How many words?

It is often said that the English language is particularly rich in vocabulary, but to make such a statement we need to know what words to count and what counts as a word.

DAVID CRYSTAL

How many words are there in English? And how many of these words does a native speaker know? These apparently simple little questions turn out to be surprisingly complicated. In answer to the first, estimates have been given ranging from half a million to over 2 million. In answer to the second, the estimates have been as low as 10,000 and over ten times that number. People are, it seems, quite happy to drop all kinds of figures into their lectures and publications (see Panel 1). The figures give the impression of great precision – though it should be noted that they are usually accompanied by such emptying expressions as 'approximately', 'on average', or 'it is thought'. Nonetheless, the vagueness does not stop organizations offering courses and exercises (at a price) that will enable readers to 'increase their word power' – without ever providing these readers with the opportunity of discovering what their current word power actually is.

How can we throw light on this apparently confusing area? Let us begin with the question of how many words there are in English – a topic which has attracted almost as many estimates as estimators. The question is complex for two reasons. It partly depends on what you count as an English word, and partly on where you go looking for them.

What counts as a word?

Consider the problems, if someone asked you to count the number of words in English. You would immediately find thousands of cases where you would not be sure whether to count one word or two. In writing, it is often not clear whether something should be written as a single word, as two words, or hyphenated. Is it *washing machine* or *washing-machine*? *school children* or *school-children*? *flower pot*, *flower-pot* or *flowerpot*? Would you count all the items beginning with *foster* as new words: *foster brother*, *foster care*, *foster child*, *foster father*, *foster home*, etc? Or would you treat them as combinations of old words: *foster* + *brother*, *care*, and so on. This is a big problem for the dictionary-makers, who often reach different conclusions about what should be done.

What would you do with *get at*, *get by*, *get in*, *get off*, *get over*, and the dozens of other cases where *get* is used with an additional word. Would you count *get* once, for all of these, or would you say that, because these items have different meanings (*get at*, for example, can mean 'nag'), they should be counted separately? In which case, what about *get it?*, *get your own back*, *get your act together*, and all the other 'idioms'? Would you say that these had to be counted separately too? Would you count *kick the bucket* (meaning 'die') as three familiar words or as a single idiom? It hardly seems sensible to count the words separately, for *kick* has nothing to do with moving the foot, nor is *bucket* a container.

If you let the meaning influence you (as it should), then you will find your word count growing very rapidly indeed. But as soon as you do this, you will start to worry about other meanings, even in single words. Is there a single meaning for *high* in *high tea*, *high priest* and *high season*? Is the *lock* on a door the same basic meaning as the *lock* on a canal? Should *ring* (the shape) be kept separate from *ring* (the sound?) Are such cases 'the same word with different meanings' or 'different words'? These are the daily decisions that any word-counter (or dictionary compiler) must make.

Whose English are we counting?

Sooner or later, the question would arise about the *kind* of vocabulary to include in your count. There wouldn't be a difficulty if the words were part of standard English – used by educated people throughout the English-speaking world. Obviously these have to be counted. But what about the vast numbers of words which are not found everywhere – words which are restricted to a particular country (such as Canada, Britain, India, or Australia), or to a particular part of a country (such as Wales, Yorkshire or Liverpool)?

They will include words like *stroller* (= push-chair) and *station* (= stock farm) from Australia, *bach* (= holiday cottage) and *pakeha* (= white person) from New Zealand, *dorp* (= village) and *indaba* (= conference) from South Africa *cwm* (= valley) and *eisteddfod* (= competitive arts festival) from Wales, *faucet* (= tap) and *fall* (= autumn) from North America, *fort-*

Varying estimates

1

Shakespeare had one of the largest vocabularies of any English writer, some 30,000 words. (Estimates of an educated person's vocabulary today vary, but it is probably about half this, 15,000.) (Robert McCrum, et al, *The Story of English*, 1986, p. 102)

He [Shakespeare] has the largest vocabulary of any writer in English, approximately 34,000 words, which is about double what an educated person uses today in their lifetime. (John Barton, in *The Story of English* episode 3)

At two years old the average vocabulary is about three hundred words. By the age of five it is about five thousand. By twelve it is about 12,000. And there for most people it rests – at the same size repertoire employed by a popular daily newspaper. (Jane Bouttell, *The Guardian*, 12 August 1986)

Graduates have an average vocabulary of about 23,000 words, fostered, I would contend, by intensive tutoring. (Jane Bouttel, also *The Guardian*)

night (= two weeks) and *nappy* (= baby wear) from Britain, *loch* (= lake) and *wee* (= small) from Scotland, *dunny* (= money) and *duppy* (= ghost) from Jamaica, *lakh* (= a hundred thousand) and *crore* (= ten million) from India, and many more.

Regional dialect words have every right to be included in an English vocabulary count. They are English words, after all – even if they are used only in a single locality. But no one knows how many there are. Several big dictionary projects exist, cataloguing the local words used in some of these areas, but in many parts of the world where English is a mother-tongue or second language, there has been little or no research. And the smaller the locality, the greater the problem. Everyone knows that 'local' words exist: 'we have our own word for such-and-such round here'. Local dialect societies sometimes print lists of them, and dialect surveys try to keep records of them. But surveys are lengthy and expensive enterprises, and not many have been completed. As a result, most regional vocabulary – especially that used in cities – is never recorded. There must be thousands of distinctive words inhabiting such areas as Brooklyn, the East End of London, San Francisco, Edinburgh and Liverpool, none of which has ever appeared in any dictionary.

The more colloquial varieties of English – and slang, in particular – also tend to be given inadequate treatment. In dictionary-writing, the tradition has been to take material only from the written language, and this has led to the compilers concentrating on educated, standard forms. They commonly leave out non-standard expressions, such as everyday slang and obscenities, as well as the slang of specific social groups, such as the army, sport, thieves, public school, banking, or medicine. Eric Partridge once devoted a whole dictionary to this world of 'slang and unconventional English'. Some of the words it contained were thought to be so shocking that for several years many libraries banned it from their open shelves!

Keeping track of slang, though, is one of the most difficult tasks in vocabulary study, because it can be so shifting and short-lived. The lifespan of a word or phrase may be only a few years – or even months. The expression might fall out of use in one social group, and reappear some time later in another. Who knows exactly how much use is still made today of such early jazz-world words as *groovy*, *hip*, *square*, *solid*, *cat*, and *have a ball*? Or how much use is made of the new slang terms derived from computers, such as *he's integrated* (= organised) or *she's high res* (= very alert, from 'high resolution'). Which words for 'being drunk' are now still current: *canned*, *blotto*, *squiffy*, *jagged*, *paralytic*, *smashed* . . .? And how do we get at the vast special vocabulary which has not grown up in the drugs world? Word-lovers from time to time make collections, but the feeling always exists that the items listed are only the tip of a huge lexical iceberg.

DAVID CRYSTAL read English at University College London, and has since held posts in linguistics at the University College of North Wales, Bangor, and at the University of Reading, where he taught for twenty years. He works currently as a writer, lecturer, and broadcaster on language and linguistics, maintaining his academic links through an honorary professorship in linguistics at Bangor. He is the editor of *Linguistics Abstracts* and *Child Language Teaching and Therapy*. Among his recent publications are *Listen to Your Child, Who Cares About English Usage?*, and *Linguistic Encounters with Language Handicap*. His most recent book is the *Cambridge Encyclopedia of Language*.

Some marginal cases

Estimating the vocabulary size of English is further complicated by the existence of hundreds of thousands of uncertain cases – words which you wouldn't feel were part of the 'central' vocabulary of the language. On the other hand, you might well feel unhappy about leaving them out.

What would you do with all the abbreviations that exist, for example? A recent dictionary of abbreviated words (the impressive *Acronyms, Initialisms & Abbreviations Dictionary* published by the Gale Research Company, 11th edition, 1987) lists over 400,000 entries. It includes old and familiar forms such as *flu*, *hi-fi*, *deb*, *FBI*, *UFO*, *NATO* and *BA*. There are large numbers of new technical terms, such as *VHS* (the video system), *AIDS*, and all the terms from computerspeak (*PC*, *RAM*, *ROM*, *BASIC*, *bit*) and space travel (*SRB* – solid rocket boosters, *OMS* – orbital manoeuvring system, etc.) And there are thousands of coinages which have a restricted regional currency, such as *RAC* (= Royal Automobile Club), *AAA* (= Automobile Association of America), or reflect local organisations and attitudes – with varying levels of seriousness – such as *MADD* (= Mothers Against Drunk Driving) and *DAMM* (= Drinkers Against Mad Mothers).

Because these forms are dependent on 'bigger' words for their existence, you might well decide not to include them in your count. On the other hand, you could argue that they are often more important than the original words – and that the original words may not even be remembered or known (as many people find with such forms as AIDS). Personally, I would include them in my word count – but some dictionaries do not.

There are other marginal cases. What would you do with the names of people, places and things in the world? Should *London*, *Whitehall*, *Paris*, *Munich*, and *Spain* be included in your word count? You might think they should – especially knowing that many of these words are different in other languages (such as *München* and *España*). However, it isn't usual to include them as part of the vocabulary of English, because the vast majority can appear in *any* language. Whichever language you speak, if you walk down Pall Mall, you can refer to where you are by using the words *Pall Mall* in your own language. The old music hall repartee relied on this point:

A: I say, I say, I say. I can speak French.
B: You can speak French? I didn't know that. Let me hear you speak French.
A: Paris, Marseilles, Nice, Calais, Jean-Paul Sartre . . .

The same applies to the names of people, animals, objects (such as trains and boats), and so on. Proper names aren't part of any one language: they are universal. However, it's important to note the usages where these words do take on special meanings – as in *Has Whitehall said anything about this?*. Here, *Whitehall* means 'the government'; it isn't just a place name. Dictionaries would usually include this kind of usage in their list. But it's not at all clear how many uses of this kind there are.

Fauna and flora present a further type of difficulty. Around a million species of insects have already been described, for example. Which means that there must be around a million

2

Lexical coverage of three unabridged US dictionaries

A hint of the extent to which any given dictionary underestimates the total word-stock of English can be obtained from the table below, which lists the bold-face words found as initial items in the entries of three unabridged American dictionaries (variants later in the entry's opening line have been excluded). Of the 48 possible items listed, coverage ranges from 70% to 35%. Only nine words appear in all three dictionaries – less than 20% overlap. This figure is not much increased even if *RH*'s proper names are excluded from consideration.

The same story emerges if pairs of dictionaries are compared. There is an overlap of 13 between *W*III and *RH*, of 11 between *RH* and *WBE*, and of 10 between *W*III and *WBE*, suggesting that, if this sample is representative, the average overlapping coverage (as defined by headwords) between any two dictionaries might be as low as 25%.

Webster III	*Random House*	*World Book Encyclopedia*
saba	saba	
	Sabadell	
sabadilla	sabadilla	sabadilla
sabadine		
sabadinine		
sabaeanl	Sabaean	Sabaean
sabaean[2]		
	Sabah	
sabai grass		
		Sabaism
		Sabaist
sabakha		
sabal		
sabalo	sabalo	
sabalote		
sabal palmetto		
sabana		
	Sabaoth	Sabaoth
	Sabata	
	Sabatier	
	Sabatini	
sabathé's cycle		
sabaton	sabaton	
sabayon	sabayon	
sabbat	Sabbat	Sabbat
sabbatarian[1]	Sabbatarian	Sabbatarian
sabbatarian[2]		
sabbatarianism	Sabbatarianism	Sabbatarianism
sabbath	Sabbath	Sabbath
		sabbath
sabbath day[1]		
sabbath day[2]		
sabbatharian		
sabbath-day house		
sabbath-day's journey		Sabbath-day's journey
sabbathless	Sabbathless	Sabbathless
	Sabbathlike	
sabbathly		
sabbath school	Sabbath School	Sabbath School
sabbatia		
sabbatian[1]		
sabbatian[2]		
		sabbatic
sabbaticall	Sabbatical	sabbatical
sabbatical[2]		
		sabbaticals
	Sabbatically	sabbatically
	Sabbaticalness	
Total: 34	22	17

designations available to enable English-speaking entomologists to talk about their subject. How much of this can be included in our word count? The largest dictionaries already include hundreds of thousands of technical and scientific terms, but none of them includes more than a fraction of the insect names – usually just the most important species. Add this total to that required for birds, fish, and other animals, and the theoretical size of English vocabulary increases enormously.

In the light of these problems, it may not be possible to arrive at a satisfactory total for English vocabulary. But one thing is plain: the core vocabulary, as reflected in the entry totals cited for such works as the unabridged *Oxford English Dictionary* or *Webster's Third New International*, is a considerable underestimate (see Panel 2). These totals focus on a figure of about half a million. However, if we allow in some of the above categories, this figure must be increased by a factor of three or four. I would never want to go below one million, for an estimate of English vocabulary, and with very little persuasion I would readily accept two.

How large is your vocabulary?

There seems to be no more agreement about the size of an adult's vocabulary than there is about the total number of words in English. Estimates do indeed vary, as we have seen. Part of the problem, I imagine, is what is meant by 'educated'. But whether we are educated or not, how can we find out the truth of the matter?

We might tape record everything we said and heard for a month, or a year, and keep a record of everything we read and wrote. Then we could tabulate all the words, mark which ones we understood and which we failed to understand, and count up. But life is too short.

An alternative, which can be carried out in a couple of hours, gives a fairly good idea. You take a medium-sized dictionary – one which contains about 100,000 entries – and test your knowledge of a sample of the words it contains. A sample of about 2% of the whole, taken from various sections of the alphabet, gives a reasonable result. In other words, if such a dictionary were 2000 pages long, you would have a sample of 40 pages. Use the following procedure.

• It's wise to break this sample down into a series of selections, say of 5 pages each, from different parts of the dictionary. It wouldn't be sensible to take all 40 pages from the letter U, for instance, as a large number of these words would begin with *un-*, and this would hardly be typical. On the other hand, prefixes are an important aspect of English word formation, so we mustn't exclude them entirely. Similarly, it would be silly to include a section containing a large number of scientific words (such as the section containing *electro-*), or rare words (such as those beginning with X).

• One possible sample, which tries to balance various factors of this kind, would take sections of 5 complete pages from each of the following parts of the dictionary: C-, EX-, J-, O-, PL-, SC-, TO- and UN-. Begin with the first full page in each case – in other words, don't include the very first page of the C section, if the heading takes up a large part of the page; ignore the first few EX- entries, if they start towards the bottom of a page; and so on.

• Draw up a table of words like the one in Panel 3. On the left-hand side write in the headwords from the dictionary, as they appear. Do not include any *parts* of words which the dictionary might list, such as *cac-* or *-caine*, but *do* include words with affixes, such as *cadetship* alongside *cadet*, even if the former is listed only as *-ship* within the entry on *cadet*. In short, include all items in bold face within an entry. Include phrases or idioms (e.g. *call the tune*). Ignore alternative spellings (e.g. *caesarian/cesarian*.

• The table has two columns: the first asks you to say whether you think you know the word, from having heard or seen it used; the second whether you think you actually use it yourself in your speech or writing. This is the difference between *passive* and *active* vocabulary. Within each column, there are three judgments to be made. For passive vocabulary, you ask 'Do I know the word well? vaguely? or not at all?'. For active vocabulary, you ask: 'Do I use the word often? occasionally? or not at all?'. Place a tick in the appropriate column. If you are uncertain, use the final column. You may need to look at the definition or examples given next to the word, before you can decide. Ignore the number of meanings the word has: if you know or use the word in *any* of its meanings, that will do. (Deciding how many *meanings* of a word you know or use would be another – much vaster – project!)

A vocabulary estimate

3

Part of one person's vocabulary estimates, using the head words of the *Longman Dictionary of the English Language* (90,000+ headwords). + = known/used.

	KNOWN			*USED*		
	Well	*Vaguely*	*No*	*Often*	*Occasionally*	*Never*
cablese		+				+
cable stich	+			+		
cable television	+				+	
cable vision		+				+
cableway			+			+
cabman		+				+
cabob			+			+
Caboc			+			+
cabochon (noun)			+			+
cabochon (abverb)			+			+
caboodle	+				+	
caboose		+				+
cabotage			+			+
cab-rank	+			+		
cabriole			+			+
cabriolet			+			+
cabstand	+					+

• When you've finished, add up the ticks in each column, and multiply the total by 50 (if the sample was 2% of the whole). The total in the first column is probably an underestimate of your vocabulary size. And if you take the first two columns together, the total will probably be an overestimate.

This procedure of course doesn't allow for people who happen to know a large number of non-standard words that may not be in the dictionary (such as local dialect words). If you are such a person, the figures will have to be adjusted again – but that will be pure guesswork.

Here are the estimates for the first two columns, as filled in by a female office secretary in her 50s:

WORDS KNOWN

Well	*Vaguely*
30,050	8,250

38,300

WORDS USED

Often	*Occasionally*
16,300	15,200

31,500

The results are interesting. Note that passive vocabulary is much larger than active. This will always be the case. You will also find that it's easier to make up your mind about the words you definitely know than the words you frequently use.

Even allowing for wishful thinking, sampling bias, and other such factors, it would seem that some of the widely quoted estimates of our vocabulary size are a long way from reality. Comparisons with Shakespeare or other past writers are meaningless, given the enormous increase in English vocabulary since his day. What I would now very much like to know is (a) whether this procedure can be tightened up in some way, or whether a better procedure can be suggested? and (b) what range of totals emerge from people of varying backgrounds and ages? *ET* will publish in due course a range of vocabulary estimates from readers who have tried out the procedure for themselves (or, if they prefer, have tried it out on a 'friend'). If you do send in these details, please make sure you include data on age, educational background, and occupation, as well as the dictionary you used. The results will always be interesting, and may be surprising. If nothing else, it can provide you with a good topic for parties. There really isn't a way of capping such observations as 'I have an active vocabulary of approximately 38,600 words'. It will be a safe conversation-stopper – unless, that is, you encounter another *ET* reader at the same party. ET