CORPUS LEXICOGRAPHY - The importance of representativeness in relation to frequency

Della Summers

This paper describes how the frequency of words in various corpora has influenced the presentation of phrases, the semantic description given in the definition, and the ordering of definitions in some entries in two recently published dictionaries. The two dictionaries are both for use by advanced foreign students of English. One, a dictionary of 80,000 words and phrases for general reference purposes, is the Longman Dictionary of Contemporary English (Third Edition, 1995). The other is a new type of dictionary, aimed at helping students to produce English themselves, using words and phrases in their appropriate meaning and context, the Longman Language Activator (1993), the first dictionary to be designated as a 'production dictionary'.

REPRESENTATIVENESS

Both dictionaries were compiled by lexicographers using on screen corpora which shared one central design principle to be representative of general language. This is a bold ambition some say one that is impossible to fulfil yet we at Addison Wesley Longman have always felt it essential to have a principled approach to corpus design, rather than following the more opportunistic path, which is usually overdependent on newspapers in electronic form or scripted or semiscripted radio broadcasts. For the purposes of designing these corpora, 'general language' has been seen as being reducible into individually distinct text types, such as fiction, scientific writing, and educational books in written language, and conversation and business meetings in spoken communication. The text types are conceptualized as amalgams of:

subject area in the written corpora (Fiction. Science. Politics. etc.)
medium (e.g. Books, Periodicals, or Unpublished in the written corpora),
level (high, medium, or low in written corpora),
context (Education, Leisure, Business, and Public or Institutional in spoken corpora)

This means that a work of literature such as *A Passage to India* has the features Fiction+ Book+ and High.

The demographic parts of our spoken corpora are arguably the most representative of all, designed as they are to mirror statistically the demographics of the target group, be it in the US or Britain, in terms of age, gender, region, and educational or social background. The design of such corpora, especially the Longman Lancaster Corpus of just under 30 million words and the British National Corpus of 100 million words, was much influenced by the work of Douglas

Biber, and Geoffrey Leech was actively involved in advising on their creation in his role as adviser on the Longman academic committee, Linglex, chaired by Lord Quirk, and as an academic partner in the British National Corpus *[1]*consortium.

The constitution of both these corpora is documented in detail elsewhere (Summers, 1993, Crowdy 1993, etc.), but the question is: how does this influence dictionary compilation?

[1] The British National Corpus is a collaborative initiative carried out by Oxford university Press, Longman, Chambers Harrap, Oxford University Computer Services, Lancaster University's Unit for Computer Research in the English Language (headed by Professor Geoffrey Leech), and the British Library. The project received funding from the UK Department of Trade and Industry and the Science and Engineering Research Council, and was supported by grants from the British Academy and the British Library

FREQUENCY AS AN AID TO THE LEXICOGRAPHER

One of the main reasons for wanting the corpus to be representative was so that reliable frequency statistics could be generated and used to aid the lexicographers in making the many linguistic judgments that lie behind the final entry for a word in the printed dictionary. All aspects of lexicography are influenced by frequency, but perhaps the most important in relation to the new edition of LDOCE was that the dictionary would attempt to truly follow frequency ordering of definitions in the entries, a feature of dictionaries much vaunted in the past, but rarely carried through in any objective or consistent way, due to the lack of reliable, balanced corpora.

The principle of frequency ordering dictates that the most frequent definition should come first, but it is not a simple operation to decide which meaning is the most frequent. One of the great advances of the BNC is that it is reliably tagged grammatically, using the latest version of the CLAWS tagging system developed under Geoffrey Leech's guidance at Lancaster. This means that we can now produce frequency counts which split words with more than one part of speech such as **pull** the verb as opposed to **pull** the noun (see later **Frequency marking**). However, no corpora are as yet tagged with meaning tags in the same way as they can be with grammatical tags. If we take the primary activity of lexicography to be the identification of the meanings of words, we can see that a parallel process is going on, with the lexicographer trying to base her analysis of the meanings of the word on frequency while at the same time framing the definitions and deciding on the meaning splits that she will eventually represent in the definitions in the dictionary.

Furthermore, experience has shown that even when working with corpora that are all based on the principle of representativeness, different corpora will present the lexicographer with different frequencies for words and repeated strings of words. There is therefore still a need to temper raw statistical information with intelligence and common sense. The corpus is a massively powerful resource to aid the lexicographer, which must be used judiciously. Our aim at Addison Wesley Longman is to be corpus based, rather than corpus bound.

FREQUENCY ORDERING AND COLLOCATIONS IN LDOCE

Because the users of the dictionary are advanced learners of English, collocation was an area of language that had been identified for special attention, but when we set out to make what we inelegantly term 'phraseology' into one of the predominant new features of LDOCE3, we had little idea of just how far reaching this policy would turn out to be for the look of the dictionary, for the organization of the entries, and for the form of the definitions themselves.

Preliminary work done by Penny Stock and Sheila Dignen into some of the most frequent words of the language, such as **day** and **eat**, alerted us to a fact that we should already have known: that is, that many such words were not just frequent by virtue of their single word uses in the corpus *[2]*, but also because they often occur in so many set phrases or chunks of language, like **one day, the other day, some day** and so on. It is commonly opined by corpus workers that although things that are discovered from corpus analysis are 'obvious', they only become obvious once the corpus has revealed them to us, i.e. they are not reliably recovered from the lexicographer's innate knowledge and understanding of the language. The 'discovery' that collocations add to the frequency of some words is such a case.

It was decided to recognize the phraseological behaviour of many words by entering them in their phrasal form in the dictionary, although still retaining them under a single word headword. We termed these phrases 'lexical units' or 'lexunits', believing that they function both semantically and grammatically as units. Quite often this resulted in the lexical unit being the first definition in the entry as the most frequent meaning. For example, under the headword liable you will find:

[2] (at this point, the Longman Lancaster Corpus)

1 be liable to do something to be likely to do or say something or to behave in a particular way, especially because of a fault or natural tendency: *The car is liable to overheat on long tops.* **2** [not before the noun] legally responsible for the cost of something: [**+for**] *Tenants have legal liability for any damage they cause.* **3** likely to be affected by a particular kind of problem, illness etc: [**+to**] *You're more liable to injury when you don't get regular exercise.* **4** *law* likely to be legally punished or forced to do something by law: [**+to**] *Anyone found trespassing is liable to a maximum fine of $100.* [**+for**] *All males between 18 and 60 are liable for military service.*

The phrase is entered as the first definition because it was identified as the most frequent use on the corpus *[3]*. The frequency of the word **liable** is 2139 occurrences out of 88,361,710 in the written part of the BNC as accessed by our concordancing system. Nearly half of those occurrences are followed by the word **to**, either as an infinitive marker or as a preposition, with 460 uses being followed by the preposition **for** Semantic and collocational anaylsis of the corpora produced the meaning splits as shown above, with the infinitival construction being shown to be the most frequent. This was borne out by the frequencies in the smaller Longman

Lancaster Corpus, which unlike the BNC contains no newspaper material. (The BNC written part is made up of 31% news media. The 10 million word spoken component of the BNC, which we access as a separate corpus, only provided 48 uses of the word **liable**, but of those 18 (or 37.5%) showed the infinitive complement. In this, as in other instances, the spoken corpus influenced the ordering of the senses in the entry and the lexical unit was given as the most frequent sense and therefore the first definition. Other cases of the phrase being shown as the first definition include **be on the lookout for, sectarian violence/conflict/murder** (showing selection restrictions for an adjective), and **arouse hope/interest, expectations** etc. (as an example of typical objects of a verb). Interestingly, the second sense of **arouse** is also a lexical unit with a different set of object preferences, **arouse anger/fear/dislike etc.**

UNITARY DEFINITIONS OF LEXICAL UNITS

Presenting the thing to be defined (i.e. the definiendum) as a lexical unit often radically alters the form of the definition, almost always with an increase of clarity and elegance. For example, these definitions:

miracle n [C] **1** something lucky that you did not expect to happen or did not think was possible: *By some miracle, we managed to catch the plane.* **it is a miracle (that) ... 3 miracle drug/cure** a very effective medical treatment that cures even serious diseases **4 work/perform miracles** to have a very good effect or result: *Maybe you should try yoga it worked miracles for me.* **5 a miracle of engineering/design etc.** something that is produced or invented that is a very impressive example of a particular quality or skill: *This new electronic notebook is a miracle of miniaturization.*

SPOKEN PHRASES IN THE ACTIVATOR

Very similar techniques to those described above were first used in the *Longman Language Activator*, the dictionary in which we first used the spoken corpus of the BNC. Spoken language was particularly relevant in this book, since its aim was to enable advanced students of English to 'activate' their passive knowledge of English and to use it more effectively in both speech and writing. The influence of the spoken corpus can be seen particularly, although not exclusively, in concepts with a functional or speech act aspect, such as PROMISE **(I promise, cross my heart**, and **I give you my word**), SOON **(as soon as possible, as soon as you can**, and **the sooner the better)**, and DON'T CARE **(not care, not give a damn, couldn't care less, be past caring, for all I care, who cares? so what?** and **shrug off**).

In the Activator, the individual semantic units are known as exponents, so that for all I care is an exponent of the concept DON'T CARE. All the above mentioned exponents would be covered in some form in dictionaries published before the Activator, but what is new is the inclusion of exponents made up mainly of delexical words such as **say what you like** under SURE or **you only have to...** under OBVIOUS, which make their first appearance in a published dictionary due to our use of the spoken corpus:

you only have to... you say that you only have to look at something, read something etc. when you think it is so obvious that anybody will notice it: *You've only got to look at the adverts for government jobs to realize that these people are paid too much....*

[3] The corpus seen as a collection of corpora, including in this case the BLOC and the Longman Lancaster.

FREQUENCY MARKING

Previously, I mentioned the value of the CLAWS grammatical tagging in generating frequency information from corpora. In LDOCE3, where different parts of speech are separate homographs, we have indicated the 3000 most frequent words of English in both spoken and written language, broken down by part of speech. This process will be described in detail at a later time, but what seemed to be a simple exercise to identify the first one thousand, the next thousand, and then the third thousand turned out to be a far more complex procedure than we envisaged at the beginning, involving a series of judgments, rather than taking raw frequency alone *[4].* These judgments included deciding which categories of word to exclude from the count, for example we began by excluding all proper names and place names, and checking the corpus for the number of texts or contexts that a word appeared in. The word **bid** as a verb, for example, has a frequency of 226 in the 10 million word spoken British corpus, which on raw frequency would have placed it in the top 3000 spoken words, earning it the **S3** label in the dictionary. However, checking of the range of spoken sources showed that all the instances of this in the spoken corpus were from one source an auction. There were no instances of the verb use in the 5 million words of general conversations in the demographically collected part of the spoken corpus and none at all in our American spoken corpus so it was decided not to give the verb a spoken frequency tag at all. The noun for bid, on the other hand, with a frequency of 3144 in the 90 million words of the written part of the BNC, did receive a tag to indicate that it was within the 3000 most frequent words in written material, in examples like: *the highest bid for the house, rival bids for the contract, and a bid for power.*

[4] For a comparison, see the frequency marking in the COBUILD Dictionary (Second edition, 1995), which gives newscaster as one of the 800 most frequent words of English, perhaps due to its dependence on news media source material, and shows all nationality names, such as Yemeni, as being in the same band of frequency as American.

CONCLUSION

Frequency is a powerful tool in the lexicographer's arsenal of resources, allowing her to make informed linguistic decisions about how to frame the entry and analyse the lexical patterns associated with words in a more objective and consistent way. However, in dictionary making editorial judgment is of paramount importance, because blindly following the corpus, no matter how carefully it may be constructed to represent the target language type accurately, can lead

to oddities. We expect our motto: 'Corpus based, but not corpus bound' to hold good for many years to come.

**Taken from the book 'Using Corpora for Language Research.' Harlow: Longman 1996*

BIBLIOGRAPHY

Burnard, L. (1995).

Users' Reference Guide to British National Corpus. Oxford: Oxford University Computing Services.

Crowdy, S. (1993)

Spoken Corpus Design. *Literary and Linguistic Computing.* Vol.8` No. 4.

Rundell, M., and Stock. P. (1994).

The corpus revolution. *English Today*. Vol.8, No.4.

Summers, D. (1993).

Longman Lancaster English Language Corpus Criteria and Design. *International Journal of Lexicography,* Vol.6, No.3.

Summers, D., Rundell M., et al

Longman Dictionary of Contemporary English.

Harlow: Longman. 1995.

Summers, D., Rundell M., et al

Longman Language Activator.

Harlow: Longman. 1993.