or "that they might in due course be arranged and classified alphabetically and significantly" (1879 version).

The main difference between manual and computer-aided lexicography seems to lie not principally in selection or in interpretation but in the superior processing capacity of machines to sort the data for human interpretation, compared to that of humans manually sorting the data for human interpretation. In a digital medium, for example, text-processing options such as left-sort (see Krishnamurthy 1987:64) can be used in order to identify formal patterns in which the potentially polysemous search word is the phrasal head; such patterns—grouped together

automatically—can in themselves be indicative of different senses. The full context for the search word is also usually just a click away.

Nevertheless, despite praising the Cobuild English dictionaries for their "innovative descriptive definitions," word-frequency bands and up-to-date entries (Newmark 2006/7: 187), in a later edition of "Translation Today," Peter argues that "the value of corpus linguistics is overrated when it is claimed that a corpus represents the "real language," since many mistakes and misunderstandings arise from contextual guestimates" (Newmark 2010/11: 24). But it is not quite clear here how a qualitative distinction can be drawn between the interpretation of 'quotations' on slips on the one hand and of a (probably) greater number of 'examples' presented in a concordance on the other hand. Perhaps the key lies in the implied indiscriminate inclusion of "deficient" texts (Newmark 2003: 58-9) in a digital corpus. Such an interpretation of the 'raw materials' is indeed consistent with Peter's summative comment that "Cobuild's deficiency is that it fails to account for linguistic ignorance, carelessness and misinformation" (Newmark 2010: 31). Hence, Peter concludes that the "book and the up-to-date […] dictionary will always be indispensible" (ibid.). Indeed, but corpus-based lexicography (with the exception of scholarly historical studies) aims to be precisely that, i.e. up-to-date by making available to the lexicographer ample contemporary examples of how a particular word behaves in context.

The issue of 'context' is, however, a fraught one, ironically not least because of a variety of understandings about its meaning. Nearly a quarter of a century ago, Peter himself distinguished four types of context: the linguistic (e.g. collocations); the referential (the topic); the cultural; and the individual (the ideolect of the writer) (see Newmark 1988: 193). Adding to the mix, a further layer of complexity was added, distinguishing, for instance, between general-language words and special-language words: "some words are more context-dependent or bound than others..." (Newmark 1991: 87; see also Rogers 1999), so context, the argument seems to go, is a complex matter. Nevertheless, when comparing the individual interpretation of a single context with the multiple contexts provided by an electronic corpus, Peter concedes that contexts produced in quantity from the text corpus can provide convincing evidence of a particular word sense much more reliably than the "dear old" one-off context of personal judgment (Newmark 2006/7: 24). But the argument begins to become circular when Peter chides himself for misinterpreting a word in (the "jolly old") context when translating "because I never bothered to look the words up in an old-fashioned but up-to-date dictionary" (Newmark 2008: 32), a dictionary which is itself likely to be based on a corpus of texts, i.e. on language use, interpreted with the aid of processing software by lexicographers based on the jolly old context (or rather, lots of them).

The so-called true meaning or division into senses is an abstraction from use, not something that exists independently. It is part of the lexicographer's professional expertise to identify semantic shifts which are underway in the lexicon, to decide when such changes should be codified, and to distinguish such changes from outliers or idiosyncratic usage.

A final word on context—and something which I never discussed with Peter—concerns a middle way between consulting an up-to-date paper dictionary (i.e. not being "lazy," ibid.) and the individual interpretation of context, namely the use of on-line corpora during the translation process or in translator training (see for instance, Bowker 1998).

Using the British National Corpus on-line, the first ten contexts for Peter's problem word 'tweak' (underlined) are as follows (the underlined code at the beginning of each context can be clicked on to reveal the source on the BNC website, which in this case includes literary and journalistic texts):

1. **A0D 1359** She bounced to the mirror to powder and tweak for a moment, catching Conroy's eye and giving him a wink.
2. **AHC 1567** Later I turned and walloped back towards Berlin, by way of Magdeburg, with just the one tweak of the Tardis en route as I saw the signs for Potsdam.
3. **AJU 614** Variants include: Italian tweak—In Italy and Belgium, you can either chose the list of candidates your party has preselected for you, or tick your favourite and tweak his or her name up the preselected list.
4. **AKY 871** SCHOOLBOYS know how to tweak gadgets.
5. **AS3 861** Perhaps someone could lure some, sportsmen, in tweeds into its tenebrous depths, where we could leap down on them from a great height to tweak their noses and fill their plus fours with cornflakes.
6. **BMC 3276** Technology is 'old-style' Philips 16 bit, the system that preceded the current and more fashionable Bitstream process, but which has been found to be easier to 'tweak' for a desired effect.
7. **BMD 478** Many a promising liaison must have been nipped in the bud by Algernon being a duffer at flower arranging or breaking the significant corn stalk with a clumsy tweak; and many a passion choked to premature death because Augusta absentmindedly pinched the petals off her posy of moss rosebuds and myrtle.
8. **BN1 1384** A tweak at the edge of her consciousness said: prey.
9. **C87 2104** Put on those flying goggles, tweak that handlebar moustache and blow the enemy to hell!
10. **C9J 900** Probably nothing a quick tweak on the truss rod and bridge adjusters wouldn't sort out, but seeing as that would have involved taking the neck off, I didn't pursue it…

Apart from example 5, and arguably example 9 (a literal sense which Peter acknowledges knowing: "tweaking someone's ear/nose/cheek," Newmark 2007/8: 28), my interpretation of the examples indicates the

following sense: the act of making minor (see 'tweak for a moment,' 'a tweak at the edge,' 'a quick tweak') adjustments to objects in order to make them fit for purpose.

Whilst writing this contribution for the Special Issue of *JoSTrans* in honour of Peter Newmark and his work, I couldn't help feeling that I was in dialogue with Peter once again. If this means that I have told some corpus-savvy readers what they already know, I hope they will forgive me.

## References


- **Ahmad, Khurshid and Margaret Rogers** (2001). "The Analysis of Text Corpora for the Creation of Advanced Terminology Databases." In Sue-Ellen Wright and Gerhard Budin (eds) *Handbook of Terminology Management*. *Vol.2: Application-Oriented Terminology Management*. Amsterdam/Philadelphia: John Benjamins, 725-760.

- **Bowker, Lynne** (1998). "Using specialized monolingual native language corpora as a translation resource: A pilot study." *Meta* 43/4: 631-651.

- **Krishnamurthy, Ramesh** (1987). The Process of Compilation. In John M. Sinclair (ed.) *Looking Up. An account of the COBUILD Project in lexical computing.* London & Glasgow: Collins ELT, pp.62-85.

- **Newmark, Peter** (1988). *A Textbook of Translation*. Hemel Hempstead: Prentice Hall International.

- **—** (1991). Word and Text: Words and their Degree of Context in Translation. In: Peter Newmark *About Translation*. Clevedon: Multilingual Matters, pp.87-100.

- **—** (1998). *More Paragraphs on Translation*. Clevedon: Multilingual Matters.

- **—** (2003). "No Global Communication without Translation." In Gunilla Anderman and Margaret Rogers (eds) *Translation Today*. *Trends and Perspectives.* Clevedon: Multilingual Matters, pp.55-67.

- **—** (2004). "Non-literary in the Light of Literary Translation." *JoSTrans* Vol. 1:8-13.

- **—** (2006/7). "Translation Now" n° 41: "Reference Books." *The Linguist*, Vol. 45/6:185-187.

- **—** (2007/8). "Translation Now" n° 48: "The case of 'tweak.'" *The Linguist*, Vol. 46/6:27-8.

- **—** (2008). "Translation Now" n° 52: "Blame the context!" *The Linguist*, Vol. 47/4:30-32.

**—** (2010). "Translation Now" n° 63: "Picaresque and corpus linguistics." *The Linguist*, Vol. 49/3:30-31.

— (2010/2011). Translation Now No. 66: Corpus linguistics and translation: A summary, *The Linguist*, Vol. 49/6:24-5.

- **Rogers, Margaret** (1999). Translating Terms in Text: Holding on to Some Slippery Customers. In: Gunilla Anderman and Margaret Rogers (eds) *Word, Text, Translation*. Clevedon: Multilingual Matters, pp.104-16.

- **Winchester, Simon** (1998). *The Surgeon of Crowthorne*. Harmondsworth: Penguin.

- **—** (2003). *The Meaning of Everything. The Story of the Oxford English Dictionary*. Oxford: OUP.

**Website**

- **British National Corpus on-line**. www.natcorp.ox.ac.uk (consulted 12.12.2011).

[1] The idea for this topic originated in a discussion with Peter in the pub over a glass of wine, against a backdrop of a lively football match on a giant TV screen.

[2] Cited in Winchester (1998: front matter) from "An appeal to the English-speaking and English-reading public to read books and make extracts for The Philological Society's New English Dictionary" issued in 1859 by the Philological Society.