How Many Words Are There?

Adam Kilgarriff

Abstract

How many words are there? We would like a nice answer - a number - but, I'm sorry to say, I am not going to give you one. The question does not stand up to scrutiny. There are many reasons why, from scientific specialisations to restaurant menus, and the chapter talks through a number of them, drawing data from chemistry, my local vegetarian restaurant, and very large databases of English to illustrate what happens when you start to enter the murky zone where things that might possibly be English words, start to be outnumbered by things that you think, really can't be.

1. Introduction

Words are like songs. The ditty a mother makes up to help her baby sleep, the number the would-be Rolling Stones belt out in their garage, the fragment in a strange dialect recalled by the octogenarian, these are all songs. The more you look, the more you find.

The dictionary, as an institution, is misleading. The big fat book has an aura of authority to it, carefully cultivated by its publishers. On the back covers of the dictionaries on my shelf we have "Full and completely up-to-date coverage of the general, scientific, literary, and technical vocabulary …", "No other single-volume dictionary provides such authoritative and comprehensive coverage of today's English", "The new authority on the world's language", "The most comprehensive and up-to-date picture of today's English". This is sales talk. They want to give their potential purchasers the impression that they have all the words in them (and more than their competitors). They also have numbers – always a bone of contention between the editorial department and the marketing department:

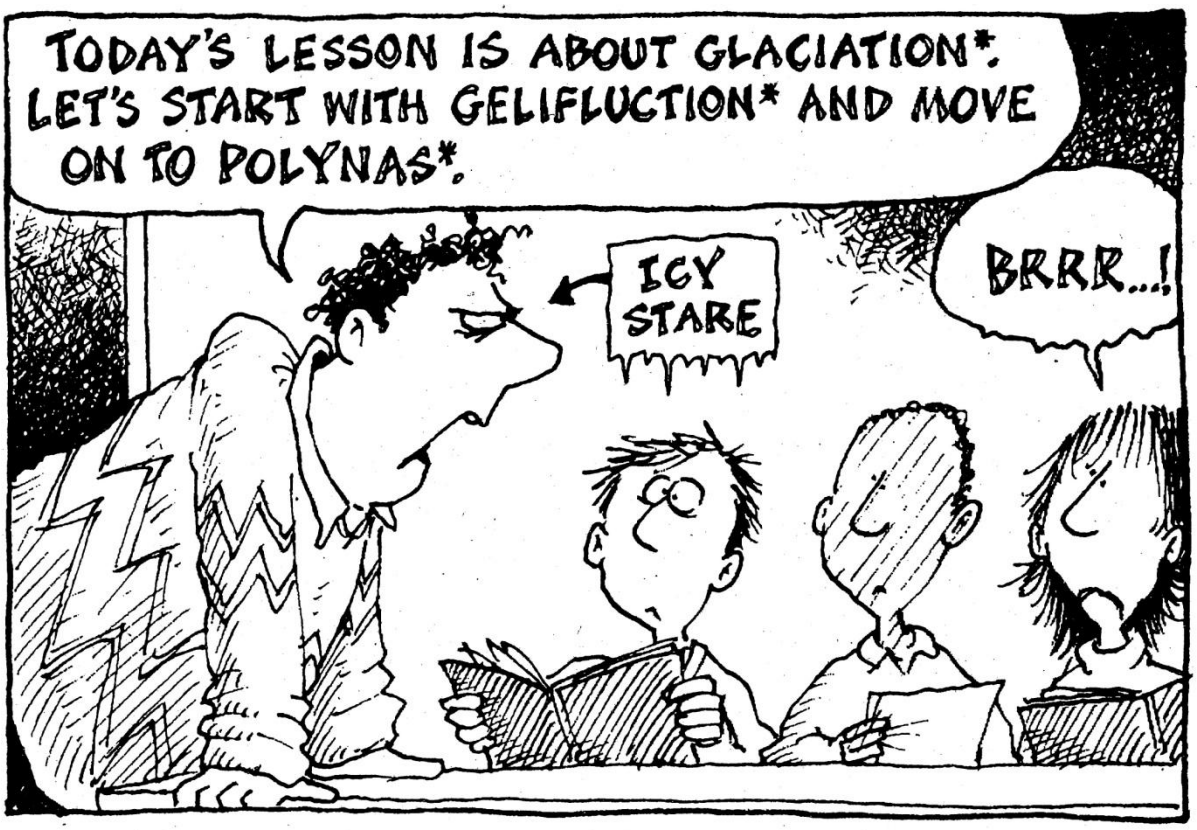
> Marketing: How many words are there, for the press release?
> Editor: Well, there are 57,000 full entries.
> Marketing: That's no good, Chambers and Websters both have far more.
> Editor: Well, we could count run-on items, the embedded compounds, phrasal verbs and phrases, that gets us up to 76,000.
> Marketing: Still not enough, I'm sure you can do better, what about these bolded bits in examples?
> Editor: But they're just common expressions, they are not even defined.

Marketing: Are you forgetting who pays your salary? We need to sell!

There is even something strange about the syntax. We don't say "Is it in a dictionary?", always "Is it in the dictionary." This is a triumph of marketing. Another word that works like that is *bible.* In the case of *bible,* it is reasonable to say that, at source, there is just one, and all editions, in all languages, are just versions of that. The use of *the* for *dictionary* suggests some Platonic ideal that any published items is a more or less true version of.

Dictionaries have a variety of uses. Consider Scrabble. The simple role of the dictionary in Scrabble is to say if a string of letters is a word. It can only do that by having all the words in it. Alongside word games, there is resolving family arguments. A dictionary that does not allow a protagonist to say "I told you so, it's not in the dictionary" is not worth the paper it is written on.

The impulse to document a language has much to do with comprehensiveness.



As the cartoon (from the 'Horrible Geography' series) indicates, there are a lot of words. All sorts of nooks and crannies of human activity have their own terms, not known to the general

public but nonetheless, straightforwardly and unequivocally, words of English. *Gelifluction* does not have an entry in the largest dictionary I had available to check, the Oxford English Dictionary, although it does occur (apparently mis-spelt *gelifiuction*) in an example sentence for the related word *solifluction*. *Polynya* (note difference of spelling) does have an entry. *Gelifluction* occurs just four times in a database of twelve billion words of text crawled for the web, *polynya* occurs 328 times, 105 times with the second 'y' and three times without it.

All this makes it hard to give a number. The primary reason is the sheer number of nooks and crannies of human activity that there are: how might we cover all of them? There are other reasons:

(a) Rules for making new words up. This is the province of derivational morphology and word formation rules (see Booij, this volume). Some specialisms even have for their own rules for generating an unlimited number of specialist words. The *Nomenclature of Inorganic Chemistry, IUPAC Recommendations* (Connelly 2005) is a collection of rules for naming inorganic compounds. If the rules are followed, then different chemists working independently will give the same name to a new compound according to its chemical composition, thereby reducing ambiguity and confusion. The rules sometimes give rise to terms with spaces in, sometimes to terms containing hyphens, brackets, numbers (Arabic and Roman), Greek letters, the + and - signs, and sometimes to long strings with none of the above. Examples (from wikipedia[1]) include *ethanidohydridoberyllium*, *bis(η^5-cyclopentadienido)magnesium*, *pentaamminechloridocobalt(2+) chloride*, *di-μ-chlorido-tetrachlorido-1κ^2Cl,2κ^2Cl-dialuminium*, and *Decacarbonyldihydridotriosmium*.

(b) Homonymy. Where there are two different meanings, when do we want to say we have two different words? Some cases are clear, e.g. *file* 'type of tool' and 'collection of documents', others less so (see Durkin, this volume).

(c) Multiwords. Do we allow in words written with spaces, like *all right*? When does a single word turn into a sequence of words, or *vice versa*? (See Wray, this volume, and Moon, this volume.)

[1] http://en.wikipedia.org/wiki/IUPAC_nomenclature_of_inorganic_chemistry_2005

(d) Imports. There can be uncertainty about the language that a word belongs to; when do words borrowed from other languages start to count? (See Grant, this volume, and Sorell, this volume).

(e)Variation: when do two different spellings, or pronunciations, start to count as two different words?

First, we present a little data, and then we say some more about imports and variation.

2. A little data

The question, "how many words are there?", may be asked of any language. All of the aspects discussed here relate to any language, though sometimes in different ways. Here, we mainly discuss English, with occasional reference to how different considerations play out differently in other languages.

enTenTen12 (Jakubíček et al 2013) is a database of twelve billion words of English gathered from the web in 2012. The twelve billion is the number of tokens, not types: that means that the 547 million occurrences of *the* count as 547 million, not as just one, as they would if I was counting types. To put it another way, how many words are there in "dog eats dog"? There are two possible answers: three, if I am counting tokens, but two, if I am counting types. The question "how many words are there?" clearly relates to types, not tokens.

Another ambiguity to draw attention to is between inflected forms of words and lemmas. Do *invade, invading, invades, invaded* count as forms of the same word, or as different words? If we say 'forms of the same word', we are talking about lemmas, or dictionary headwords. If we say 'four different words' we are talking about word forms. For English, the difference between the two is not so great since very few lemmas are associated with more than four forms (the standard number for verbs, like *invade*), with nouns having just two (singular and plural). For many languages, the numbers are higher, sometimes running into hundreds or thousands. In this section all discussions are of word forms, largely because they are easier to count (and counting them avoids complications like the errors made, particularly for rare words, by automatic lemmatizers).

There are 6.8 million different types in enTenTen12 (including only items comprising exclusively lower-case letters, separated by spaces and punctuation). Their distribution is

Zipfian: the commonest items occur far, far more often than most, and very many occur only once (see Sorell, this volume). Here there are 1,096 words that occur over one million times, and 3,745,668 words occurring just once. The distribution can be broken down as follows.

Frequency band	# words	Random sample from lower edge of frequency band)
1,000,000+	1096	active expensive floor homes prior proper responsible round shown title
1,000-999,999	60,789	ankh attunements diatom dithered limoncello mobilisations sassafras seemeth softgel uremic
100-999	109,362	alledge dwellin faceing finacee frackers neurogenetic sacralized shl symbole vigesimal
10-99	511,714	abbut arquebusses bundas carcer devilries feace hotu petronel taphophiles theaw
5-9	611,146	athambia dowter hazardscape humanracenow kernelled noatble producest stancher sullens trattles
4	307,309	boarwalk intercousre layertennis locutory meritest nonhumanistic pitiyankees scapularies starbeams uitrekenen
3	483,720	rokas faraa cuftucson cremosas topboard brahmanam samuebo messenblokken

		regenica
2	941,181	androgynized bolibourgeoisie lascomadres lowspot neoliberialism nonmorbid oapmaking projectst salesm whatsoevery
1	3,745,668	cirumscriptions digatel dramturgy figurability frelks inactivazed mixtore shunjusha teires wrider

At the one million point we have mainstream core-vocabulary words.

At the one thousand point we have

(a) words from specialist domains, found in large dictionaries:

- an *ankh* is an Egyptian symbol usually meaning ‘life’, or ‘soul’
- a *diatom* is a single cell alga
- *sassafras* is a species of tree with aromatic leaves and bark, and the extract drawn from it
- *limoncello* is an Italian alcoholic drink made from lemons. Also note that *limoncello* is on the margins of being a name, and in addition to 1000 lowercase occurrences, there are 729 capitalized. On the borderline between regular words and names, see Anderson (this volume), also see restaurants section below
- a *softgel* is an oral dosage form for medicine similar to capsules

(b) inflected forms for familiar, if not specially common, words: *attunements, dithered, mobilisations;* also *seemeth*, an archaic inflected form of a common word; and *uremic* (relating to the disease *uremia)*

At the 100 point we have

- *vigesimal*, a number system based on twenty, present in the larger dictionaries
- one simple spelling error, *faceing* (the target form was *facing* in all cases that I checked)

- *finacee*, target form: *fiancé, fiance, fiancée, fiancée*, depending on gender and the tricky business of how accented characters in imported words relate to English spelling. One thing is clear: the *a* should be before the *n.*
- spelling errors mixed with old or other non-standard forms: *alledge, dwellin.* A mixture is a case where some of the instances are of one kind, e.g., spelling errors:

> "If you hear or read anyone in the United States assert or alledge that we have a democracy , a representative democracy or anything short of a kleptocracy"

while others are of another kind, e.g., an old form:

> "That the Debts either by Purchase, Sale, Revenues, or by what other name they may be call 'd, if they have been violently extorted by one of the Partys in War, and if the Debtors alledge and offer to prove there has been a real Payment, they shall be no more prosecuted , before these Exceptions be first adjusted"

- a spelling error mixed with a foreign word: *symbole*
- inflected forms of derived forms of words: *frackers* is plural of *fracker*, "someone who fracks", where fracking is a process of extracting oil from underground reserves, currently a politically and environmentally contentious topic; *sacralized*, past tense of "made sacred" or "treated as sacred"
- a prefixed form: *neurogenetic* (where *neuro* is mid-to-low frequency prefix)
- *shl*: a mixture of programming language command, url-parts, shortened *shall*, abbreviations

At the ten point, *abbut, arquebusses, devilries*, and *taphophiles* are recognizably words of English, albeit obscure and/or mis-spelt and/or inflected/derived forms, while the remaining six are not even that, and so it is as we carry on down to the items occurring just once. These are like the residue at the bottom of a schoolboy's pocket: very small pieces of a wide variety of substances, often unsavoury, all mixed together, often unidentifiable. One would rather not have to look into them too closely.

In sum – at the top of the list – at least the top 1000 – we have core vocabulary. By the time we have reached 60,000 we have obscure vocabulary and marginal forms. Another 100,000 items, and dictionary words are thin on the ground, though we still often have their inflected and derived forms, and their mis-spellings. After a further half million, half the items no longer even look like English words, but are compounded from obscure forms,

typos, words glued together and other junk, and so on down to *bolibourgeoisie*, *whatsoever*, and *frelks.*

3. Imports

3.1 Restaurant English

As explained by Douglas Adams in the Hitchhikers' Guide to the Galaxy, a distinct form of mathematics takes over in restaurants at that moment when it comes to working out each person's contribution to the bill. Likewise, a distinct form of English. Let us make a linguistic visit to the grandest of our local vegetarian restaurants, 'Terre a Terre'. A sample of their menu:

> Red onion, mustard seed, cumin crumpets with coconut curry leaf and lime sabayon, ginger root chilli jam and a fresh coriander, mint salsa sas. Served with thakkali rasam of tamarind and tomato, nimbu bhat cardamom brown onion lemon saffron baked basmati rice with our confit brinjal pickle.

The peculiar thing about this form of English is that, while the language is English, most of the nouns don't seem to be. They form a subtext to the history of the population itself, with:

- indigenous: *onion, mustard, seed, crumpet, leaf, root, jam, mint, pickle*
- fully naturalized: c*umin, coconut, curry, lime, ginger, coriander, tamarind, tomato, cardamom, saffron, rice*
- recent (within my lifetime): *salsa, bhat, basmati, confit, brinjal*
- novel*: sabayon, sas, thakali, rasam, nimbu*

A restaurant like Terre a Terre is at the leading edge of both culinary and linguistic multiculturalism. All sorts of other areas have their borrowings too: wherever we share artefacts or ideas or practices with another culture, we import associated vocabulary, for example in music (*bhangra, didgeridoo*), clothes (*pashmina, lederhosen),* or religion *(stupa, muezzin)*. The question "but is this word English" feels narrow-minded and unhelpful. To give a number to the words of English, we would need to be narrow-minded and unhelpful.

3.2 Naturalization

A side-effect of importing words is: how much naturalization do we do?

There are assorted reasons – some good, some bad, most contentious – for having immigration policies and controlling which people are allowed into a country, and those policies are then strenuously policed. For words, some countries (famously France, with its

Académie Française) have, or have had, policies, and we may argue about the reasoning behind those policies being good or bad. They are also hard to police. English does not have such a tradition. We welcome in all sorts of words – but are often not sure how to say or spell them. The Nepalese staple lentil soup, in enTenTen12, is found as *dal baht, dal bat, dahl bat, dahl baht, dahl baat.* If the source language did not use the Latin alphabet, then the imports will also suffer vagaries depending on the source-language writing system and transliteration schemes. The *dal baht* case suggests a problematic mapping for the /a:/ sound between Nepali (usually written in Devanagari script) and English (written in Latin). Arabic usually does not write vowels, which is a main reason why there are so many options for how *Mohammed* is spelt in English. *Mohammed, Mohammad, Mohamed, Mahmoud, Muhammed, Mehmet, Mahmud, Mahmood, Mohamad, Mahomet,* and *Mehmood* all occur more than 1000 times in the enTenTen12 corpus.

English can be seen as an imperialist language, currently the world's pre-eminent imperialist language, with its words marching into other cultures and taking over. English speakers, at least so far as their language is concerned, have no anxieties about being taken over and fading out. But the situation looks quite different from the other side. All over the world languages are threatened and are dying, usually where, more and more often, bilingual speakers choose the alternative over their indigenous language (Crystal 2000). One part of this process is at the level of vocabulary, with speakers, even when speaking the indigenous language, using imports more and more often, either in preference to a local term, or because there is no well-established local term. Many languages have government-supported terminology committees, charged with identifying, or creating, local language terms where as yet there is nothing well-established in the local language. Most often the non-local term is an English one.

The question, "do we include this word in the count for *our* language?" is an interesting one for English – but for many languages it is also a political one, closely related to the very survival of the language.

3.3 Variants

Most English words have a single standard spelling and if the word is spelt in any other way, it is a spelling error. We all learnt that at school. We are troubled by the few exceptions: does *judg(e)ment* have an *e* in the middle? Answer: it can. There are also the transatlantic exceptions, including the *or/our* group (*colo(u)r, favo(u)r, hono(u)r,* etc.) and the *ise/ize* group. In our count of the words of the language, do we treat variants as different words?

Many languages have far less stabilized spelling than English, in particular languages which do not have a long written tradition.

There is interplay between standardization, pronunciation, and dialects. How far can a word stray and still be the same word? When I first came across *eejit*, when working with Glaswegians, I was puzzled as I felt I did not know the word. It was some months before I discovered it was a variant of *idiot.*

Some alphabetical languages, like English, have partial correspondences between spelling and pronunciation. Other, like Italian, have close to full correspondence. In those, if the pronunciation varies from region to region, the spelt form often will too, so leaving us less certain about two items being variants of the same word. The *eejit* cases will abound.

4. Conclusion

"How many words are there?" begs a set of further questions about what a word is: across time, across languages, across variation in meaning and spelling and spaces-between-words, across morphological structure. There is also the question of whether we are talking about the core of the language, or about the whole language including all the specialist corners where some small group of people have developed their own terms and usages.

Dictionaries are no help. They have pragmatic solutions to the question that they face, namely, "how many words shall we include", and the answer varies from dictionary to dictionary. Whatever they say on the back cover is to be treated with the greatest scepticism.

"How many words are there?" is not a good question. A better question is "how many words do various different speakers of a language (of various levels of education, etc.) typically know", or, moving on from a purely academic perspective to one where the answer has practical implications, "how many words do you need?" For that, we pass you on to the chapter by Paul Nation (this volume).

References

Connelly, N. G. (2005). *Nomenclature of Inorganic Chemistry: IUPAC recommendations.* Royal Society of Chemistry (Great Britain). London.

Crystal, David. (2000). *Language Death*. Cambridge: Cambridge University Press. ISBN 0-521-65321-5.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013). The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster.