Adam Kilgarriff
Longman Dictionaries

The Myth of Completeness and Some Problems with Consistency

(The Role of Frequency in Deciding What Goes in the Dictionary)

Abstract

At the core of lexicography there is an ill–acknowledged, subjective notion of importance. Important words need fuller treatment. When people talk about consistency in dictionaries, this wrinkle is often overlooked. There is an analogous situation in theoretical lexicology and NLP, where the lure of elegant, rule–based theories has taken precedence, and "importance" largely ignored. If lexicography is to become more objective, the concept of importance must be closely studied with a view to finding an objective surrogate for it. This will come from corpora. As yet we have very little idea of which facts from which corpora are relevant, but if we do not take on the challenge, lexicography is forever doomed to ineffable subjectivity.

1. Introduction

As every scrabble–player and crossword enthusiast knows, the dictionary is – or damn well ought to be – complete, listing everything that is an English word (and nothing that isn't).[1] In the literature, one finds rather more considered voices asserting that a dictionary, or dictionaries in general, are incomplete in a particular area – or, more frequently, that they are inconsistent, with words displaying similar behaviour receiving different treatments, this said with shaking head and expression of world–weary disillusion. The argument of this paper is that, firstly, completeness is a myth. A dictionary can no more be complete than incomplete; the concept simply does not apply. And secondly, claims of consistency or inconsistency are less straightforward than they might seem, and are in general unproven until grounded in corpus evidence.

2. The evidence

A first indication of the problem is exposed by comparison of the following two entries from LDOCE2 (Summers 1987):

whisky 1 a strong alcoholic drink 2 a glass of whisky
bourbon a type of whiskey

"Why", the advocate of completeness and consistency might say "is the 'glass–of' sense of *bourbon* not presented? The dictionary is incomplete. It is also inconsistent. The two words can both be used in both ways, so should receive the same treatment in the dictionary." ('Alternations' such as that between the 'drink' sense of drink–words and the glass–of sense are discussed, under various names by various authors, inter alia Apresjan (1973), Leech (1981), Ostler and Atkins (1991), Kilgarriff (1992).)

A second example: *opportunity, privilege* and *indignity* are all nouns with interesting subcategorisation behaviour. But whereas *opportunity* has over 2000 entries in the Longman Lancaster corpus, *privilege* has 300, and *indignity* just 51. The dictionary user is much more likely to encounter – or, for the non–native speaker, to want to know how to use – the full range of subcategorisations with *opportunity* than with *privilege* or *indignity*. There is a strong case for giving *opportunity* the fullest treatment, *privilege* an intermediate one, and *indignity*, a summary one.

Of course every lexicographer recognises the problem and understands that, in short, more important words require fuller treatment. But every lexicographer also knows that there is nothing black and white about what is an important word, and many of the most difficult lexicographic judgements lie around questions of "is this sense (or multi–word unit or grammatical pattern) sufficiently important to require its own treatment?" Because 'importance' is a matter of degree, well–designed dictionaries have a variety of strategies for indicating various shades of grey. In LDOCE2, a sense that is the result of an alternation is rolled into the same sense as the primary meaning for the word (using a bracketing convention) where the alternate meaning is not very important. Where it is more important, it gets its own sense. In the Longman Language Activator, a more important multi–word unit is treated as a distinct headword with a definition, whereas a less important one is presented in bold, sometimes with a gloss, with an example but without a definition. In Kilgarriff (1992) I describe eight formally distinct strategies used in LDOCE2 for indicating different shades of grey.

3. Theoretical considerations: rule systems

Completeness is a well–defined term in the mathematics of rule–systems. A theory is complete if every true theorem can be derived, from axioms and rules of inference. Since Chomsky's Aspects, the study of syntax has been virtually synonymous with the study of a particular variety of rule systems, so the mathematics of rule systems has been central to the field of study.

The competence/performance distinction allows people working in this tradition to view the data as evidence for a rule system of one kind or another. For the study of syntax, the competence/performance distinction has been amply defended and the paradigm of syntax–as–rule–system has been highly successful. But the distinction does need a lot of defending; it is an

idealisation which allows us to disregard some of the things people say which might cause problems for our theory, simply by throwing them in the rubbish bin of 'performance'. If the distinction is too easily invoked, it undermines objectivity and makes theories unfalsifiable. For syntax, the distinction has proved its worth (though see Sampson (1987) for arguments against). For lexis, it has not.

Rule systems are very attractive to researchers. They allow us to describe and encode a range of phenomena elegantly and concisely. They are particularly enticing for NLP: a small core lexicon can describe a large number of word senses if it includes rules which multiply a single lexical entry out as a number of senses. For example, rule–based procedures can add in the 'glass–of' senses of *bourbon* and *whisky* when the core lexicon contains only the 'liquid' sense. This approach, espoused in, for example, Pustejovsky (1991) and Levin (1993), offers the prospect of greatly increased coverage, efficiency savings, and theoretical elegance. Does it offer the prospect of a complete lexicon?

It would work like this. Each lexical entry would have, in addition to its core meaning, an account of its class membership (or memberships). For each class, the alternations (whereby, for example, the 'liquid' sense gives rise to the 'glass–of' sense for all 'drinks' words) are listed. Provided we capture all the class–memberships for all the words, and all the alternations that apply to each class, the fully multiplied–out dictionary would be complete.

The prospect is alluring. But it is grounded in the competence/ performance distinction and a model of lexis as rule–system. When Morticia pours herself a hemlock from the Addams Family bar, *hemlock* is operating in a 'glass–of' sense. But it would take a large amount of forethought on a lexicographer's part to classify *hemlock* as a 'drink' word, in order that it might participate in the alternation. And if *hemlock* is to count as a drink, what other unusual word–uses must we anticipate? There is a risk that we shall want to say any word might be used for anything. Anything goes. Our lexicon is quite useless; it tells us not only that *horse* can mean 'horse', but also that it can mean 'cow', 'cheese', 'lexicon' or anything else.

3.1 Levin"s *English Verb Classes and Alternations*

The rule–system approach will only work if both words and alternations can be satisfactorily classified, and a satisfactory method is found for identifying nonce cases, like Morticia's use of *hemlock*, so they can be set aside for separate treatment. One substantial piece of work that addresses two of these three issues is Levin's *English Verb Classes and Alternations* (1993). Levin has classified over 3000 English verbs into 192 classes, and has identified 80 alternations. Associated with each class is a list of alternations that can be applied, and a list of those that cannot. Levin's hypothesis is that a verb's meaning determines its syntactic behaviour, so the alternations she considers are those which relate two different syntactic frames for a class of

verbs, as in the relation between the following:

Martha carved a toy out of wood for the baby.
Martha carved the baby a toy out of wood.

Levin's work is a major contribution to our understanding of the behaviour of English verbs, and is a resource for lexicography and NLP alike. We join her in believing it will "pave the way toward the development of a theory of lexical knowledge" (1993: 1). But work of this kind requires a complement. She is committed to the notion of lexis as rule–system, and the competence/performance distinction, arguing that

> "native speakers can make extremely subtle judgments concerning the occurrence of verbs with a wide range of possible combinations of arguments and adjuncts in various syntactic expressions." (1993: 2)

Her discussion makes no mention of nonce cases, or the dubious status of cases where native speakers are less than unanimous. She is addressing those areas of lexical semantic knowledge most closely linked to syntax, and correspondingly, the methods and assumptions used for studying syntax serve her well. The only alternations she addresses are those manifested in distinct subcategorisation frames, so she has chosen a domain where the identification and individuation of alternations is relatively straightforward. But a theory of lexical knowledge must cover also those areas where alternations are not so easily identified, and must make sense of the kline from standard cases to nonce cases and one–offs.

4. The Horns of the Dilemma

"There are two kinds of science: physics and stamp collecting" (Rutherford). In this context, rule–systems bear the hallmark of physics. They allow for the concise statement of generalisations, and have predictive power. If they are inappropriate, must we be stamp–collectors, merely recording lexical facts, generalisations forbidden, immune to accusations of inconsistency because, with generalisations renounced, there are never any grounds for asserting consistency or inconsistency? If all claims that two words fall in the same class are suspect, debates concerning whether two words ought to be treated similarly are of little concern.

The theoretical issues lead back to the lexicographical question. There are rules and generalisations at play in lexis, and these have a role in theoretical lexicology, practical lexicography and NLP. But their role is not un–trammelled: it is constrained by the elusive property of importance. Because *whisky* is more important than *bourbon*, it is reasonable for LDOCE2 to describe the glass–of sense of the one but not the other. Because 'drink' is a more important classification for *bourbon* than for *hemlock*, it is

appropriate for an NLP lexicon to include the 'glass–of' sense among the implicit, rule–derived senses of the former but not the latter.

5. Enter the corpus

The advent of the corpus offers a promise of objectivity, and thus theoretical well–foundedness. In short, importance is frequency. But which frequency? In what corpus? The corpus has opened a hornet's nest.

The most obvious and pressing question is representativeness. If statements about general language are to be based on frequencies in a particular corpus, then if that corpus is skewed, so are the claims made on the basis of it. For general language, it is far from clear what larger population a sample should be representative of. These are urgent questions currently receiving a substantial amount of attention (Biber 1993; Summers 1993). But there are many other questions which would remain even if the corpus were representative.

At Longman we have experimented with the idea that the number of corpus lines a lexicographer should look at should be related to the frequency of the word. High frequency words tend to have more meanings, be more 'important', and are worthy of more corpus study. This seemed straightforward. I gave the matter some thought, determined that a logarithmic relationship was appropriate, and drew up a table stating, for corpus frequency X, what sample size Y should be. But when I tried to apply it, it was rapidly thrown back at me. I had not taken the sophistication of how lexicographers work with the corpus into account. For most words, a substantial proportion of their corpus lines can be accounted for by a small number of collocates. Thus half the corpus lines for *ranks* are accounted for by a preceding *close* or *closed*. From the point of view of not missing interesting corpus lines, it is the lines not accounted for by *close* or *closed* which must be closely studied: once the set expression is noted, there is no need for the lexicographer to spend much longer looking at examples of it. So, for purposes of focusing the lexicographer's attention to maximum advantage, X should be the frequency of *ranks* appearing without a preceding *close* or *closed*, and Y should be drawn from that population. But that population cannot be defined for the general case, and even where it can be defined it cannot readily be counted. The moral of the story is that frequencies rarely have a simple story to tell, and while lexicography desperately needs them, they raise innumerable difficult practical and theoretical questions.

6. Conclusion

Completeness is not the sort of thing that applies to dictionaries. Scepticism is in order in relation to some accusations of inconsistency. Is LDOCE2 inconsistent between its treatment of *bourbon* and *whisky*? Only

on a narrow-minded and, in the last analysis, untenable account of consistency. We need new models of what it is for a dictionary to be consistent, which go beyond the model offered by boolean rule-systems and take corpus evidence into account. We have as yet very little understanding of how corpus frequencies might be integrated with rule-systems. But at present, at the foundations of lexicography we find the shifting sands of 'importance'. Only as we clarify this notion, by finding its objective correlates in the corpus, will the foundations be secure.

Notes

1 Or any other language: my work has been on monolingual English dictionaries, so here I apologise for speaking as if English were the only language.

References

Apresjan, J. 1974. "Regular Polysemy". *Linguistics* 142: 5–32.

Biber, D. 1993. "Using Register–Diversified Corpora for General Language Studies." *Computational Linguistics* 19(2): 219–242.

Kilgarriff, A. 1992. *Polysemy.* Dphil Thesis, University of Sussex.

Leech, G. 1974. *Semantics.* Penguin.

Levin, B. 1993. *English Verb Classes and Alternations.* University of Chicago Press.

Ostler, N. and Atkins, B. T. S. "Predictable Meaning Shift: Some linguistic properties of lexical implication rules." In J. Pustejovsky and S. Bergler (eds.) *Lexical Semantics and Knowledge Representation:* ACL SIGLEX Workshop, Berkeley, California.

Pustejovsky, J. 1991. "The Generative Lexicon." *Computational Linguistics* 17(4): 409–441.

Sampson, G. 1987. "Evidence Against the Grammatical–Ungrammatical divide." In Meijs, W. (ed.) *Corpus Linguistics and Beyond.* Rodopi, Amsterdam.

Summers, D. 1987. *Longman Dictionary of Contemporary English, New Edition.* (LDOCE2) Harlow: Longman.

Summers, D. 1993. "Longman/Lancaster English Language Corpus – Criteria and Design." *International Journal of Lexicography* 6(3): 101–208.