

# The Science of Words

Miller

Chapter 7



◀ Familiar written words are perceived as units, not as strings of letters. Jasper Johns plays on this automatic and immediate recognition in his 1959 painting, False Start.

▼ A quarter-century earlier, experiments had found this reflex word recognition to be so powerful that people asked to name the colors in which a series of words were printed faltered when, say, the word in green ink was R-E-D.

<b>RED</b>	<b>RED</b>	<b>BLUE</b>
<b>BLUE</b>	<b>BLUE</b>	<b>GREEN</b>
<b>GREEN</b>	<b>GREEN</b>	<b>BLACK</b>
<b>BLACK</b>	<b>BLACK</b>	<b>YELLOW</b>
<b>YELLOW</b>	<b>BLUE</b>	

CHAPTER



## The Mental Lexicon

*W*illiam James (1842–1910), the father of scientific psychology in America, told a story about a practical joker who, “seeing a discharged veteran carrying home his dinner, suddenly called out ‘Attention!’ whereupon the man instantly brought his hands down, and lost his mutton and potatoes in the gutter.” People who know a word are like well-drilled veterans. They may not lose their dinner in the gutter when they hear it, but they cannot help but respond.

The reflex recognition of words is a topic of much importance to scientists who study the mental lexicon, but a caveat is needed before that story is told. In general, it is easier to explore people's knowledge of words using written rather than spoken materials, simply because inscriptions are easier than sounds for an investigator to control. This is one reason that so much experimental information about the mental lexicon is available only for the written word—and for alphabetically written English words, at that. This limitation is ethnocentric and generally deplorable, but until cross-cultural replications are available there is little that can be done but to report results for this special case. A thoughtful reader will be cautious in drawing generalizations.

### *The Word-Superiority Effect*

More than a hundred years ago James McKeen Cattell (1860–1944), another pioneer American psychologist, reported an unexpected finding: Letters are easier to read when they form a word than when they do not. Cattell compared haphazard strings of letters with short words by measuring the shortest exposure time that was needed for correct recognition. He found that at short exposure durations, where only four or five random letters can be recognized, it is possible to read two or three short words that together contain more than five letters. For example, the nine letters,

FONHGTAEW

are much harder to read when shown in that arrangement than when presented as:

FOG HAT NEW

Words are seen as individual units, not as strings of letters.

This phenomenon waited more than fifty years for a plausible explanation. It came in the form of probability theory. Students of the statistical properties of written messages observed that written words are highly redundant (see Chapter 2). That is to say, a string of nine letters conveys more selective information when any letter can occur in any position than when the same nine letters are constrained to spell familiar words. If one assumes that selective visual information is received at the same rate for both displays, words should be recognizable faster (after less information has been assimilated) than should nonredundant strings of letters. In other words, Cattell's subjects had a much better chance of guessing the letters correctly when they saw them in words than when they were random strings. Someone who saw only

FO★ H★T ★★W

had a much better chance of filling in the missing letters when they spelled words than when they did not.

That explanation was accepted for another twenty years until the experimental psychologist Gerald M. Reicher figured out how to test it. The trick is to eliminate the effects of guessing. People try to read a short string of letters that is flashed briefly, then answer a question about the final letter in the string. For example, they might see

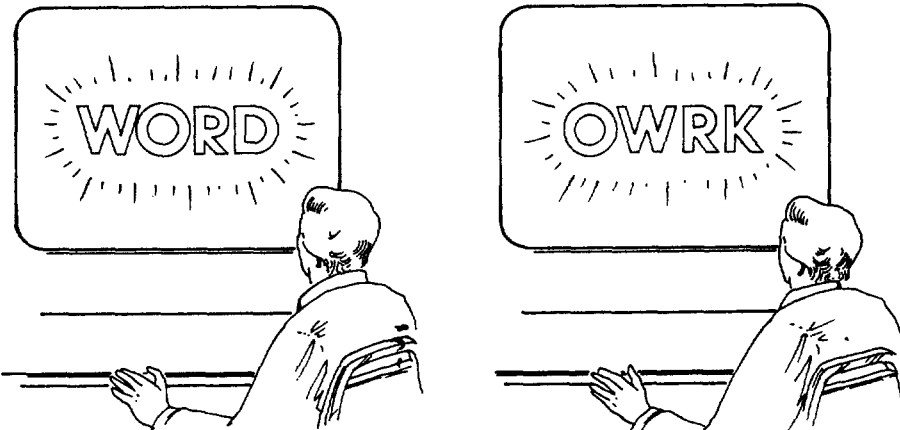
HEAR

very briefly and then be asked whether the final letter was a D or an R. This question cannot be answered by guessing the final letter that forms a word, because both alternatives form a word. For the purpose of comparison, other people see

AEHR

and are asked the same question: Was the final letter D or R? Obviously, this question cannot be answered by guessing the final letter that forms a word, because neither letter forms a word. If Cattell's phenomenon were simply a matter of guessing from the context, then under these conditions—where context provides no help for either condition—the difference should disappear.

In fact, however, the phenomenon does not disappear. Accuracy in reporting the final letter is significantly better in words than in nonwords. In fact, the final letter is reported more accurately in a word than when it is presented in isolation. This finding, known to psycholinguists as “the word-superiority effect,” put a sharp point on Cattell's observation.



*A string of letters is flashed on the screen for a tenth of a second and viewers are asked whether the final letter was a D or a K. Their responses are significantly more accurate when the letters spell a familiar word than when they do not.*

After much experience with printed words, people come to see them as unitary wholes—not as strings of letters, but as integrated chunks of information whose constituent parts are not identified separately. The precise nature of these learned patterns is not yet established, but the important implication of this explanation is that literate adults have acquired, one way or another, a large store of complex visual units that are immediately available in reading.

Evidence that these unitary percepts are involuntary, as well as immediate, comes from another observation, called the Stroop effect after its discoverer, J. Ridley Stroop. For his doctoral dissertation in 1935, Stroop printed color names in different colored inks: The word *red* might be printed in blue, the word *yellow* in green, *brown* in red, and so on through a long list of colored color words. Then he asked people either to read aloud the list of words or to name the sequence of colors. He found that people could read words printed in colored inks almost as rapidly as they could read them in black ink, but they had great difficulty naming the colors of the inks. When they looked at the letters R-E-D printed in green ink, they could not avoid reading *red*, which interfered with saying “green.” Reading the words was so automatic that they were unable to suppress this reaction and concentrate on the task at hand.

Familiar words are coherent perceptual units, so immediately and automatically available that a literate person is no longer able to control their recognition. The totality of these acquired perceptual units has been likened to a dictionary that people carry around in their heads.

## ***The Familiarity Effect***

The ability to recognize words as coherent units is acquired through learning, and like most learned abilities, it improves with practice. The more times a word is encountered, the more familiar it becomes and the faster it can be recognized.

This relation holds even for meaningless words. To demonstrate the familiarity effect, try the following experiment. Take ten 7-letter words in, say, Turkish and make up a deck of 86 cards in such a way that two words are printed on 25 cards each, two more on 10 cards each, two on 5 cards each, two on 2 cards each, and the final two on only 1 card. Then shuffle the deck, hand it to a friend who knows nothing about Turkish, and ask him or her to go through it one card at a time, spelling each word aloud and then pronouncing it. After the entire deck has been read in that manner, announce a surprise test: Measure the shortest exposure duration required to recognize each of the ten words. If you do the experiment correctly, you will find that the more frequently a word was seen the more rapidly it could be recognized. It will take about three or four times as long to recognize words seen only once as to recognize words seen twenty-five times.

It is well known that everyday language provides optimal conditions for the development of large differences in the familiarity of different words. One

<b>BLUE</b>	<b>BLUE</b>	<b>BLUE</b>
<b>GREEN</b>	<b>GREEN</b>	<b>GREEN</b>
<b>BLACK</b>	<b>BLACK</b>	<b>BLACK</b>
<b>YELLOW</b>	<b>YELLOW</b>	<b>YELLOW</b>
<b>BLUE</b>	<b>BLUE</b>	<b>BLUE</b>
<b>BLACK</b>	<b>BLACK</b>	<b>BLACK</b>
<b>RED</b>	<b>RED</b>	<b>RED</b>
<b>GREEN</b>	<b>GREEN</b>	<b>GREEN</b>

A. Control

B. Matched

C. Mismatched

*The Stroop effect. People are first shown column A and asked to read the words aloud as fast as they can; that calibrates their reading speed. Then they are shown column B and asked to name aloud the colors of the words as fast as they can; their speed for naming the colors in B is the same as their reading speed for A. Finally, they are shown column C and again asked to name the colors of the words as fast as they can. At this task, people go much more slowly and make many mistakes, frequently reading the word instead of naming its color.*

of the most firmly established statistical facts about words is that some of them are used far more than others. For example, Hartvig Dahl has counted the frequency of different words (word types) in a transcript of 1,058,888 running words (word tokens) of spoken conversation. Of course, his definition of "word" was crude (any string of letters between successive spaces; see definition D1 in Chapter 2), because that is the easiest unit for a computer to count. Thus, for example, the *uh* that fills pauses was counted as a word, and *be*, *am*, *are*, *is*, *was*, and *were* all were counted as different words, not as different forms of the same word, *be*. The results were so massive, however, that no refinements in the definition of "word" would have changed them significantly. Dahl found that the most frequently spoken word was the first person singular pronoun; on the average, every sixteenth word was *I*. The top twenty words are listed in the table on the next page—taken together, those twenty made up more than 37 percent of all the words uttered. Note, incidentally, that only one of the frequent words, *know*, is an open-class, or content, word; all the others are closed-class words, little words that give grammatical shape to phrases and sentences. As the list continues beyond the twentieth word, more content words begin to appear, but slowly at first. Just 42 different word types made up 50 percent of the word tokens counted; 848 different word types made up 90 percent of the corpus.

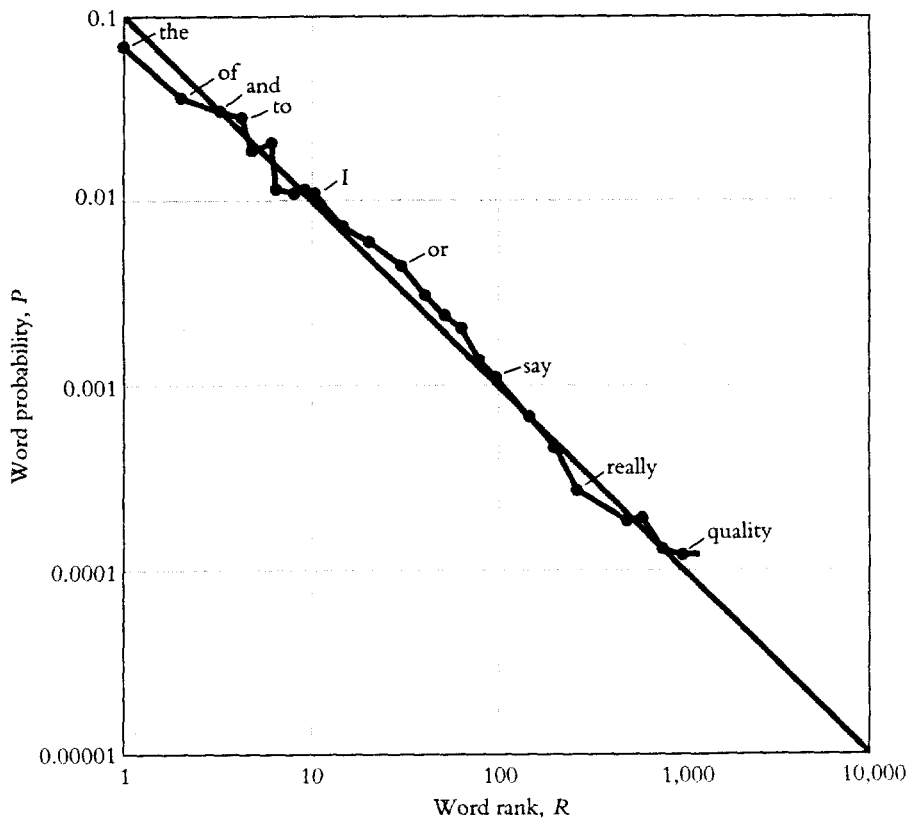
***The Twenty English Words Occurring Most Frequently  
in Personal Discourse***

Rank	Word type	Frequency	Cumulative frequency	Percentage
1	<i>I</i>	65,213	65,213	6.2
2	<i>and</i>	38,020	103,233	9.7
3	<i>the</i>	29,753	132,986	12.6
4	<i>to</i>	29,653	162,639	15.4
5	<i>that</i>	27,558	190,197	18.0
6	<i>you</i>	26,598	216,795	20.5
7	<i>it</i>	20,542	237,337	22.4
8	<i>of</i>	20,290	257,627	24.3
9	<i>a</i>	19,385	277,012	26.2
10	<i>know</i>	15,285	292,297	27.6
11	<i>was</i>	15,091	307,388	29.0
12	<i>uh</i>	14,017	321,405	30.4
13	<i>in</i>	12,964	334,369	31.6
14	<i>but</i>	9,799	344,168	32.5
15	<i>is</i>	8,875	353,043	33.3
16	<i>this</i>	8,815	361,858	34.2
17	<i>me</i>	8,506	370,364	35.0
18	<i>about</i>	8,377	378,741	35.8
19	<i>just</i>	8,318	387,059	36.6
20	<i>don't</i>	8,307	395,366	37.3

From H. Dahl, *Word Frequencies of Spoken American English*. Essex, Conn.: Verbatim, 1979.

Similar data for written texts show greater variety in the choice of words. For example, whereas Dahl found only 17,871 different word types in his transcript of 1,058,888 spoken words, Henry Kučera and W. Nelson Francis at Brown University counted 50,406 different word types in their sample of 1,014,232 written words. But the general picture is the same. A few words are overworked, most are neglected. From such statistics it is inevitable that some words will become much more familiar than others.

It is a general observation that the more familiar a word is, the less time people require to read it. Faster recognition is a consequence of highly familiar words being seen as perceptual wholes, not as strings of letters. For example, if people are handed a printed text and asked to scan it for all occurrences of the letter *t*, they are more likely to overlook a *t* in *the* than in other less frequently used words. That is to say, people see the highly familiar *the* as a complete unit, not as a string of three letters. Moreover, it is not the meaning of the text that makes people overlook *t* in *the*, because the same thing happens when they are asked to scan a haphazard list of words. The familiarity effect is strong and



The standard curve for English words in written texts. The probability  $P$  of a word occurring is plotted as a function of the word's rank  $R$  when ordered with respect to frequency of occurrence. The product  $PR$  is approximately constant, which yields the straight line with a slope of  $-1$  when plotted on doubly logarithmic coordinates.

pervasive; psycholinguists systematically design their experiments in such a way that the familiarity effect does not swamp other variables in which they are interested.

An obvious implication of the familiarity effect is that the dictionary in your head must be very different from the dictionaries that are sold in bookstores. How long it takes you to find something on a printed list depends on the length of the list, but how long it takes you to recognize a word does not seem to depend on how many different words you know. If you look up such infrequently used words as *tun* or *ire* in a hand-held dictionary, it does not take any longer than it takes to look up the frequently used words *the* or *but*. But when you look up these words in your mental lexicon, the less used words take much longer to find.

The comparison is questionable, of course, because it takes so long to find anything in a hand-held dictionary, but it does pose a question about how a mental lexicon might be organized. Are familiar words somehow imprinted on the brain in larger letters? Perhaps as words are used they are returned to the



top of a pile, so that frequently used words are always near the top. Perhaps frequently used words are easy to find quickly because they are stored in many different places in the brain.

It is no great trick to demonstrate that the mental lexicon is not organized the way a hand-held dictionary is. What is not so easy to figure out is how the mental lexicon IS organized. It is not even obvious how many mental lexicons there are.

### *Multiple Vocabularies*

The word *lexicon* has two senses. One is synonymous with dictionary: a printed book containing an alphabetized list of words and their meanings. The other is more abstract: the words of a language, whether or not they have been written down. An unabridged, printed dictionary can be regarded as a rather tedious theory—or a very detailed description—of the abstract lexicon. That, at least, is what a good dictionary aspires to be. But it is not a satisfactory description of a mental lexicon.

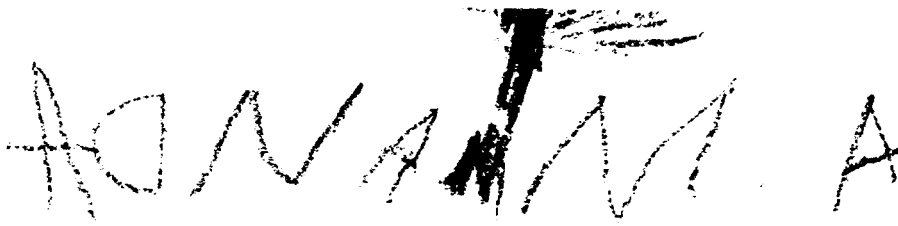
The term *mental lexicon* introduces still a third, more personal, sense. What you know, your personal word knowledge, is but a subset of the abstract lexicon, the lexical component of the language. The abstract lexicon can be thought of as the sum total of all the different words in all the mental lexicons of all the people who know and use the language. Nobody knows every word, but somebody knows each one.

What does it mean to say that someone knows a word? Does it mean that they use it in speaking? In writing? Does it mean that they can define it? Or does it mean merely that they are sure they have seen it before? There are many words that a person can recognize in reading and might even use in writing, but would never utter or expect to hear in ordinary conversation. In a printed dictionary, a word is either on the list or it is not; in a mental lexicon, the edges are fuzzy.

One way to describe these differences in how words are known is in terms of multiple vocabularies. A literate person has at least two vocabularies, a phonetic vocabulary for talking and listening and an orthographic vocabulary for reading and writing; an illiterate person, in contrast, has only the phonetic vocabulary.

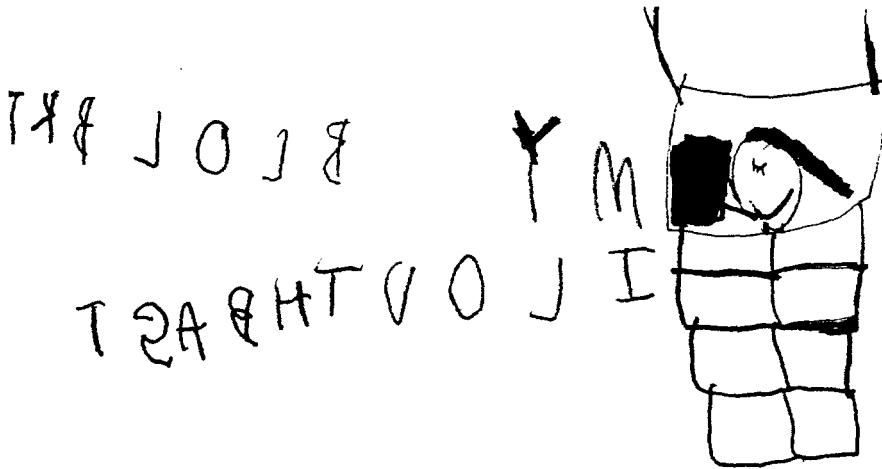
Once this notion of multiple vocabularies is introduced, it is natural to ask how many there are. In addition to the distinction between spoken and written, is there not also a difference between input and output? Combining the two distinctions gives four vocabularies: a phonetic input vocabulary for listening, a phonetic output vocabulary for speaking, an orthographic input vocabulary for reading, and an orthographic output vocabulary for writing.

These are not trivial or unimportant distinctions. Neuropsychologists, who study patients with brain injuries that interfere with speech or language in various ways, claim that they need at least those four vocabularies to describe



A handwritten name 'AMANDA' written in mirror image, appearing as 'ADNANA'.

Presumed signs of dyslexia such as mirror writing are often seen in the early stages of learning to write. Here Amanda, in preschool, mirror-writes her name and about "my blanket I love the best."



Handwritten text and a drawing of a blanket. The text includes mirrored words like 'TRF J O J R' and 'T Q A R H T V O J I'. The drawing shows a blanket with a face and a black square, possibly representing a button or a patch.

the clinical symptoms that they see. A type of disorder known as *dyslexia* can serve to illustrate how independent the different vocabularies are. Dyslexia denotes a reading difficulty; when it results from brain injury it is called acquired dyslexia to distinguish it from the apparently innate reading difficulties of certain children. Loss of the ability to read is known as *alexia*. Moreover, since many patients with acquired dyslexia also show *agraphia* (inability to spell or write), the more interesting cases for the present discussion are those designated as having alexia without agraphia. These patients can hold a conversation and they can write, but they read only with great difficulty.

Each clinical case has its own unique features that make generalization difficult, but studies of alexic patients have shown that their reading of letters is usually better than their reading of words. When asked to read a word, they may try to build it up from its letters in a slow and labored way. Shown the printed word *ball*, they might say aloud "B, A, L, L, . . . ball." How successful they are depends on how good they are at spelling. When shown handwritten words they have even more difficulty, because individual letters are harder to isolate and identify in cursive script than in printed form. Finally; when groups of letters are presented briefly, alexic patients have no greater success in reading words than in reading haphazard strings of letters or digits—no word-

superiority effect is obtained. It might be possible to explain these symptoms as consequences of difficulty in recognizing letters, but a more plausible theory is that there is a particular area in the brain where visual word forms are stored and recognized. When that area is damaged, the patient tries to compensate for the loss with letter-by-letter spelling.

The converse of patients showing dyslexia without agraphia are those showing agraphia without dyslexia—patients showing impairments of the writing process without serious difficulties in speaking, listening, or reading. Lexical agraphia is probably the simplest form. In a language whose written form is regular, these patients may not be seriously handicapped, but in English, where a variety of spellings sometimes correspond to the same spoken



*The Berlin Wall in its last days (December 1989) suggests the polyglot multiples that contemporary history can impose on the mental lexicon.*

utterance, their difficulties are very noticeable. Patients afflicted in this way are not simply poor spellers; they can spell regular words and even nonwords perfectly well by relying on a nonlexical phonological route. It is only irregular spellings that give them trouble. Similar evidence argues for a distinction between input and output processes for spoken language; clinical neurology seems to provide evidence for at least four different vocabularies.

Four vocabularies may seem like a lot, but why stop there? Why not go on? Why not include tactile input and output vocabularies for those who read and write braille? Or telegraphic input and output vocabularies for those who send and receive Morse code? And that is only for one language. Someone who knows two languages could double the number, and someone who knows three could triple it. There is almost no limit to the number of vocabularies a determined polyglot might accumulate.

At this point a thoughtful reader will become uncomfortable with this way of describing the situation. A vocabulary is a large store of information: It contains tens of thousands of words, most of them with multiple meanings. Building just one vocabulary is a major learning task. Is it credible that some people would acquire dozens of these elaborate knowledge structures? And, if so, is there a separate vocabulary matrix for each one? How would polyglots make room in their heads for anything else?

Obviously, all these different vocabularies cannot be totally independent. Consider an analogy. Everyone knows that different signals can carry the same message. An acoustic signal corresponding to the spoken word *hello* is picked up by a microphone and transduced into an electronic signal; the signals are different, but the message remains invariant. A handwritten note is typed into a computer and transmitted to a remote computer screen; several different signals convey the same message. Are the different vocabularies that have been distinguished by neuropsychologists like that? Can they be regarded as little more than different collections of signals for transmitting the same messages?

A test of this analogy would be whether messages remain invariant under transformation from one kind of signal to another. In some cases, such invariance must obtain. A word as spoken and as heard cannot be associated with different meanings, if only because speakers hear their own speech: It would be totally confusing if, when you uttered *table*, you heard yourself saying something else. Or if, when you wrote *table*, you saw something different on the page. Input and output vocabularies must be closely related. Moreover, in languages that are written alphabetically, the spoken/heard "table" is related to the written/read *table* by well-learned rules of spelling. Literacy would be even harder to acquire than it is if *table* could be spelled by some arbitrary string of letters. Even people who know a second language do not have totally independent vocabularies: For someone who knows both English and Italian, *tavola* and *table* will not be drastically different in meaning. And for a familiar English word like *table*, it seems safe to assume that the same set of meanings is associated with the spoken, heard, written, and read representations. The real ques-

## Lexical Access and Positron Emission Tomography

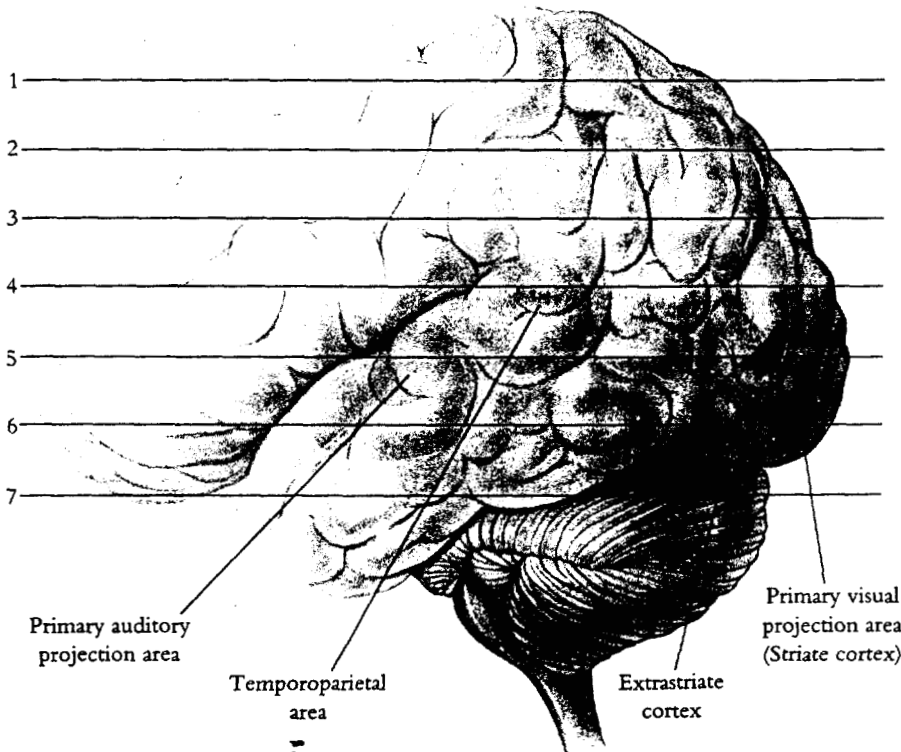
Computerized tomography is the construction of a three-dimensional image of a bodily structure from a series of X-ray pictures. Positron emission tomography (PET) adds a measure of blood flow to this imaging technique. If a subject—patient, volunteer, animal—is given an intravenous injection containing a radioactive isotope, the resulting

radiation can be recorded tomographically. As metabolism at a site increases, blood flow increases; as more blood flows to it, radiation from that site increases; as the radiation increases, it is registered on the tomographic image.

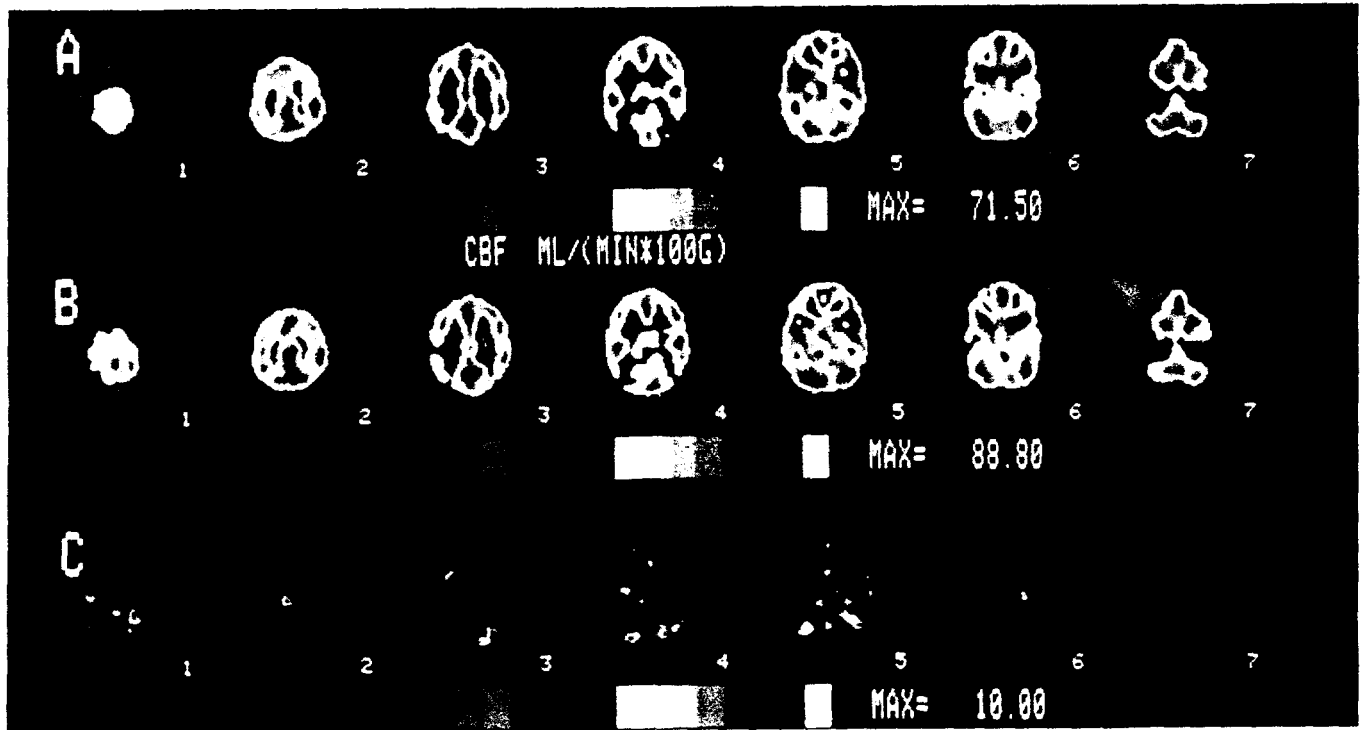
This technology has been used to study blood flow in the brain during simple verbal tasks. Volunteers received intravenous injec-

tions of water that contained oxygen-15 (half-life 122 seconds). Then PET scans were taken for 40 seconds while the subjects stared at a fixation point (the control condition) or while they passively listened to or looked at a series of familiar English nouns (the experimental condition). The effect of this passive perceptual processing on blood flow was estimated by subtracting the scans for the control condition from the scans for the experimental condition.

Listening to words increased the blood flow in the primary auditory projection areas in both hemispheres and in the nearby temporoparietal areas in the left hemisphere. Looking at words increased activity in the primary visual projection area (the striate cortex) and in the nearby extrastriate cortex in both hemispheres, although more intensely in the left hemisphere. When subjects were asked to look at pairs of words and press a key if they rhymed, increased activity was observed in both the extrastriate and temporoparietal areas. These observations were consistent with the belief that the left temporoparietal area is where auditory word-images are formed and the extrastriate area is where visual word-images are formed.



*The left hemisphere of the human brain, locating the slices made by the PET scans pictured.*



A subtraction procedure makes it possible, using PET scans, to identify brain areas related to lexical processing. Here, for example, the top row (A) shows the brain blood flow measured while the person viewed a fixation point (the control state). Each image is a slice through the brain, going from the top of the brain (slice 1) to the bottom (slice 7). The top of each slice is the anterior part of the brain and the bottom is the posterior. The middle row (B) shows the blood flow while the person looked at words that were presented at a rate of one per second (the experimental state). The bottom row (C) is obtained by subtracting the control images from the experimental images; the difference shows the change in blood flow induced by visual word presentation. It can be seen from slices 4 and 5 in row C that the peak response occurred in the posterior part of the brain, the visual input center.

tion is not how many vocabularies there are, but how so many different signals can all gain access to the same message.

In short, to speak loosely of multiple vocabularies can be misleading. Dyslexics do not lose the words they are unable to read; agraphics still know the words they cannot spell. Such patients are simply unable to gain access to what they know via the usual associations. What psycholinguists have in mind is a single lexical matrix with multiple ways of getting in and out of it.

## ***Vocabulary Size***

A lexical matrix is too large to imagine building totally new ones for each use. It is a sobering thought to realize how much lexical knowledge you have acquired. Some of it you know firmly, but a lot is known only at the level of recognition—and often held so tentatively that it might better be called lexical belief, rather than lexical knowledge. But it is obvious that you know a great deal. It is a challenging problem to estimate how much.

The standard procedure for estimating an individual's vocabulary size is to administer a multiple-choice test. Words are presented and the test-taker is asked to choose correct definitions from lists of four or five alternatives. Since the person being tested merely has to recognize the right defining phrase, the results of the test might be called the size of the person's reading vocabulary. The problem is to develop a test in such a way that the test score can be translated into an estimate of vocabulary size. Dictionary sampling is the popular method for achieving that result. The basic assumption (and the source of most of the disagreements among estimators) is that the number of words in the language is given by the number of words in a dictionary. For this basic assumption to be even marginally plausible, it is necessary to use the largest dictionary available.

Consider the following arithmetic. Suppose you start with a dictionary that contains 500,000 words. If you sample 500 of them at random to estimate the size of your friend's mental lexicon, then your sampling factor is 1,000. That is to say, for every word that your friend recognizes, you give credit for knowing 1,000 words that you might have sampled but did not. If your friend recognizes 100 of the 500 words, the estimated vocabulary size is  $100 \times 1,000$ , or 100,000 words. But note, however, that if you had started with a dictionary containing only 100,000 words, your friend would have to recognize every test word to achieve the same estimated size. The general rule is: The larger the dictionary on which your test is based, the larger the estimates that you are likely to obtain.

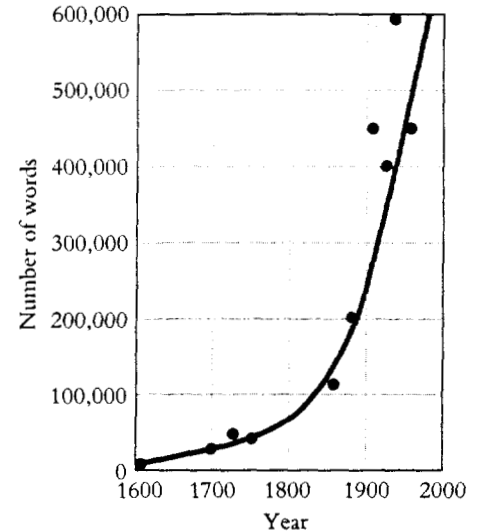
Since the size of the dictionary that you sample is so important, you might ask what the largest English dictionary is. The answer depends on when you ask—over the past four hundred years there has been a steady increase in the number of words that dictionaries contain.

Author/editor	Brief title	Date	Approximate number of words
Robert Cawdrey	<i>Table Alphabeticall</i>	1604	2,500
John Kersey	<i>New English Dictionary</i>	1702	28,000
Nathan Bailey	<i>Dictionarium Britannicum</i>	1730	48,000
Samuel Johnson	<i>Dictionary</i>	1755	40,000
Noah Webster	<i>American Dictionary</i>	1828	70,000
Noah Porter	<i>Dictionary of English, unabridged</i>	1864	114,000
William D. Whitney	<i>Century Dictionary</i>	1891	200,000
Isaac K. Funk	<i>New Standard Dictionary</i>	1913	450,000
James A. H. Murray	<i>Oxford English Dictionary</i>	1928	400,000
William A. Neilson	<i>Webster's New International</i>	1934	600,000
Philip B. Gove	<i>Webster's Third New International</i>	1961	450,000

If you extrapolate this growth, it rapidly approaches infinity, taking the sampling factor with it. Fortunately, the number seems to have leveled off in the twentieth century at around half a million words. But your friend may still know some perfectly acceptable words that are not on a list of 500,000.

Another hazard for such estimates is that, even after you have chosen the largest dictionary you can find, you still have to estimate how many words it contains. You might think that you could rely on the publisher's claims. For example, the dust jacket of one best-selling collegiate dictionary says that it has "almost 160,000 entries and 200,000 definitions." But remember, this is advertising, and most customers think that more is better. So take a look inside. This particular dictionary has 1,373 pages, which should work out to  $160,000/1,373 = 115$  entries per page. If you sample a few pages at random, however, you will find only 50 to 60 headwords per page. (A headword, sometimes called the main entry, is the uninflected form, or citation form, that identifies the entry and is used to place it in alphabetical order with other entries.) Where are the rest of the entries? They are there, but you have to read the headword entries to find them because the "160,000 entries" are not all headwords. In this case, there are about 71,000 headwords. To get the count up to 160,000 you have to count as a word everything that is printed in bold-face. For example, inside the entry alphabetized under the headword **obfuscate** are the inflected forms **obfuscating** and **obfuscated** and also the run-on (appended) derivatives **obfuscation** and **obfuscatory**.

This format may be a good way to publish and advertise a dictionary, but think what it does to the person who wants to sample its entries in order to



The size of large English dictionaries has grown exponentially over the past four hundred years, indicating how questionable it is to assume that the number of words in the English language is given by the number of words in any particular English dictionary.



construct a vocabulary test. If the dictionary is assumed to contain 160,000 words, then the test will be counting all five forms of *obfuscate* as separate words. Most test makers have recognized that the dictionary's operational definition of "word" was not what they had in mind when they set out to estimate how many words people know.

In short, what seem on the surface to be straightforward questions—How many words are there? How many words does the average person know?—turn out, on closer inspection, to be rather complicated. And the ultimate complication—that some words have many different meanings—has not yet been mentioned. It is a curious fact that the most familiar words tend to have the most meanings; people perversely persist in using most frequently those polysemous words most likely to be ambiguous. Should polysemy be taken into account in estimating vocabulary size? The word *press*, for example, has dozens of meanings, both as a noun and as a verb. Does it count no more than, say, *press agent*, which has only one meaning?

Faced by such problems, vocabulary estimators have been forced to make some arbitrary decisions. Suppose the question is slightly rephrased: not "How many words does the average person know?" but rather, "How many words has the average person learned?" That is to say, some expert looks at every root word and asks, "If you learned this word, what other related words would you probably understand?" The answer would define as "one word" an entire family of words that, according to the judge's lexical intuition, people come to understand when they master the central word of the family. For example, *obfuscate* would count as one word, not five.

What kinds of decisions would such a judge need to make? Some of them are easy. A person who speaks English should not have to learn inflected forms as if they were totally new words; if you know *book*, for example, you should also know *books* and *book's*. But what about derivatives and compounds? Most derivatives formed with such regular affixes as *#ness* or *#ly* can be understood in context if the stem is understood, but derivatives formed with morpheme boundaries should be considered one by one. For example, consider some of the problems with the *#er* suffix. A person who has learned *run* will understand *runner*, but a person who has learned *walk* could easily miss one sense of *walker*, and a person who has learned *tell* will not understand *teller* at all. So knowing *run* and *runner* would count as knowing one word, knowing *walk* and *walker* could count as either one or two, and knowing *tell* and *teller* should count as two. By such a criterion, most compounds count as new words—it is a characteristic feature of compound words that their compound meaning is not given by the meanings of their parts, although sometimes it is possible to guess what the coiner had in mind.

At the University of Illinois, William Nagy and Richard Anderson went through a list of 227,553 different words using this learning criterion. According to their count, the list contained 45,453 headwords. They judged that 139,020 of the remaining 182,100 derivative and compound forms could be understood in context by someone who knew their root forms, but that 42,080

## Notable Lexicographers of English: Samuel Johnson

It is sometimes assumed that a good dictionary of a language should contain all the words in that language. No English dictionary meets that requirement—it is not even obvious that the number of English words is finite. And it was certainly not the intention of the great lexicographers of English to create an archive for every word that has been or could be uttered. Their goals were ambitious, but not THAT ambitious.

The most famous English lexicographer was Dr. Samuel Johnson (1709–1784), the poet, essayist, literary critic, and conversationalist. In 1746 Johnson, always in need of money, signed a contract for the *Dictionary of the English Language*. The following year he published his *Plan of a Dictionary of the English Language*, which was addressed to Lord Chesterfield in hope of enlisting support for the project. It was a well-reasoned plan, showing familiarity with the best lexicographic practices of the day. He discussed criteria for including words and set his policies for dealing with spelling, pronunciation (by the use of rhyming words), morphology, contexts, and idiomatic expressions. And he planned to hire experts to include encyclopedic material in some entries.



Samuel Johnson.

Johnson hoped that his dictionary would serve to “fix” the English language in a pure state—a project that he knew Chesterfield favored. But to no avail. Chesterfield soon lost interest. Lacking financial support, many of Johnson’s plans proved too ambitious for one man to implement.

The *Dictionary* was published in 1755, after nine years of prodigious effort. The work is so important in the history of English lexicography that it is often cited as the first to have had this or that feature. But Johnson was not an innovator. There is no ingredient

of his *Dictionary* that had not been introduced already by other lexicographers. Johnson simply did the standard things better than they had ever been done before. His spellings were traditional, his treatment of pronunciation was as sketchy as that of other dictionaries, and his etymologies were uncertain, but his definitions were lucid gems, illustrated by a choice of literary quotations drawn from his own vast scholarship. By illustrating every sense with quotations from great authors, he hoped to preserve “the wells of English undefiled” to serve as a permanent standard of good writing.

Today, Johnson’s *Dictionary* is remembered mostly for a few highly quotable definitions. His famous definition of *lexicographer* as a harmless drudge was certainly modest. And his definition of *oats* as a grain that in England is generally given to horses, but in Scotland supports the people, was, he later confessed, meant to vex the Scots. At the time Johnson’s *Dictionary* appeared, however, it was unequaled, and for more than a century it remained the most authoritative dictionary in English. But even Dr. Johnson’s great prestige was not enough to halt the irresistible process of linguistic change.

were semantically opaque. To master the complete list of 227,553 different words, then, a student would have to learn  $45,453 + 42,080 = 88,533$  word families. The Illinois team went on to ask how many of these 88,533 lexical elements most people know. They estimated that the average high school graduate knows about 45,000 of them.

The Illinois estimate is conservative. It excludes proper names, numbers, foreign words, acronyms, and many undecomposable compound words that occur regularly in newspapers. If these were included, the average high school graduate would probably be found to have learned some 60,000 different "words." Superior students, because they do more reading, would probably know twice that many.

It is not worth arguing over these numbers, however, because so many subjective and intangible factors contributed to the estimates. Shouldn't there be some more objective way to decide which words are learned together and can be regarded as a family?

### *Retrieval from the Mental Lexicon*

If you are a good reader, as your eyes skim along the lines of print, you set in motion a sequence of complex interpretive processes whose outcome is the conscious appreciation of meaning. Fortunately for you, but unfortunately for linguistic scientists, the information processing required to produce that awareness does not clutter your mind or obscure the meaning. The process is simply unavailable to introspection. To build a picture of what is going on behind the scenes, it is necessary to make inferences on the basis of the performance itself or to conduct psychological experiments designed to choose among different hypotheses.

Anyone who considers such matters in detail, however, quickly realizes that recognizing the words is a critical component of the reading process. It is easy to see how that part of the process could be studied experimentally: Simply flash the words and see how long an exposure is required to read them. When Cattell tried it he found that words can be read much faster than nonwords, and his discovery was generalized to the principle that the more familiar a word is, the less time people need to recognize it.

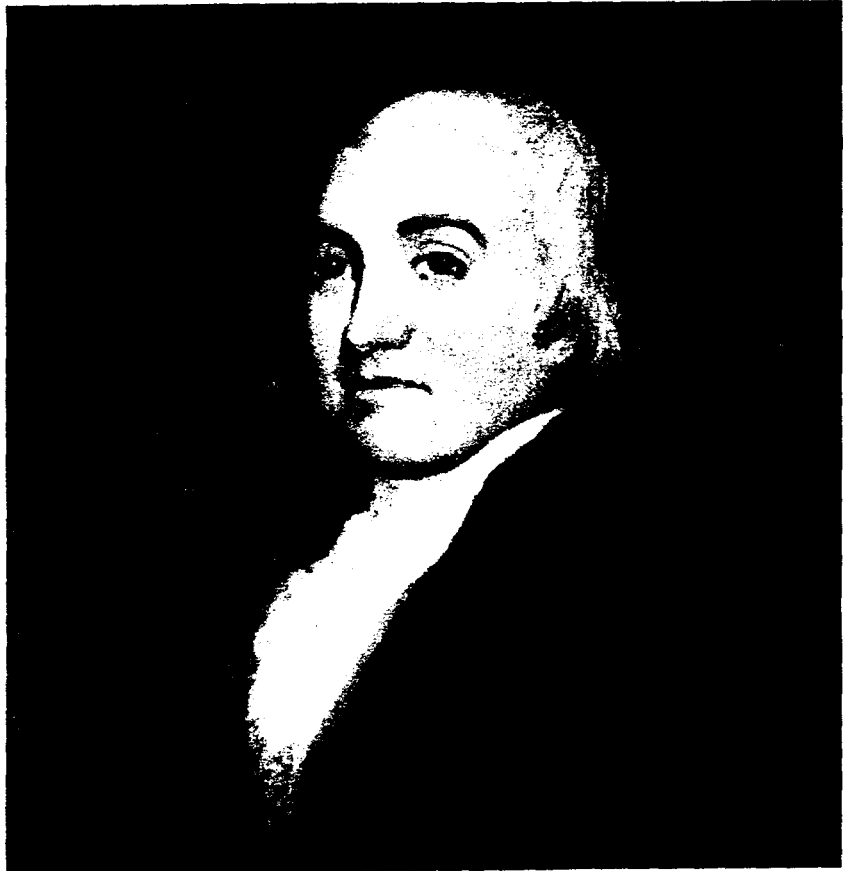
A variety of techniques have been used to demonstrate that familiarity breeds speed, but perhaps the most popular is the lexical decision task. People are asked to indicate as rapidly as possible whether or not a string of letters spells an English word. Subjects do not have to say what word it is or what it means—just "Yes, it is a word," or "No, it is not a word." The reaction time is the time between the instant that the word appears and the instant that people answer yes or no. (Reaction times for mistaken responses are discarded.)

## Notable Lexicographers of English: Webster and Worcester

Whereas Samuel Johnson hoped to “fix” the language, the goal for Noah Webster (1758–1843) was to replace Johnson’s *Dictionary* as the American standard. Webster first became famous for his blue-backed speller, *The American Spelling Book* (1783), which sold more copies than any other schoolbook that had ever been published.

Webster had no respect for Johnson. He criticized Johnson’s choice of words, simplified his spellings, decided that literary citations are unnecessary (he made up his own examples), and vigorously pressed his claim that American English needed its own dictionary. Webster’s attempt to fill that need, his two-volume *American Dictionary of the English Language*, was published in 1828. Historians have judged it a minor contribution to lexicography that would have disappeared had it not been actively promoted and heavily revised by its publishers.

Two years later, when the more conservative lexicographer Joseph Emerson Worcester (1784–1865) published his *Comprehensive Pronouncing and Explanatory Dictionary of the English Language*, a battle began. Webster and Worcester were natural antagonists. The brash Webster, contemptuous of tradition and proudly American,



Noah Webster.

was associated with Yale; the scholarly Worcester, who admired British lexicography, was associated with Harvard. But the real “war of the dictionaries” was the commercial rivalry between their publishers. Webster’s publishers eventually won by commissioning

a German philologist to rewrite Webster’s etymologies in light of the recent growth of linguistic knowledge in Europe. In 1864 their new dictionary appeared and rapidly gained international fame—ironically beating Worcester at his own conservative game.

The lexical decision task consistently shows faster response times for high-frequency, high-familiarity words—as expected. An early finding that was not expected, however, was that people respond faster to homographs than to nonhomographs. That is to say, words like *crane* or *chest* that have more than one sense were recognized as words slightly faster than equally familiar words like *neighbor* or *cliff* that have only one sense. Indeed, the more meanings a word has, the faster it is recognized as a word. Even though subjects are not asked to identify the word or think of its meaning(s), they obviously cannot prevent themselves from doing so, since the variety of meanings influences the test results. The natural inference is that a homograph is really two or more entries in the mental lexicon and that the response time is the time it takes to find any one of them.

To illustrate how this task can be used to probe into the workings of the mental lexicon, consider an experiment in which native speakers and readers of Serbo-Croatian (the principal language of Yugoslavia) made rapid lexical decisions about inflected singular nouns in three cases: nominative, dative/locative, and instrumental. Serbo-Croatian has a complex case system, in that there is no simple relation between the form of the affix and the case that it marks. Some cases, moreover, are used more frequently than others. For example, the feminine noun *frula* (flute) occurs 31 percent of the time in the nominative case (written *frula*), 10 percent of the time in the dative/locative case (both are written *fruli*), and less than 1 percent of the time in the instrumental case (written *frulom*). So it is possible to ask the following question: Does the familiarity effect for Serbo-Croatian nouns depend on the stem frequency or on the frequency of the inflected form? It was found that the nominative form could be recognized as a word slightly (but significantly) faster than could the dative/locative or the instrumental forms, but there was no difference in response times between the dative/locative and the instrumental.

What could this mean? Consider one possible hypothesis, derived from the morpheme-based theory of morphology described in Chapter 6. Suppose that the mental lexicon contains only a list of morphemes and that words containing two or more morphemes cannot be looked up, but must be synthesized on the fly, so to speak. Then the nominative singular noun *frula* must be synthesized out of the root *frul* and the suffix *-a*. Since *frul* is shared by all cases, the activation threshold for *frul* cannot explain why the nominative singular is recognized faster than are the other forms. So the difference must be attributable to the suffix *-a*. But that is improbable, because *-a* has that effect only when it is used to mark the nominative case. So the morpheme-based hypothesis can be dismissed. The mental lexicon must contain more than a list of morphemes.

A subtler method for exploring the role of morphology in lexical organization involves a variation of the lexical decision task known as repetition priming. If a word or nonword is presented twice (with an intervening lag), the second lexical decision time will be faster than the first. It is assumed that the first presentation (the prime) facilitates the decision on the second presenta-

## Notable Lexicographers of English: James Murray

James Murray's goal was to establish the histories of English words by arranging literary quotations in chronological order. In 1857 Richard Trent presented a proposal for such a dictionary to the Philological Society, which decided to sponsor *A New English Dictionary on Historical Principles*. Temporary editors began the task, and a network of volunteer readers was assembled to contribute quotations. But the real work did not begin until 1879, when Murray (1837–1915) was persuaded to become the editor.

He worked diligently, and by 1884 the first volume had been published. To speed the work, three more editors were eventually added, but the final volume did not appear until 1928. By that time what had come to be called the *Oxford English Dictionary* contained 240,000 headwords and 400,000 entries, filled 15,487 large pages, and was based on a file of more than 5,000,000 quotations. Not only were sense divisions precise and detailed, with clear definitions, but the history of every sense was documented with quotations. The etymologies were the best that existed up to that time. The wonder is not that it took fifty years to complete, but that it was ever completed at all.



James A. H. Murray.

Although he edited the largest English dictionary, it is clear that Murray had no ambition to include all the words of English. Vulgar words were excluded, and the growing vocabularies of science, technology, commerce, and indus-

try were largely omitted—all in keeping with the nineteenth century's conception of good taste. But the goal that Murray and his companions set for themselves—the creation of a valid historical record—was achieved in a manner that evoked such adjectives as “monumental,” “massive,” “indispensable,” and “without parallel.” Perhaps the magnitude of the task was best described by Murray himself in his presidential address to the Philological Society:

Only those who have made the experiment know the bewilderment with which editor or sub-editor, after he has apportioned the quotations for such a word as *above* . . . among 20, 30 or 40 groups, and furnished each of these with a provisional definition, spreads them out on a table or on the floor where he can obtain a general survey of the whole, and spends hour after hour in shifting them about like pieces on a chess-board, striving to find in the fragmentary evidence of an incomplete historical record, such a sequence of meanings as may form a logical chain of development. . . . Those who think that such work can be hurried, or that anything can accelerate it, except more brain power brought to bear on it, had better try.

tion (the target), and the size of the difference is taken as the priming effect. For example, when the singular dative/locative form of the feminine Serbo-Croatian noun *rupi* (hole) was repeated, the second lexical decision time was 90 milliseconds shorter than the first. When the prime was changed to the nominative form *rupa*, the priming effect on *rupi* was 79 milliseconds. And when the instrumental *rupom* was used as a prime, the priming effect was 69 milliseconds. Regular inflected forms of the same word do prime each other, indicating that there is a close association among them.

The results in English are even stronger: Inflected words prime their uninflected forms just as well as the uninflected forms prime themselves. This result provides objective support for the intuitive impression that someone who has learned, say, *pour* or *burn* does not have to learn *pours* or *burned* as separate words. The vocabulary estimators are right in counting all the inflected forms as a single entry in the mental lexicon.

But what about derivative words? The vocabulary estimators seem to have been on the right track there, too. For example, the inflected form *manages* and the derivative forms *manager* and *management* all facilitate a subsequent recognition of *manage* as much as *manage* facilitates itself. By contrast, repetition priming does not occur between morphologically unrelated words whose initial letters coincide; for example, *cancel* does not prime *can*.

In general, therefore, the results of experiments using repetition priming with a lexical decision task support the general idea that morphologically related words are stored together in the mental lexicon. Activate any member of a morphological family and all the others are ready to spring into action. Moreover, these effects are not limited to reading printed words—the same kinds of results have been obtained with auditory priming, although the temporal duration of the priming effect seems to be shorter. Experts still argue over details, but the general conclusion has been that the organization of the mental lexicon reflects the way different morphological forms are learned together.

Those who want to estimate vocabulary size in terms of the number of root words that must be learned in order to understand all the different but morphologically related words can take comfort from this picture. But they should not overlook the fact that lumping all morphologically related forms together as a single word in a single lexical entry leaves the psycholinguist with a very unappealing characterization of the entries in the mental lexicon. What use is a lexical entry that fails to differentiate inflected and derived forms? It cannot be assigned to any single syntactic category. It cannot be used in the statement of morphological or syntactic rules. It cannot be associated with any single definition. And how differences in the familiarity of the different forms are to be registered is left a mystery.

In the end, therefore, a theorist is driven back to the conception of the mental lexicon as a lengthy list of individual words, not a collection of undifferentiated word families. But on top of this lengthy list there must be an elaborate network of morphological associations among words. When a word

is used—activated—the activation spreads over this network of morphological associations. Words are not only associated with meanings. They are associated with one another.



It is a general observation that the human brain seems to have more storage capacity than computing power, so the idea of storing separately every form of every word may not be too outrageous. But it is a puzzle to understand why the brain stops where it does. When people encounter a new word they list it in their mental lexicons and associate it with its morphological relatives, but when a new, regular syntactic phrase is heard, it is not listed in memory. Presumably there comes a point when even the vast storage capacity of the human brain can no longer cope with the exponential principle.





▲ In James Murray's Scriptorium, built in his back garden, the first edition of the Oxford English Dictionary (OED) took shape. More than half of the 44-year project, which drew on a file documenting word usage in over five million quotations, was his own work.

▶ This entry (for abaptistan, an obsolete instrument for cranial surgery) in Murray's hand was for the first installment (A-ANT) of the OED, published in 1884.

+ Abaptistan *Ab.* [ Gr. ἀβάπτιστον, not im-  
 14 mersed. of ἀψω + βάπτω: dip ] 'the crown of the old  
 15 capon, which was conical, or had some contrivance to prevent it  
 16 from penetrating the cranium too suddenly, and so injuring the brain'  
 17 (described by Galen) Spal. Ins. des. (So called by Galen, and invented  
 18 early, Sextimianus  
 19 in Chamber's Spal. Ins. des. as the Gr. name, whence in subseq. dict., but having  
 20 no claim to the name)  
 + 1696 Philips, Abaptistan = Anabaptistan, a Surgeon's Instrument.