The Mental Lexicon

Ask anyone what the basic building blocks of a language might be - Words
This much is obvious, but the rest is a complete mystery.

Prior to Chomsky's generative theory of syntax (1964), linguists gave priority to the description of words.

Field workers had a concern in collecting words for undocumented languages.
Historical linguists inferred genetic relations between languages by comparing words.
Psychologists assumed children acquired language by learning words

Chomsky's theories brought phonology and syntax to the fore, stressing the structural relations between phonological and syntactic categories.

Linguistics is now in the midst of a rediscovery of the word; started by Joan Bresnan's work on a Lexical-Functional Grammar that claimed to be a 'realistic theory of grammar'. This position has been adopted by Chomsky as a cornerstone of his 'minimalist' program.

This theory puts the word at the heart of a linguistic description by recognizing the role words play in unifying information from different parts of the grammar.

Syntax Phonology [$_P$ ___ NP] [In]
Word *in*
| |
Semantics Location wrt concavity

Acquire a language by acquiring words & their associated grammatical features.

We'll investigate these relations later on,
For now focus on general properties of our mental lexicons

Size

Psychologists have long known vocabulary size was correlated with measures of intelligence.
Most intelligence tests rely heavily on measures of vocabulary size.
More recently Sue Kemper et al. have found a correlation between vocabulary size and risk for Alzheimers amongst the elderly

Let's start with a simple question and ask how many words do we know?
5,000 10,000 50,000 100,000 500,000 1,000,000?

Make a guess. How do we test?

One solution starts by checking how many words you know in a dictionary.

Start by checking how many of the words on a page you recognize

No. of words = proportion of recognized words x no. of words in dictionary

(Seashore & Eckerson. 1940. The measurement of individual differences in general English vocabularies. J. of Educational Psychology 31.14-38)

Now we can start asking about the technical details

1. How does the size of our dictionary affect our estimate?

The size of English dictionaries has increased exponentially over the last 400 years

Robert Cawdrey	Table Alphabeticall	1604	2,500 words
John Kersey	New English Dictionary	1702	28,000
Samuel Johnson	Dictionary	1755	40,000
Noah Webster	American Dictionary	1828	70,000
Noah Porter	Dictionary of English	1864	114,000
Isaac K. Funk	New Standard Dictionary	1913	450,000

(Miller, p. 135)

2. What counts as a word? What does your dictionary count as a word?

Divide no. of words by no. of pages = average no. of words per page

How does this estimate correspond to what you find on the page?

Dictionary counts all boldface entries

forget, *forgot*, *forgotten*, *forgetting*, *forgettable*, *forgetter* & *forget oneself*

forgetful, *forgetfully* & *forgetfulness*

1. Should we count all of these? 2. Why are they arranged in two separate entries?

Start with the first question. What counts as a word? What is a word?

Free form / free morpheme

a. occur in isolation

b. separable from other words

Are all of the entries under *forget* free forms? Are they separate, but equal?

They are inflected forms of the word *forget*. Demonstrate various types of affixation:

inflection	forg**o**t, forgott**en**, forgett**ing**
derivation	forgett**able**, forgett**er**, forget**ful**, forget**fulness**
compounding	forget oneself

Reason to think these forms are stored differently in our minds

inflection does not change meaning significantly-predictable

derivation sometimes produces unpredictable semantic changes

compare *forgetter*; *walker*-2 entries; *teller* (bank)

exceptions **a* $forget_N$ (c.f. *a* $report_N$); **unforget*; **reforget*

compounding usually results in unpredictable semantic change
forget oneself: lose self-restraint

K'iche' Maya compounds

saqa b'ala:m	raxa:l q'ana:l
white jaguar = ocelot	greeness yellowness = glory
b'aqwach tap	rax te:w
eye crab = hangnail	green cold = malaria

Inflected forms are predictable from their part of speech and meaning
In other words, they can be generated on the fly by the grammar

Many derived forms and most compounds have unpredictable meanings
They must be stored in the mental lexicon.

Another problem words frequently have more than one meaning, e.g., *net*, *press*, *break*
Should we count these different uses as different words?

There are many other processes we use to make up new words:
Conversion, e.g., target (V) from target (N)
Acronyms, e.g., radar (from radio detecting and ranging)
Abbreviations, e.g., tv, id, vd, oj, aids
Blends, e.g., spam (spiced/ham), smog (smoke/fog), spork (spoon/fork)
Clippings, e.g., Alex (from Alexander), doc (from doctor), rent (from parent)
Backformations, e.g., enthuse (from enthusiasm), donate (from donation), pea from pease
Coinage, e.g., xerox, kleenex
Onomatopoeia, e.g., meow, cheep, ribbit

3. Make a distinction between words we use and words we recognize
= production vs. comprehension lexicons

For all of these reasons, estimates of average vocabulary size vary considerably

Seashore & Eckerson:	58,000 basic words
	1,700 rare basic words
	96,000 derivatives and compounds
	> 150,000 total
Nagy & Anderson	45,453 basic words
	42,080 semantically opaque derivatives & compounds

estimate average high school graduate knows 45,000

Diller 1978 estimated ~ 250,000 total words for college students
Language Files (tenth edition, p. 381): 40,000-60,000

These estimates are 100 to 1,000 times greater than the most optimistic counts of animal signs. Underline the quantitative difference between human language & forms of animal communication

---

Speed

Once we have an estimate of vocabulary size we can begin to estimate the speed of lexical look up

Imagine going through an unorganized list of words to see if it contains the word *boat*
You find the information faster in a dictionary, but this still takes time

Normally speak at a rate of 6 σs/sec ~ 3-4 words
Native speakers recognize a word in 200 msecs (1/5 of a sec) from the beginning of the word
Often well before all of the word is heard

A speech shadowing task is a traditional technique to test access speeds
Subjects repeat what they hear in headphones
Good shadowers can repeat with a delay of 250-275 msec
Subtract 50-75 msec for time to convert word to speech
-> 200 msec for word recognition (= 1/5 sec.)

Ask yourself whether *plid* is a word
Took my 90 MHz laptop about 2 secs to respond Its lexicon is much smaller than mine

Assuming that you have a basic vocabulary of 60,000 words
and search through them at the rate of 100 words/sec
It would take 10 minutes to search through your entire lexicon

The size of the average lexicon and the speed of lexical access
(I'll add the ease of lexical acquisition, too)
point to a systematic organization for the lexicon

There are tradeoffs between the size of the lexicon and access
Imagine trying to cram books into a room. Cram the maximum no. by simply stacking the room full. That technique won't improve access speed, though.

Accuracy

We occasionally make mistakes in the retrieval process

Mistake nonwords for real words: *concision*

Confuse *reluctant* (unwilling) and *reticent* (unwilling to speak)
deprecate (disapprove) and *depreciate* (lower in value)
foreboding (ominous) and *forbidding* (dangerous)
effect (cause) and *affect* (influence)

Speech errors or slips of the tongue (or pen) show that lexical access isn't a simple mechanical search.

Most confusions due to similarities in pronounciation and/or meaning.

More frequently used words are accessed faster and more accurately

We are also constantly updating our lexical store to reflect the lexical environment
- What do you put groceries into at Dillons (a bag or a sack?)
- No computer has the power to invent new words: *spork* (spoon + fork)

We recognize the gaps in our dictionaries
- starve s.o. out by denying them food; what about denying s.o. water?

Intelligently organized, dynamic store of lexical information
available to every user of a human language.

Production

What do we know about the word production process?

Evidence from slips of the tongue:

MEANING	The *white* (=black) sheep of the family.
	They've *ended* (=started) the third week of their strike.
SOUND	A *reciprocal* (= rhetorical) question.
	The *audience* (ordinance) survey map.
MEANING AND SOUND	You're a *destructive* (= disruptive) influence.
	Look at this *badger* (= beaver).

The Stepping-Stone Model
Assume the parts of words are activated in sequence:

	Meaning		Sound	
---->	BEAVER	---------->	bivər	------->

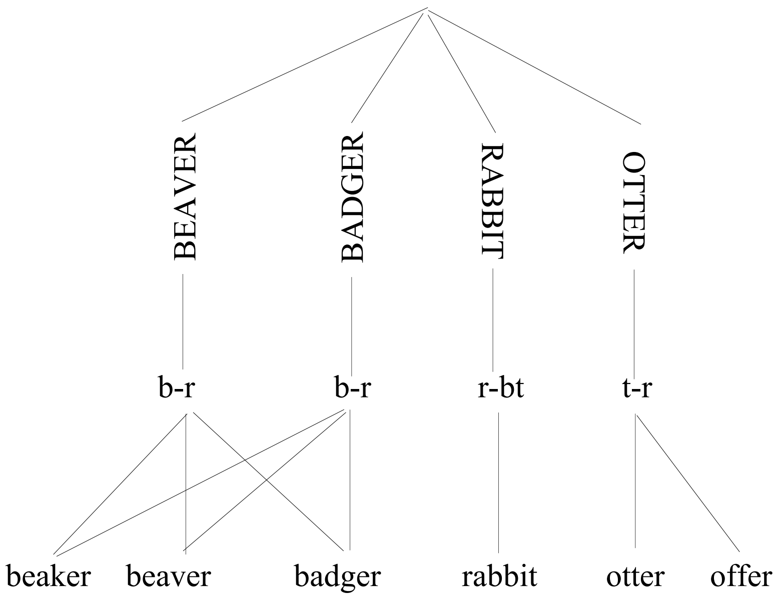
Assume multiple candidates are activated at each stage:

OTTER	beaker
BEAVER	beaver
BADGER	badger
RABBIT	bearer
	begger

The model predicts these errors will appear at each stage.
The model does NOT predict an interaction between meaning and sound
Most errors feature a similarity in meaning and sound.

The Waterfall Model

The waterfall or 'cascade' model (McClelland 1979) makes all of the information from the semantic stage available to the phonological stage. Once a set of meanings has been activated, the information cascades down to the activation of sounds.



Problem: the waterfall model doesn't allow information to flow backwards. It's common to prompt people to recall a word by giving them an initial sound, e.g., think of a small woodland animal whose name begins with a *b*. The waterfall model shows how meanings activate sounds, but not how sounds activate meanings.

Neural Networks

The key to capturing lexical activation is allowing activation to spread in multiple directions: from meaning to sound and from sound to meaning. The progressive activation of possible candidates and the suppression of unlikely candidates continues until one word reaches a threshold. Frequently used words reach this threshold faster than infrequently used words.

Comprehension

The Cohort Model provides one representation for the process of lexical recognition. It assumes lexical recognition proceeds in a left to right fashion sifting through all of the words which share the same initial sounds:

BEE
BEAVER -----------> BEAVER
BEETLE
BEGIN
bI vr

The Cohort Model is supported by word recognition studies that selectively erase sounds in the word. Erasing the initial sounds has a greater effect than erasing other sounds.

Psychologists have done most of the research on lexical access. Unfortunately, psychologists do not investigate lexical access in other languages. English has many suffixes and few prefixes. The root or base of English words almost always comes at the beginning of the word. Consider applying the Cohort Model to Navajo.

Navajo Verb (Young & Morgan 1987)

Disjunct				Conjunct						
1	2	3		4	5	6	7	8	9	
ADV	ITER	DIST_PL	#	DIR_OBJ	DEIC_SUBJ	ADV	MODE	SUBJ	CLASS	STEM
a	ná	da		shi	ji	di	i	sh	0	
ba				ni	’a	hi	yi	ni	ł	
ch’í				yi	hwi	li	ni	0	d	
cha				bi		si	si	iid	l	
kéé’				ha		yi	o	oh		
k’í				’a						
na				nihi						
so				di						
tá										
ta’										
yá										

ADVERBIAL	Manner, direction and indirect object
ITERATIVE	‘over and over’ or ‘back again’
DISTRIBUTIVE PLURAL	‘each one separately’
DIRECT OBJECT	Number and person
DEICTIC SUBJECT	Indefinite (‘someone’ or ‘people in general’)
ADVERBIAL	Adverbial/aspectual
MODE	Aspect (perfective, imperfective, progressive, optative)
SUBJECT	Person and number
CLASSIFIER	Voice and Transitivity

Dictionaries

Navajo speakers are frustrated by dictionaries that list words by stem without the inflectional prefixes. This frustration suggests that a speaker’s recognition lexicon may not correspond directly to the semantic lexicon. Documenting a spoken language requires sampling all of the words a speaker accepts as part of the language. What lexical domains need to be documented?

Lexical Categories in Kaufman and Justeson’s Mayan Etymological Dictionary

Kinship and social organization
Thought and feeling/perception and evaluation
Color

Body arts and other parts; bodily processes
Animal parts
Location (place names)
Earth
Sky
Fire
Water and Liquides
Mammals
Birds
Reptiles
Turtles
Snakes
Lizards
Batrachians
Fish
Mollusks
Bugs
Speech and Interaction
Play, Dance, Music
Trade and Property
Manipulation
Tools
Buildings
Furniture
Containers
Clothing and Adornment
Agriculture
Plant Parts
Maize
Beans
Trees
Palms
Vines
Herbs
Grasses
"Pineapples"
Fungus
Eating
Taste
Life and Existence
Movement
Sickness
Magic
Quality
Shape
Quantity

Manner
Speed
Numerals
Names
Sounds (onomatopoeia)
Exclamations

References

Aitchison, J. 1987. Words in the Mind. Oxford: Basil Blackwell.
Diller, K. C. 1978. The Language Teaching Controversy. Rowley, Mass.: Newbury House.
Kaufman, T. & Justeson, J. A Preliminary Mayan Etymological Dictionary. Unpublished, The University of Pittsburgh.
Miller, G. A. 1991. The Science of Words. New York: W. H. Freeman.
McClelland, J. L. 1979. On the time relations of mental processes: an examination of systems of processes in cascade. Psychological Review 86.287-330.
Nagy, W. E. & Anderson, R. 1984. The number of words in printed school English. Reading Research Quarterly 19.304-330.
Seashore, R. H. & Eckerson, L. D. 1940. The measurement of individual differences in general English vocabularies. Journal of Educational Psychology 31.14-38.
Young, R. & Morgan, W. 1987. The Navajo Language: A Grammar and Colloquial Dictionary, second edition. Albuquerque: University of New Mexico Press.