CHAPTER 6

Chunking in Language Usage, Learning and Change: I Don't Know

Nick C. Ellis

6.1 Chunking: The Foundations

6.1.1 Letter Chunks

How many letters can you apprehend from a single presentation? Look at the first of the four stimuli below, just one quick glimpse, and immediately afterwards, write down what you saw. Now do the same for the next three, one at a time.

CVGJCDHM
RPITCQET
UMATSORE
VERNALIT

I suspect that you perceived more letters from later strings than from earlier ones. But given that each stimulus was the same eight letters long, why should that be?

Miller et al. (1954) showed Harvard undergraduates pseudoword letter strings like those above for very brief presentations (a tenth of a second) using a tachistoscope. The average number of letters correctly reported for the four types of stimuli were, in order, 53, 69, 77 and 87 percent. The pseudowords differed in their 'order of approximation to English' (AtoE). CVGJCDHM exemplifies zero-order AtoE strings – they are made up of letters of English, but these are sampled with equal probability of occurrence (1 in 26). RPITCQET exemplifies first-order AtoE strings – made up of letters of English, but sampled according to their frequency of occurrence in the written language (as in opening a book at random, sticking a pin in the page, and choosing the pinned letter [e.g. 'r']; repeat). UMATSORE exemplifies second-order AtoE – these reflect the

probabilities of two-letter sequences in English (bigrams) (open a book at random, stick a pin in the page, choose the pinned letter and the one following it in a word [e.g. 'um'], find another random example of the last letter, choose that letter and the one following it [e.g. 'ma']; repeat; . . .). VERNALIT exemplifies fourth-order AtoE – these reflect the probabilities of four-letter sequences in English (4-grams) (open a book at random, stick a pin in the page, choose that letter and the three following it in a word [vern], find another random example of the final trigram, choose that trigram and the one following it in a word [e.g. 'erna']; repeat; . . .).

The fact that we can better perceive higher-order AtoE strings means that our perceptual system is sensitive to the probabilities of occurrence of letters and letter sequences in English. It has expectations about 4-grams and better perceives stimuli which meet those expectations. At a lower level, it has expectations about 2-grams and better perceives stimuli which meet these those expectations. Generally, it has expectations about the relative frequencies of occurrence of English orthographic sequences and better perceives stimuli which meet those expectations. Miller, Bruner and Postman summarized such results, showing that performance can be predicted from knowledge of the statistical structure of English, saying 'the more frequently a trace has been embedded in a trace aggregate, to use the language of Gestalt psychology, the greater the probability that the aggregate will be aroused when the component is activated'.

Two years later, George Miller (1956) introduced 'chunk' as his preferred technical term to replace 'trace aggregate'. In his classic paper, 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', he reviewed the span of short-term memory (STM) – how long a sequence (of digits, letters or words) you can repeat back in order, having just heard it – and observed that for young adults it was 7±2 chunks of information. He noted that memory span is approximately the same for stimuli with very different amounts of information (binary digits have 1 bit of information each; decimal digits have 3.32 bits each; words about 10 bits each). So memory span is not limited in terms of bits of information but rather in terms of chunks, these being the largest units in the presented material that a person recognizes.

What counts as a chunk depends on the knowledge of the person being tested. For example, a word is a single chunk for a speaker of the language, but is many chunks for someone who is totally unfamiliar with the language: compare *breakfast* and *parakuihi*, or *sushi* and 寿司. Chunks are subjective: 'We are dealing here with a process of organizing or grouping the input into familiar units or chunks, and a great deal of learning has

gone into the formation of these familiar units' (Miller 1956: 91; see also my Chapter 4 on salience in this volume). The units of perception are influenced by prior association; in William James's words, 'The chief cerebral conditions of perception are the paths of association irradiating from the sense-impression, which may have been already formed' (James 1890b: 82).

6.1.2 Word Chunks

How many words can you apprehend from a single presentation? Look at the first of the four stimuli below, just one quick read through, and immediately afterward, write down what you saw. Now do the same for the next three, one at a time.

inducted avidity slaughtered renewed dharma authentically
she that empire the line letter
any dominant intelligent species to believe
delivers to the writer a magnificent

I suspect that you recalled more words from later strings than from earlier ones. But given that each stimulus was the same six words long, again, why?

I created these stimuli using the same definitions of order of approximation to English as above, but this time using words as units rather than letters. Instead of a pin and a book, I used the Natural Language Toolkit (Bird 2006; Bird et al. 2009) and searched the Brown corpus (Francis and Kucera 1979). The exercise is inspired by Miller and Selfridge (1950), who built zero- to seventh-order of approximation to English word lists, read them to students and had the students recall as much as they could. Reading from their graph, approximate recall correctness for 10-word lists was 99 percent for seventh-order AtoE, 98 percent for fourth-order, 95 percent for second-order, 62 percent for first-order, and 50 percent for zero-order. They concluded that 'when short-range contextual dependencies are preserved in nonsense material, the nonsense is as readily recalled as is meaningful material. . . . [I]t is these familiar dependencies, rather than the meaning per se, that facilitate learning' (1950: 184).

As with letters, the fact that we can better perceive higher-order AtoE word strings means that our perceptual system is sensitive to the probabilities of occurrence of words and word sequences in English. It has expectations about 4-grams, and better perceives stimuli which meet these expectations. At a lower level, it has expectations about 2-grams and better perceives stimuli which meet these expectations. It has expectations about the relative frequencies of occurrence of English words and better perceives

higher-frequency words. We have rich knowledge of chunks of words. Generally, our perceptual system is tuned to perceive higher-frequency words and word sequences that are higher in transitional probability. We have never explicitly counted such frequencies, but our perceptual system has automatically tallied their probabilities over our history of using English. The relevant knowledge is induced from usage. It is incidentally acquired while our consciousness is focused upon communication.

Note also, from the comparison of zero-order and first-order word sequence stimuli, another important range of phenomena of usage that we will return to in more detail in Sections 6.4, 6.5 and 6.8. We preferentially process frequent words over infrequent words. Infrequent words are much longer than frequent words. Infrequent words also tend to have unusual orthographic sequences, unusual spelling-to-sound correspondences and unusual pronunciations. Consider for example, *assuage, egregious, epithalamion, inefficacious, internecine, omniscient, puerile, regicidal, synecdoche, terpsichorean*. Usage shapes words.

6.1.3 Grammar Chunks

In 1957, Miller began Project Grammarama at Harvard University in order to study the learning of chunks and rules. Miller (1958) developed a laboratory analogue of grammar learning using an artificial language (AL) consisting of a set of well-formed strings that could be generated by an underlying finite-state grammar shown in Figure 6.1.

This type of finite-state system is formally simple but psychologically complex, since the underlying grammar is not readily apparent from its surface forms. Participants were shown redundant strings of letters (e.g. SSXG, NNXSXG) generated by the underlying grammar. No mention

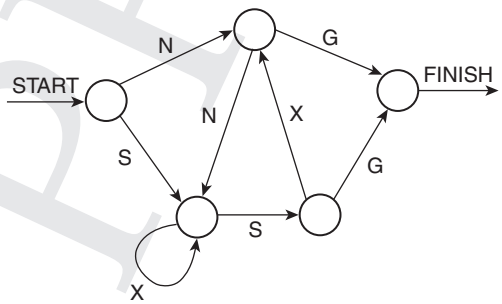

Figure 6.1 Diagram of a finite-state grammar of the type used by Miller (1958).

was made of rules or structure. They were simply asked to memorize the strings. In a comparison condition, they were asked to learn random strings which were drawn from the same letter pool but which did not follow the sequential statistics of the grammar. The results were that, although participants knew nothing of the rules of the language, the redundant, 'grammatical' strings were more easily memorized.

This study was the foundation for what is now the most widely studied paradigm of implicit learning: artificial grammar learning. The standard AL experiment involves two phases: learning and testing. In the learning phase, subjects are shown strings of letters (e.g. MXRMXT, VMTRRR) generated by an underlying grammar or rule system, usually a finite-state system that generate strings of symbols in a left-to-right, non-hierarchical fashion, often referred to as a Markov grammar. The subjects are asked to memorize the strings; no mention is made of rules or structure. After subjects have memorized the list they are informed that the strings conformed to a covert rule structure and are asked to make well-formedness (grammaticality) judgments about a set of novel strings, half of which are grammatical and half of which contain grammatical violations. The typical finding here is that subjects are able to make judgments at significantly better than chance levels without being able to articulate detailed information about what the rules governing the letter strings are, or which ones they were using in guiding their decisions. Thus it has been argued that the task demonstrates implicit learning. The paradigm has been developed and refined over the years and continues to form the basis for a considerable amount of experimental research into grammar and sequence learning (for reviews see Reber 1993; Ellis 1994b; Rebuschat 2015; Stadler and Frensch 1998; and Perruchet and Pacton 2006).

In his very fertile decade of the 1950s, Miller revealed our knowledge of chunks across the scale of language and he showed how this knowledge was used in the learning and processing of language. His work had a profound influence upon both psycholinguistics and the psychology of learning.

6.2 Learning Chunks

Chunking is the development of permanent sets of associative connections in long-term storage. Following Miller's lead, subsequent work in cognitive psychology and in Artificial Intelligence simulations of human learning and cognition in production systems such as ACT-R (Anderson 1983, 1996) and Soar (Laird et al. 1987, 1986) incorporated chunking as the primary mechanism of learning, and chunks as the units of memory.

Newell (1990) argued that chunking is the overarching principle of human cognition:

> A chunk is a unit of memory organization, formed by bringing together a set of already formed chunks in memory and welding them together into a larger unit. Chunking implies the ability to build up such structures recursively, thus leading to a hierarchical organization of memory. Chunking appears to be a ubiquitous feature of human memory. Conceivably, it could form the basis for an equally ubiquitous law of practice. (Newell 1990: 7)

From its very beginnings, psychological research has recognized three major experiential factors that affect cognition: frequency, recency and context (e.g. Anderson 2009; Ebbinghaus 1885; Bartlett [1932] 1967). Learning, memory and perception are all affected by frequency of usage: the more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed. (Hence your more fluent reading of the prior sentence than the one before). The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization, so a stimulus and its interpretation becomes associated to a context and we become more likely to perceive it in that context (hence your recognition of a colleague in their familiar stomping ground, but not unexpectedly in the street). These three aspects of usage drive chunking.

6.2.1 Frequency

The power law of learning (Anderson 1982; Ellis and Schmidt 1998; Newell 1990) describes the relationships between practice and performance in the acquisition of a wide range of cognitive skills – the greater the practice, the greater the performance (including strength of memory, likelihood of recall and fluency of production or comprehension), although effects of practice are largest at early stages of learning, thereafter diminishing and eventually reaching asymptote. This applies to recognition and recall across our world of experience: people and places, birds and bees, chalk and cheese, and, of course, linguistic constructions, too.

Consider words – though the same is true for letters, morphemes, syntactic patterns and all other types of construction. Through experience, a learner's perceptual system becomes tuned to expect constructions

according to their probability of occurrence in the input, with words like *one* or *won* occurring more frequently than words like *seventeen* or *synecdoche*. A learner's initial noticing of a new word can result in an explicit memory that binds its features into a unitary chunked representation, such as the phonological sequence 'wun' or the orthographic sequence *one*. As a result of this, a detector unit for that word is added to the learner's perceptual system, the job of which is to signal the word's presence when its features are present in the input (Morton 1969).

Every word detector has a set resting level of activation and some threshold level which, when exceeded, will cause the detector to fire. When the component features are present in the environment, this increases the activation of the detector; if this takes the level above threshold, the detector fires. With each firing of the detector, the new resting level is slightly higher than the old one: the detector is primed. This means it will need less activation from the environment in order to reach threshold and fire the next time that feature occurs. Priming events sum to lifespan-practice effects: features that occur frequently acquire chronically high resting levels. Their resting level of activity is heightened by the memory of repeated prior activations. Thus our pattern-recognition units for higher-frequency words require less evidence from the sensory data before they reach the threshold necessary for firing. So the perceptual system is tuned by experience of usage.

6.2.2 Recency

Human memory is sensitive to recency: the probability of recalling an item, like the speed of its processing or recognition, is predicted by time since last occurrence. The power function relating probability of recall (or recall latency) and recency is known as the 'forgetting curve' (Baddeley 1997; Ebbinghaus 1885). Language processing also reflects priming effects. Priming may be observed in our phonology, conceptual representations, lexical choice and syntax (McDonough and Trofimovich 2008). Syntactic priming refers to the phenomenon of using a particular syntactic structure, given prior exposure to the same structure. This behavior has been observed when speakers hear, speak, read or write sentences (Bock 1986; Pickering 2006; Pickering and Garrod 2006). (See Chapter 6 of this volume, on priming). Priming is an essential part of conversation partners' aligning and co-constructing meanings.

6.2.3 Context

Human memory is also context dependent: a stimulus (and its interpretation) becomes associated to a context, and we become more likely to perceive it in that context (Baddeley 1997; Godden and Baddeley 1975). A large body of research has shown that memory performance is reduced when an individual's environment differs from encoding to retrieval, as compared to when the two environments are the same (Tulving and Thomson 1973). The context can be environmental (places or cultures), social (speakers or cultures) or linguistic. For an example of linguistic context effects upon processing, Schooler (1993) showed that word fragment completion was faster for the second word of a strong context collocation (as in PROFOUND-IGN____?) than when the word was shown alone (IGN____?). Miller would have talked of this in terms of chunks.

Frequency, recency and context are basic forces in all contemporary psycholinguistic models of language perception and processing (Christiansen and Chater 2001; Jurafsky 2002; Traxler and Gernsbacher 2011; McClelland and Elman 1986; Xu and Tennenbaum 2007; Ellis 1996). We find structure in time (Elman 1990, 2004). Learning is statistically informed, and interpretation is probabilistic: 'Learners FIGURE language out: their task is, in essence, to learn the probability distribution P(interpretation|cue, context), the probability of an interpretation given a formal cue in a particular context, a mapping from form to meaning conditioned by context' (Ellis 2006a). Chunks are probabilistic in their nature and in their processing.

6.3 Chunking is Rational

Rational analysis (Anderson 1990) aims to answer *why* human cognition is the way it is. Its guiding principle is that the cognitive system optimizes the adaptation of the behavior of the organism to its environment in the sense that the behavior of the mechanism is as efficient as it conceivably could be, given the structure of the problem space or input-output mapping it must solve. Determining optimality for rational behavior requires a quantifiable formulation of the problem. For the case of memory, the criterial factor is the optimal estimation of an item's need probability. Rational analysis considers the way that human memory corresponds to this needs function. Anderson's (1990) rational analysis implicated three factors in determining *information need*. You met them before in the previous section, describing the forces of learning: frequency, recency and context. Consider the relative availability of items in the mental lexicon.

6.3.1 Frequency

The probability of a word occurring in a particular source is predicted by its past frequency of occurrence in that source. It works for all sorts of information. Whether organizing a library, a mental lexicon or a tool shop, you should have the most-used items nearer to hand. The power law of learning is rational in that it follows this trend. Memory performance is tuned to the world.

6.3.2 Recency

There is a power function relating the probability of a word occurring on day *n* to how long it has been since the word previously occurred. The probability of a word occurring in, say, speech to children (from the CHILDES database), or the *New York Times*, or the e-mail a person receives is predicted by its past probability of occurrence – there is a power function relating the probability of a word occurring in the headline in the *New York Times* on day *n* to how long it has been since the word previously occurred there. The forgetting curve is rational in that it follows this trend. Things happen in bursts (Barabási 2005, 2010). You can see this in your e-mail – you do not hear from someone for a while, then there is a flurry of correspondence, and then things quiet down on that front again.

6.3.3 Context

A particular word is more likely to occur when other words that have historically co-occurred with it are present. In analysis of the *New York Times* and the CHILDES databases, Schooler (1993) showed that a particular word was more likely to occur when other words that had occurred with it in the past were present. For instance, a headline one day mentioned Qaddafi and Libya, and sure enough a headline the next day that mentioned Qaddafi also mentioned Libya. I am sure you could think of parallel examples from this week's news. Schooler collected likelihood ratio measures of association between various words in order to assess the effect of this local context factor on memory and processing. As already described, in both the child language and the *New York Times* databases, a word was more likely to occur if it had occurred previously, but additionally a word was more likely to occur in a headline if a string associate of it occurred, and these effects are additive in the way predicted by Bayesian probability.

See how these three aspects of *information need* in the problem space are satisfied by their *cognitive* counterparts in learning, memory and perception summarized in Section 6.2. Chunking provides a rational representation of usage. It both builds the representations and organizes their relative availability according to need.

6.4 Psycholinguistics: Everything in Language Comes in Chunks

Ellis (2002) reviewed how the 50 years of psycholinguistic research from 1950 onward demonstrated language processing to be exquisitely sensitive to chunk frequency at all levels of language representation: phonology and phonotactics, reading, spelling, lexis, morphosyntax, formulaic language, language comprehension, grammaticality, sentence production and syntax. Usage shapes *every aspect* of language. There is space here for just a few illustrative examples.

6.4.1 Phonotactic Chunks

Frisch et al. (2001) asked native speakers to judge, using a 7-point rating scale, non-word stimuli for whether they were more or less like English words. The non-words were created with relatively high- or low-probability legal phonotactic patterns, as determined by the logarithm of the product of probabilities of the onset and rime constituents of the non-word. The mean wordlikeness judgments for these non-word stimuli had an extremely strong relationship with expected probability (r = .87).

6.4.2 Lexical Chunks

High-frequency words are named more rapidly than low-frequency ones (Balota and Chumbley 1984), they are more rapidly judged to be words in lexical decision tasks (Forster 1976) and they are spelled more accurately (Barry and Seymour 1988). When naming pictures, people are more successful and faster on items with higher-frequency names (Oldfield and Wingfield 1965).

Auditory word recognition is better for high-frequency than low-frequency words (Luce 1986; Savin 1963). There are also cohort effects in spoken word recognition. Hearing the initial phoneme of a word activates the set of all words in the lexicon that have this same phoneme. Then, as the speech signal unfolds over time and more information is received, the set is narrowed down. In the cohort model of speech recognition

(Marslen-Wilson 1990), activation in the cohort varies so that items are not simply 'in or out'. Rather, higher-frequency words get more activation from the same evidence than do low-frequency words. This underlies lexical similarity effects whereby a whole neighborhood of words is activated but the higher-frequency words get more activation and so listeners are slower to recognize low-frequency words with high-frequency neighbors because the competitors are harder to eliminate (Lively et al. 1994). Thus, the language processing system is sensitive both to the frequency of individual words and to the number of words which share the same beginnings (at any length of computation).

6.4.3 Orthographic Chunks

English spelling-to-sound mapping is a 'quasi-regular' domain. Words consisting of orthographic chunks with regular or consistent mappings to pronunciation are read better than those with irregular or inconsistent mappings. For the case of adult fluency in English, after controlling for overall word frequency, words with regular spelling-sound correspondences (like *mint*) are read with shorter naming latencies and lower error rates than words with exceptional correspondences (cf. *pint*) (Coltheart 1978); in development, exception words (*blood, bouquet*) are acquired later than are regular words (*bed, brandy*) (Coltheart and Leahy 1996).

Similarly, words which are consistent in their pronunciation, in terms of whether this agrees with those of their neighbors with similar orthographic body and phonological rime (*best* is regular and consistent in that all *-est* bodies are pronounced in the same way), are named faster than inconsistent items (*mint* is regular in terms of its grapheme-phoneme conversion (GPC) rule, but inconsistent in that it has the deviant *pint* as a neighbor) (Glushko 1979). The magnitude of the consistency effect for any word depends on the summed frequency of its friends (similar spelling pattern and similar pronunciation) in relation to that of its enemies (similar spelling pattern but dissimilar pronunciation) (Jared et al. 1990). Adult naming latency decreases monotonically with increasing consistency on this measure (Taraban and McClelland 1987). Because of the power law of learning, these effects of regularity and consistency are more evident with low-frequency words than with high-frequency ones, where performance is closer to asymptote (Seidenberg et al. 1984).

In development, Laxon et al. (1991) have shown that within regular words, consistent (*pink*, all *-ink*) and consensus (*hint*, mostly as in *mint*,

but cf. *pint*) items are acquired earlier than ambiguous ones (*cove* vs. *love, move*) and that within irregular words, those in deviant gangs where the several items sharing that rime are all pronounced in the same irregular fashion (like *look, book, cook*, etc., or *calm, balm, palm*) are acquired earlier than ambiguous ones (*love*). As with the learning of other quasi-regular language domains, these effects of consistency/ambiguity of spelling-sound correspondence within a language have been successfully simulated in connectionist models (Harm and Seidenberg 1999; Seidenberg and McClelland 1989)

6.4.4 Morphological Chunks

Morphological processing, like reading and listening, also shows effects of neighbors and false friends where regular inconsistent items (e.g. *bake-baked* is similar in rhyme to neighbors *make-made* and *take-took*, which have inconsistent past tenses) are produced more slowly than entirely regular ones (e.g. *hate-hated, bate-bated, date-dated*) (Daugherty and Seidenberg 1994; Seidenberg and Bruck 1990).

Psycholinguistic studies of the statistical patterning of chunks and their associations have been central in connectionist models of morphological processing and acquisition. These models capture the regularities that are present (1) in associating phonological form of lemma with phonological form of inflected form and (2) between referents (+past tense or +plural) and associated inflected perfect or plural forms (Cottrell and Plunkett 1994; Ellis and Schmidt 1998), simulating error patterns, profiles of acquisition, differential difficulties, false-friends effects, reaction times for production and interactions of regularity and frequency that are found in human learners (both L1 and L2), as well as acquiring default-case-allowing generalization on 'wug' tests.

Token frequency counts how often a particular form appears in the input. Type frequency, on the other hand, refers to the number of distinct lexical items that can be substituted in a given slot in a construction, whether it is a word-level construction for inflection or a syntactic construction specifying the relation among words. For example, the 'regular' English past tense *-ed* has a very high type frequency because it applies to thousands of different types of verbs, whereas the vowel change exemplified in *swam* and *rang* has much lower type frequency. The productivity of phonological, morphological and syntactic patterns is a function of type rather than token frequency (Bybee and Hopper 2001). This is because (1) the more lexical items that are heard in a certain position in a construction,

the less likely it is that the construction is associated with a particular lexical item and the more likely it is that a general category is formed over the items that occur in that position; (2) the more items the category must cover, the more general are its criterial features and the more likely it is to extend to new items; and (3) high type frequency ensures that a construction is used frequently, thus strengthening its representational schema and making it more accessible for further use with new items (Bybee and Thompson 1997). In contrast, high token frequency promotes the entrenchment or conservation of irregular forms and idioms; the irregular forms only survive because they are high frequency.

6.4.5 Collocation Chunks

Reading time is affected by collocational and sequential probabilities. Durrant and Doherty (2010) used lexical decision to assess the degree to which the first word of low-frequency (e.g. *famous saying*), middle-frequency (*recent figures*), high-frequency (*foreign debt*) and high-frequency and psychologically associated (*estate agent*) collocations primed the processing of the second word in native speakers. The highly frequent and high-frequency-associated collocations evidenced significant priming.

The British linguist Firth emphasized the importance of collocational knowledge in our understanding of word meanings: 'You shall know a word by the company it keeps' (Firth 1957). Forty years later, Landauer and Dumais (1997) presented a computational analysis of this maxim. Their Latent Semantic Analysis (LSA) model simulates a language learner's acquisition of vocabulary from text. The model simply treats words as being alike if they tend to co-occur with the same neighboring words in text passages. By inducing global knowledge indirectly from local co-occurrence data in a large body of representative text, LSA acquired knowledge about the full vocabulary of English at a rate comparable to that of school children. After the model had been trained by exposing it to text samples from more than 30,000 articles from *Groliers Academic American Encyclopedia*, it achieved a score of 64 percent on the synonym portion of the Test of English as a Foreign Language (a level expected of a good ESL learner). The performance of LSA is surprisingly good for a model which had no prior linguistic or grammatical knowledge and which could not see or hear, and thus could make no use of phonology, morphology or real-world perceptual knowledge. In this model, lexical semantic acquisition emerges from the analysis of word co-occurrence. Figure 6.2 panel 6 compares *butterfly* and *don't* for the information latent in their collocational contexts.

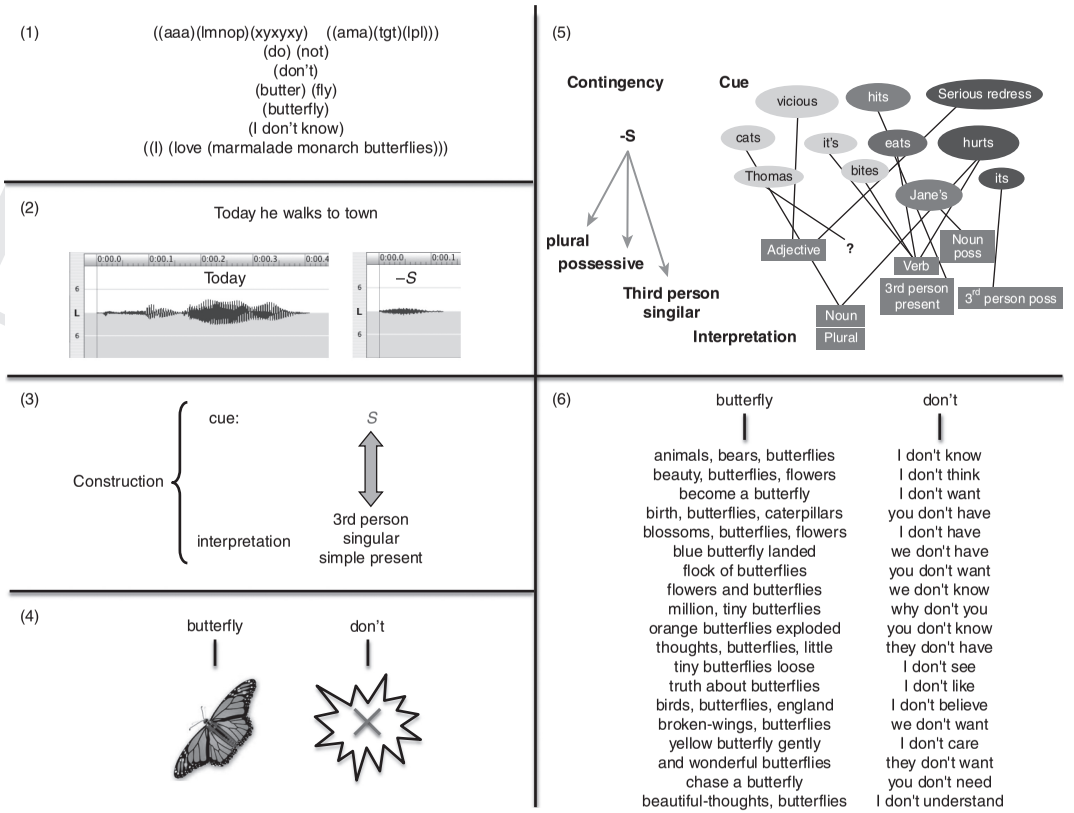

Figure 6.2 Some factors affecting language chunking. (See text for details.)

6.4.6 Phrasal Chunks

Arnon and Snider (2010) used a phrasal decision task (Is this phrase possible in English or not?) to show that comprehenders are also sensitive to the frequencies of compositional four-word phrases: more-frequent phrases (e.g. *don't have to worry*) were processed faster than less-frequent phrases (*don't have to wait*), even though these were matched for the frequency of the individual words or substrings. Tremblay, Derwing, Libben and Westbury (2011) examined the extent to which lexical bundles (LBs, defined as frequently recurring strings of words that often span traditional syntactic boundaries) are stored and processed holistically. Three self-paced reading experiments compared sentences containing LBs (e.g. *in the middle of the*) and matched control sentence fragments (*in the front of the*) such as *I sat in the middle/front of the bullet train*. LBs and sentences containing LBs were read faster than the control sentence fragments in all three experiments.

Maintenance of material in short-term memory and its accurate subsequent production is also affected by knowledge of formulaic sequences. Bannard and Matthews (2008) identified frequently occurring chunks in child-directed speech (e.g. *sit in your chair*) and matched them to infrequent sequences (e.g. *sit in your truck*). They tested young children's ability to produce these sequences in a sentence-repetition test. Three-year-olds and two-year-olds were significantly more likely to repeat frequent sequences correctly than to repeat infrequent sequences correctly. Moreover, the three-year-olds were significantly faster to repeat the first three words of an item if they formed part of a chunk (e.g. they were quicker to say *sit in your* when the following word was *chair* than when it was *truck*). Tremblay, Derwing, Libben and Westbury (2011) similarly used word and sentence recall experiments to demonstrate that more sentences containing LBs (the same ones as in their earlier-mentioned comprehension experiments) were correctly remembered by adults in short-term memory experiments.

What about L2 learners? Jiang and Nekrasova (2007) examined the representation and processing of formulaic sequences using online grammaticality judgment tasks. English as a second language speakers and native English speakers were tested with formulaic and non-formulaic phrases matched for word length and frequency (e.g. *to tell the truth* vs. *to tell the price*). Both native and non-native speakers responded to the formulaic sequences significantly faster and with fewer errors than they did to non-formulaic sequences.

Ellis and Simpson-Vlach (2009) and Ellis, Simpson-Vlach and Maynard (2008) used four experimental procedures to determine how the corpus linguistic metrics of frequency and mutual information (MI, a statistical measure of the coherence of strings) are represented implicitly in native and non-native speakers, thus to affect their accuracy and fluency of processing of the formulas of the Academic Formulas List (AFL, Simpson-Vlach and Ellis 2010). The language processing tasks in these experiments were selected to sample an ecologically valid range of language processing skills: spoken and written, production and comprehension, form-focused and meaning-focused. They were (1) speed of reading and acceptance in a grammaticality judgment task where half of the items were real phrases in English and half were not; (2) rate of reading and rate of spoken articulation; (3) binding and primed pronunciation – the degree to which reading the beginning of the formula primed recognition of its final word; and (4) speed of comprehension and acceptance of the formula as being appropriate in a meaningful context. Processing in all experiments was affected by various corpus-derived metrics: length, frequency, and mutual information (MI). Frequency was the major determinant for non-native speakers, but for native speakers it was predominantly the MI of the formula which determined processability.

Repetition also leads to automatization and fluency of production (Segalowitz 2010; DeKeyser 2001; Anderson 1992). Forms that are used highly frequently become phonologically eroded. 'Words used together fuse together' (Bybee 2003) (after Hebb's (1949) research often summarized by the phrase 'Cells that fire together, wire together') (see also Ellis, this volume, Chapter 4, Section 4.2).

6.4.7 Chunks in Sentence Processing

There is a substantial literature demonstrating sensitivity to such sequential information in sentence processing (see MacDonald and Seidenberg 2006 for review).

Consider sentences (1) beginning 'The plane left for the . . .' Does the second word refer to a geometric element, an airplane or a tool? Does the third imply a direction, or is it the past tense of the verb leave in active or in passive voice?

(1) a. The plane left for the East Coast.
b. The plane left for the reporter was missing.

What of the likelihood of the past tense passive interpretation of leave in the sentence beginning 'The note left for the . . .' as in (2). Is it greater or less than that for the plane sentence (1b)?

(2) The note left for the reporter was missing.

Psycholinguistic experiments show that fluent adults resolve such ambiguities by rapidly exploiting a variety of probabilistic constraints derived from previous experience (Seidenberg 1997). There is the first-order frequency information: *plane* is much more frequent in its vehicle meaning than in its other possible meanings, *left* is used more frequently in active rather than passive voice. The ambiguity is constrained by the frequency with which the ambiguous verb occurs in transitive and passive structures, of which reduced relative clauses are a special type (MacDonald 1994; MacDonald et al. 1994). On top of this there are combinatorial constraints: sentence (2) is easier to comprehend as a reduced relative clause than sentence (1b) because it is much more plausible for a note to be left than for it to leave (Trueswell et al. 1994).

These psycholinguistic studies of sentence processing show that fluent adults have a vast statistical knowledge about the behavior of lexical items and the chunks they inhabit in their language. Fluent comprehenders know the relative frequencies with which particular verbs appear in different tenses, in active vs. passive and in intransitive vs. transitive structures, the typical kinds of subjects and objects that a verb takes, and many other such facts. This information is relevant at all stages of lexical, syntactic and discourse comprehension (Seidenberg and MacDonald 1999). Frequent analyses are preferred to less-frequent ones.

Eye-movement research shows that the fixation time on each word in reading is a function of the frequency of that word (frequent words have shorter fixations) and of the forward transitional probability (the conditional probability of a word given the previous word P(wk|wk–1): for example, the probability of the word *in*, given that the previous word was *interested*, is higher than the probability of *in* if the last word was *dog*) (McDonald and Shillcock 2003a, 2003b). Parsing time reflects the more-frequent uses of a word (e.g. the garden-path effect caused by *The old man the bridge*, in which *man* is used as a verb). Phrase-frequency affects parsing in a similar way. For example, ambiguity resolution is driven not only by how often a verb appears as a past participle and how likely a noun is to be an agent, but also by the exact frequencies of the noun-verb combination. Reali and Christiansen (2007) demonstrate such effects of chunk frequency in the processing of object relative clauses. Sentences such as *The person*

who I met distrusted the lawyer, are easier to process when the embedded clause is formed by frequent pronoun-verb combinations (*I liked* or *I met*) than when it is formed by less-frequent combinations (*I distrusted* or *I phoned*).

Generally, analyses of large corpora of eye-movements recorded when people read text demonstrate that measures of surprisal account for the costs in reading time that result when the current word is not predicted by the preceding context. Measuring surprisal requires a probabilistic notion of linguistic structure (utilizing transitional probabilities or probabilistic grammars). The surprisal of a word in a sentential context corresponds to the probability mass of the analyses that are not consistent with the new word (Demberg and Keller 2008).

6.4.8 Hierarchies of Chunking

This research demonstrates that associative learning from usage results in chunking at all levels of language. Language knowledge involves statistical knowledge, so humans learn more easily and process more fluently high-frequency forms and 'regular' patterns which are exemplified by many types and which have few competitors. Usage-based perspectives of acquisition thus hold that language learning is the implicit associative learning of representations that reflect the probabilities of occurrence of form-function mappings. Frequency is a key determinant of acquisition because 'rules' of language, at all levels of analysis from phonology through syntax to discourse, are structural regularities which emerge from learners' lifetime unconscious analysis of the distributional characteristics of the language input.

6.5 Connectionism and Statistical Language Learning

Psycholinguistics demonstrates the ubiquity of chunking in language; connectionism and statistical language learning approaches investigate chunking in acquisition and processing. Connectionist theories are data-rich and process-light: massively parallel systems of artificial neurons use simple learning processes to statistically abstract information from corpora of representative input data (Elman et al. 1996; Christiansen and Chater 2001; Rumelhart and McClelland 1986). The work of Elman (1990) on 'finding structure in time' was influential in demonstrating the types of syntagmatic and semantic structures that are emergent from linguistic sequences.

6.5.1 Phonological Sequences

Elman (1990) used a simple recurrent network (SRN) to investigate the temporal properties of sequential inputs of language. The network was fed one letter at a time and had to predict the next letter in the sequence. It was trained on 200 sentences where there was no word or sentence boundary information. The network abstracted a lot of information about the structure of English. It learned about orthographic sequential probabilities; it learned that there were common recurring units (which we might identify as morphemes and words); it extracted word sequence information, too. At times, when the network could not predict the actual next phoneme, it nonetheless predicted the correct category of phoneme: vowel/consonant, etc. Thus it moved from processing mere surface regularities to representing something more abstract, but without this being built in as a pre-specified constraint: linguistically useful generalizations emerged. Simple sequence learning processes learned regular chunks like words, bound morphemes, collocations and idioms; they learned regularities of transition between these surface chunks; and they acquired abstract generalizations from the patterns in these data.

Such chunks are potential labels, but what about reference? The more any word or formula is repeated in phonological working memory, the more its regularities and chunks are abstracted, and the more accurately and readily these can be called to working memory, either for accurate pronunciation as articulatory output or as labels for association with other representations (e.g. Ellis 1994a). It is from these potential associations with other representations that other interesting properties of language emerge. I will return to this is Section 6.6.

6.5.2 Syntactic Sequences

Learning the grammatical word-class of a particular word, and learning grammatical structures more generally, involves the automatic implicit analysis of the word's sequential position relative to other words in the learner's stock of known phrases which contain it. Elman (1990) trained a recurrent network on sequences of words following a simple grammar, the network having to learn to predict the next word in the sequence. At the end of training, he cluster-analyzed the representations that the model had formed across its hidden unit activations for each word+context vector. This showed that the network had discovered several major categories of words – large categories of verbs and nouns, smaller categories of

inanimate or animate nouns, smaller-still categories of human and nonhuman animals, etc. (for example, 'dragon' occurred as a pattern in activation space in the region corresponding to the category animals and also in the larger region shared by animates, and finally in the area reserved for nouns). The category structure was hierarchical, soft and implicit. The network moved from processing mere surface regularities to representing something more abstract, but without this being built in as a pre-specified syntactic or other linguistic constraint and without provision of semantics or real-world grounding. Relatively general architectural constraints gave rise to language-specific representational constraints as a product of processing the input strings. These linguistically relevant representations were an emergent property of the network's functioning (see Redington and Chater 1998 for larger analyses of this type on corpora of natural language). Learning the grammatical categories and requirements of words and word groups reduces to the analysis of the sequence in which words work in chunks.

6.5.3 Statistical Language Learning

Saffran, Aslin and Newport (1996) demonstrated that eight-month-old infants exposed for only 2 minutes to unbroken strings of nonsense syllables (for example, *bidakupado*) are able to detect the difference between three-syllable sequences that appeared as a unit and sequences that also appeared in their learning set but in random order. Statistical language learning has since become a major research field, demonstrating, in infant language acquisition (Molnar and Sebastian-Galles 2014) and child language acquisition (Rebuschat and Williams 2012), how language learners implicitly learn the statistics of the language to which they are exposed, and how the representational chunks that emerge from this implicit learning form a rich system that, through 'repeated cycles of integration and differentiation' (Studdert-Kennedy 1991), associates phonology, syntax and semantics 'in richly structured and productive ways' (MacWhinney and O'Grady 2015).

There is much research into the types of statistical learning that are possible, both implicitly and explicitly. Figure 6.2 panel 1 illustrates some of the factors that affect sequential associative learning, including the transparency of the underlying structure and the units over which learning is taking place. In particular, while sequential dependencies can be implicitly learned, discontinuous dependencies are more problematic and may require working memory representation and explicit learning (Rebuschat

and Williams 2012). Figure 6.2 panel 2 illustrates how some units in the speech stream are much more salient than others, and therefore are more likely to enter into implicit learning (see my Chapter 4 in this volume, Section 4.1.3.1 – Psychophysical Salience).

6.6 Chunks, Symbols and Constructions

Chunking does not only take place within the sequences of language form. Chunking binds form with meaning in symbolic constructions. Many of the examples in Section 6.4 related to cross-modal association between, for example, print and sound in reading, or form and meaning in sentence processing. Constructions are form-meaning mappings, conventionalized in the speech community, and entrenched as language knowledge in the learner's mind. They are the symbolic units of language relating the defining properties of their morphological, syntactic and lexical form with particular semantic, pragmatic and discourse functions (Croft and Cruise 2004; Robinson and Ellis 2008; Goldberg 1995, 2003, 2006; Croft 2001; Tomasello 2003; Bates and MacWhinney, 1987; Langacker 1987; Lakoff 1987; Bybee 2008b).

Broadly, Construction Grammar argues that all grammatical phenomena can be understood as learned pairings of form (from morphemes, words and idioms, to partially lexically filled and fully general phrasal patterns) and their associated semantic or discourse functions. Figure 6.2 panel 3 illustrates such a form-function mapping. Whereas sequential learning may well take place implicitly, at least for adjacent elements, the seeding of cross-modal associations is usually a result of conscious, explicit processing, although thereafter, the strengths of these associations are also implicitly tuned during usage.

6.6.1 Learning Novel Form-Meaning Associations

Research on explicit learning (e.g. Ellis 2005) has shown how conscious processing promotes the acquisition of novel, explicit, cross-modal, form-meaning associations. These breathe meaning into the processing of language form. Form-meaning chunks are symbolic constructions. Learning a new symbol, for example a lexical construction, as an explicit declarative memory from a sound-image episode such as 'étoile'-★ involves explicit learning (Ellis 1994a). The primary conscious involvement in language acquisition is the explicit learning involved in the initial registration of pattern recognizers for constructions that are then tuned and integrated into the system by implicit learning during subsequent input processing.

Neural systems in the prefrontal cortex involved in working memory provide attentional selection, perceptual integration and the unification of consciousness. Explicit learning results in explicit memories.

Neural systems in the hippocampus bind disparate cortical representations into unitary episodic representations (Ellis 2005: 305). By forming unitized memory representations, the hippocampal region performs the information-processing function of forming pattern-recognition units for new stimulus configurations and of consolidating new bindings; these are then adopted by other brain regions in the neocortex where they subsequently partake in implicit tuning (Gluck et al. 2003). Once such cross-modal chunks have been consolidated, these representations are also then available as units of implicit learning in subsequent processing, allowing statistical learning and tallying of form-meaning contingencies. Some of the cross-modal associations are much richer in their perceptual imagery than others (see my Chapter 4 in this volume, Section 4.1.3.2, Salient Associations). Figure 6.2 panel 4 illustrates this.

The function relating strength of association and frequency of experience is the power law of practice. Like other stimuli, linguistic forms are typically ambiguous. The same form can attract different meanings. So there is competition between the different meaning candidates when it comes to interpretation (see Chapter 12 in this volume, on ambiguity). As described in Section 6.4.7, parsing and comprehension are probabilistic processes. The resolution of this competition depends upon the contingency of form and functions in prior experience.

6.6.2 Contingency

Because linguistic forms are ambiguous and carry multiple meanings with varying strengths of association, it is not just the frequency of encounter of a construction that determines its acquisition. The degree to which animals, human and other alike, learn associations between cues and outcomes depends upon the contingency of the relationship as well. In classical conditioning, it is the reliability of the bell as a predictor of food that determines the ease of acquisition of this association (Rescorla 1968). In language learning, it is the reliability of the form as a predictor of an interpretation that determines its acquisition (MacWhinney 1987a). The last thirty years of psychological investigation into human sensitivity to the contingency between cues and outcomes (Shanks 1995) demonstrates that when, given sufficient exposure to a relationship, people's judgments match quite closely the contingency specified by ΔP (the one-way dependency statistic, Allan

Table 6.1 *A contingency table showing the four possible combinations of events, showing the presence or absence of a target cue and an outcome*

	Outcome	No outcome
Cue	a	b
No cue	c	d

1980), which measures the directional association between a cue and an outcome, as illustrated in Table 6.1.

In the table, a, b, c, and d represent frequencies, so, for example, a is the frequency of conjunctions of the cue and the outcome, and c is the number of times the outcome occurred without the cue.

$$\Delta P = P(O|C) - P(O|\neg C) = \frac{a}{a+b} - \frac{c}{c+d}$$

ΔP is the probability of the outcome, given the cue (P(O|C) minus the probability of the outcome in the absence of the cue ($P(O|\neg C)$). When these are the same, when the outcome is just as likely when the cue is present as when it is not, there is no covariation between the two events and ΔP = 0. ΔP approaches 1.0 as the presence of the cue increases the likelihood of the outcome and approaches –1.0 as the cue decreases the chance of the outcome – a negative association.

Some cues, especially grammatical functors, are ambiguous in their interpretations, and this makes them difficult to learn (Figure 6.2, panel 5). Connectionist and psycholinguistic research shows that the strength of association between a linguistic form and an interpretation is also updated implicitly from usage, and the likelihood that a particular interpretation comes to mind is a function of the relative strengths of association of the various possible outcomes.

6.7 Chunking in Language Change

I have described the learning theory and psycholinguistic evidence of chunking: how each episode of usage strengthens the relevant associations, and how these effects cumulate into syntagmatic frequency effects whereby more-frequent linguistic forms are preferentially recognized and more fluently produced, as well as associative frequency effects whereby

interpretation and expression of form-function mappings reflect the satisfaction of statistical constraints. Chunking provides a rational representation of usage. It both builds the representations and organizes their relative availability and fluency according to need.

6.7.1 The Principle of Least Effort Shapes an Artisan's Tools

The work of Zipf (1935, 1949) provides comprehensive empirical evidence of the effects of these processes in language structure, usage, and change. Zipf's (1949) groundbreaking analysis of the ecology of language centers upon communicative function, where linguistic constructions are tools for sharing meanings. He laid the foundations by carefully crafting a tool analogy to illustrate the operation of the *principle of least effort*. It is a productive and provocative metaphor, *go with it*:

> An artisan works at his bench, with *n* different tools of various sizes and weights arranged on a straight board in front of him as he sees fit. His occupation is to perform for us certain jobs using his tools as economically as possible. We do not care how many tools he uses, nor how he alters their shape, weight, and usage, nor how he arranges them upon the board.

The work of using a tool consists of transporting it from its place on the board to the artisan's lap and then back again after its use. Over time, he adapts the arrangement of his tools according to their usage, taking account of the mass *m* of the tool, the distance *d* away on the board and the frequency of use *f* (Zipf 1949: 59). In order to use his tools with the maximum economy *'he must arrange the n tools of his shop in such a way that the sum of all of the products of* $f \times m \times d$ *for each of the n tools will be a minimum'* (1949: 59). You will be reminded now of rational analysis as discussed in Section 6.3.1 – clearly, the most frequently used tools should be kept closer to hand. But various additional economic principles apply. Ideally, there should be a 'close packing' of tools, for then *d* is reduced. From this follows the *principle of the abbreviation of the size* – the smaller the size *s* of the tools and their mass *m* can become, the more closely they can be packed together – as well as the *principle of the abbreviation of mass* – reducing the mass *m* of the tools will also lessen the work ($m \times d$) (1949: 61). The redesigning of tools according to these principles should take account of their frequency of use: 'the artisan will lay a premium upon the reductions of the sizes of all of the tools in proportion to their nearness to him' (1949: 61). The arrangement aims to more closely pack together the tools as well as to reduce their *n* number.

So much for the forms of the tools. But what of their functions? The more functions a tool can perform, the fewer the total necessary number *n* of tools. In their redesign, it is economical to adapt the easiest tool so that it absorbs the jobs of other less-easy tools and thereby increase its own frequency of use still more. In increasing the frequency of the easiest tool, the easier its use is made by abbreviation – and the easier the tool's use is made by abbreviation, the more frequently it is used. *'In short, greater frequency makes for greater ease which makes for greater frequency and so on'* (1949: 62).

> As a result of the artisan's redesignings, every one of the tools can have been altered in form and function from its original state beyond all present recognition. Some tools may have changed their form but preserved their usage; By definition, this is a *formal change*. Some tools may have preserved their form but changed their usage; by definition this is a *semantic change*. And some may have done both and others may have done neither.
>
> Nevertheless, whatever alterations were or were not undertaken from moment to moment in the course of the shop's history, they were all undertaken, or not undertaken, as a response to the minimizing of the total work of the shop, according to the 'minimum equation', which directly or indirectly refers to all form, function, and arrangement. Therefore we may say that, from moment to moment, the shop was seeking to preserve by definition a *formal-semantic balance in the forms or usages of its tools*. (1949: 63)

Over time, the more frequent tools tend to become lighter, smaller, older, more versatile and more thoroughly integrated with the action of other tools because of their permutations of use with them. They are also the most valuable tools in the system, in that their permanent loss would cause the relatively greatest cost of redesigning and retooling. Hence, it is most economical to conserve the most frequent tools.

6.7.2 The Principle of Least Effort Shapes Linguistic Tools

Zipf's artisan's board is straight in order to allow his tool analogy to parallel the one-dimensionality of the serial speech stream. (*Get it?* – How bland are these two words individually, but how potent in the holophrase, in the context.) In the next 150 pages, he extends the principle of least effort to the economy of language and how this shapes language change. In the evolution of human behavior and 'all trades, their gear and tackle and trim' (Hopkins 1918), flint, bone and rocks have undergone formal and semantic changes, emerging as spades, Swiss-Army knives, Brown (#4) Robertson-head screwdrivers and all manner of specialist tools. Likewise the evolution of language involves formal changes – for example, *telephone, gasoline* and

omnibus have become *phone, gas* and *bus* – and semantic changes where shorter words have been substituted for longer ones – for example, *car* for *automobile*, or *juice* for *electricity* – and the shorter substitutions have taken on the specialized meaning of the longer word: *juice* now means 'electricity', among other things (Zipf 1949: 67).

His theoretical insights are matched by admirable empirical effort and precision. He reports extensive analyses of many corpora of some thousands of words from different authors, genres, languages and eras. He counts words as tools, measuring their frequency, their packing as a function of length and other aspects of psychophysical mass, their semantic versatility, their degree of collocation or permutation with other words and their age in the language. This work is particularly impressive, given that it was performed when *computer* meant 'a person who makes calculations', well before the digital age. His analyses reveal several *universal laws of language*.

6.7.3 Universal Laws of Language

The most fundamental of these laws, now eponymous as *Zipf's law*, describes the frequency distribution of words: the frequency of words f decreases as a power function of their rank r in the frequency table. Thus the most frequent word (in English, *the*, with a token frequency of ~60,000/million words) occurs approximately twice as often as the second-most-frequent word, three times as often as the third-most-frequent word, etc. The rank-frequency relationship, since $r \times f = C$, 'appears on doubly logarithmic chart paper as a succession of points descending in a straight line from left to right at an angle of 45°' (Zipf 1949: 24[1]). Zipf demonstrated that this scaling law holds across a wide variety of language samples. It has been confirmed repeatedly since.

Next is the *law of abbreviation of words*:

> Every language is demonstrably undergoing formal and semantic changes which act on the whole in the direction of shortening the sizes of longer words, or of increasing the frequency of shorter words. Moreover, as far as we know, every language shows an inverse relationship between the lengths and frequencies of the usage of its words. (Zipf 1949: 66).

The longer the period of usage over which this shortening can apply, the shorter the resultant word. Zipf (1949: 111) shows graphs relating word length (number of syllables, 1–6), period (Old English, Middle English, 15th, 16th, 17th, 18th, 19th century) and percentage coverage of the newspaper text, whereby the sizes and frequencies of words are inversely related

to their age, with the result that the longer and less-frequent words tend to be the younger ones.

The *principle of the economical versatility of words* describes how the number of different 'meanings' of a word decreases according to the square root of its frequency, i.e. $mr = \sqrt{Fr}$. His first demonstration of this, figure 2.2 (Zipf 1949: 29–30), concerns how the average number of different meanings of the twenty successive sets of 1,000 words in the Thorndike Frequency Count of English, when ranked in order of decreasing frequency, decrease in proportion to the square root of the rank.

The *principle of the economical permutation of words* describes how the number of different permutations into which a word enters, as along with the frequencies with which the permutations occur, is directly related to the frequency of the word. Thus, for example, the more frequently a word appears in the language, the ever more frequently it is used in holophrases.

Many of these principles work at other sizes of grain. For example, the *Economical Permutation of Morphemes in Words* describes how the magnitudes of morphemes decrease as their frequencies increase.

The impulse behind these various principles is the economy that comes from the 'close packing' of tools and the reduction of their *n* number. Because of this influence,

> there is a tendency for old age, small size, versatility of meaning, and a multiplicity of permutational associations all to be directly correlated with high frequency of usage . . .
>
> All the Principles are all constantly operating simultaneously, for the preservation of a dynamic equilibrium with a maximum of economy. That is, in dynamic processes, words are constantly being shortened, permuted, eliminated, borrowed, and altered in meaning" (Zipf 1949: 121).

These are serious linguistic universals. Zipf's work is astonishing in so many ways: it pioneered the ubiquity of power law relationships in complex systems, corpus analysis and empirical linguistics, dynamic systems theory, rational analysis and the recognition that psycholinguistic processes and the structures and functions of language are inextricably linked in usage.

6.7.4 A Zipfian Analysis of Contemporary English

For the case of language change in English as it relates to chunking, it is instructive to update his law of abbreviation analysis using more-modern corpora and statistical techniques. Figure 6.3 shows the relation between length in phonemes (top) and alphabetic letters (bottom) and log frequency of occurrence for the 54,447 word form types constituting more than

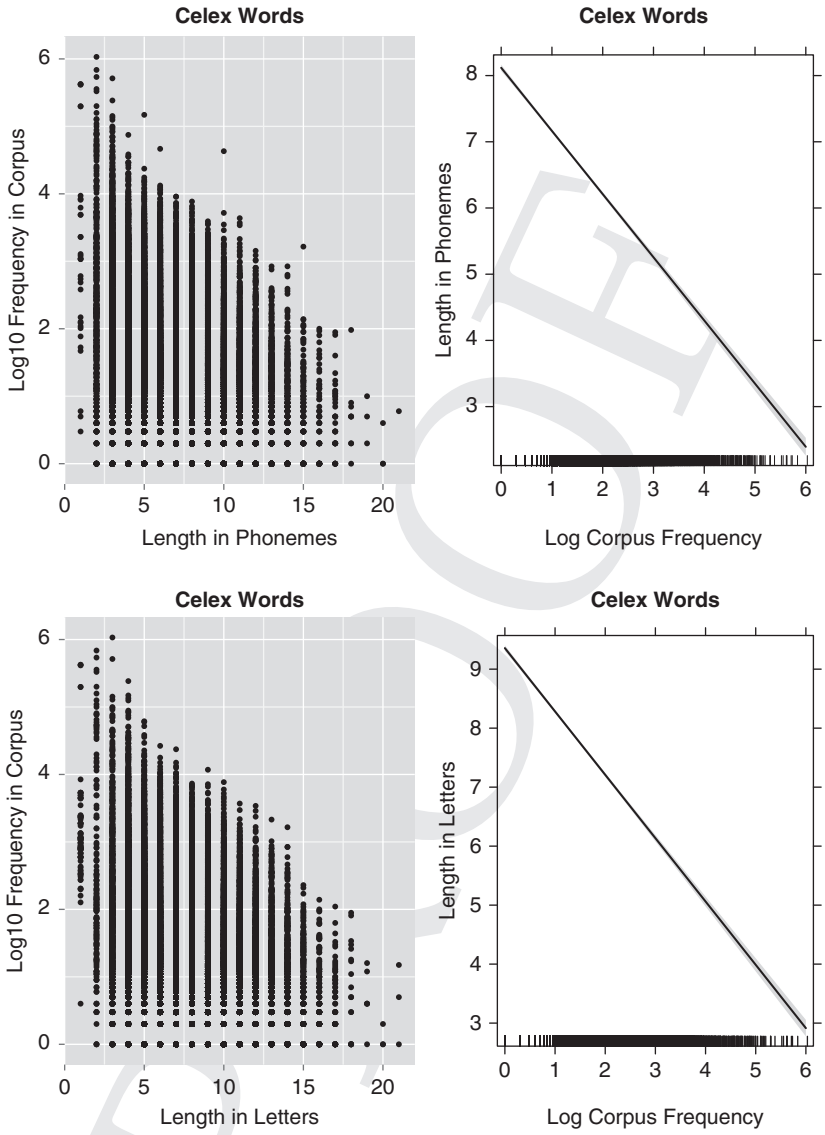

Figure 6.3 The relation between length in phonemes *(top)* and in letters *(bottom)* and log frequency of occurrence for English words in the CELEX database. Note that in each column there are many more observations at lower frequencies than at higher ones. Indeed, in each column there is a Zipfian frequency distribution (though you cannot see the long tail here). So these figures are driven by two of Zipf's universals. The right-hand plots show the regression line relating length and log frequency.

18 million word tokens in the CELEX lexical database of English. The law of abbreviation of words clearly applies equally to spoken and written word forms.

Figure 6.4 shows the relation between length (in letters) and log frequency of occurrence for the top 5,000 words, 2-word, 3-word, 4-word,

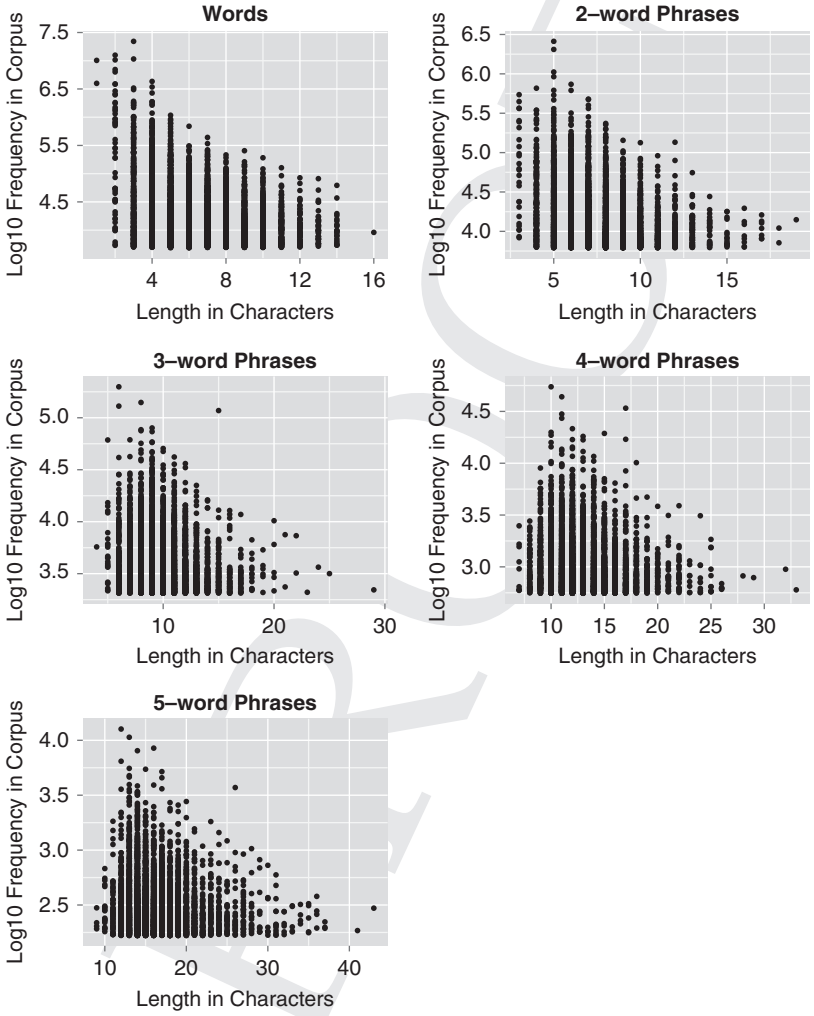

Figure 6.4 The relation between length (in letters) and log frequency of occurrence for the top 5,000 words, two-word-, three-word, four-word and five-word phrases of English from COCA.

and 5-word phrases from the largest publicly available, genre-balanced corpus of English: the 450-million-word Corpus of Contemporary American English (COCA)[2]. I use letter length here rather than phoneme length for convenience, given the same effects upon spoken and written form in Figure 6.3. Again, the law of abbreviation of words clearly applies. The inverse association of length and frequency clearly holds for 2- to 5-gram phrases, too. The shaping of these words is clearly shown in Table 6.2, which shows the top six along with six examples from the

Table 6.2 *Rank, frequency, example ngram and length of 1-, 2-, 3-, 4- and 5-grams in the 450-million-word Corpus of Contemporary American English (COCA)*

Rank	Frequency	Ngram	Length
Words (1-grams)[1]			
1	22,038,615	The	3
2	12,545,825	Be	2
3	10,741,073	And	3
4	10,343,885	Of	2
5	10,144,200	A	1
6	6,996,437	In	2
. . ..			
99,914	17	Septicaemia	11
100,033	17	slum-dwellers	13
100,112	16	Mistily	7
100,200	16	ill-bred	8
100,223	16	Spackle	7
100,439	16	Reappointed	11
2-grams[2]			
1	2,586,813	of the	5
2	2,043,262	in the	5
3	1,055,301	to the	5
4	920,079	on the	5
5	737,714	and the	6
6	657,504	to be	4
. . .			
1,020,318	23	zoo Atlanta	10
1,020,319	23	zoo director	11
1,020,337	23	zoom is	6
1,020,348	23	zooms out	8
1,020,375	23	zulu nation	10
1,020,380	23	Zurich to	8

Table 6.2 *(cont.)*

Rank	Frequency	Ngram	Length
3-grams			
1	198,630	I do n't	6
2	140,305	one of the	8
3	129,406	a lot of	6
4	117,289	the United States	15
5	79,825	do n't know	9
6	76,782	out of the	8
. . .			
1,011,682	25	youth may be	10
1,011,694	25	youths who had	12
1,011,726	25	Zero to Three	11
1,011,736	25	zip code and	10
1,011,750	25	zone is the	9
1,011,760	25	Zoning Board of	13
4-grams			
1	54,647	I do n't know	10
2	43,766	I do n't think	11
3	33,975	in the United States	17
4	29,848	the end of the	11
5	27,176	do n't want to	11
6	21,537	the rest of the	12
. . .			
1,001,168	13	you've got to get	13
1,001,170	13	Yugoslav republic of Macedonia	27
1,001,177	13	zero in New York	13
1,001,191	13	Zoe Baird and Kimba	16
1,001,200	13	zucchini and yellow squash	23
1,001,201	13	Zukerman joins us now	18
5-grams			
1	12,663	I do n't want to	12
2	10,663	at the end of the	13
3	8,484	in the middle of the	16
4	8,038	I do n't know what	14
5	6,446	I do n't know if	12
6	5,551	I do n't think it	13
. . .			
989,575	6	Zero Tolerance Approach to Punctuation	34

Table 6.2 *(cont.)*

Rank	Frequency	Ngram	Length
989,576	6	zest cup fresh lemon juice	22
989,577	6	Ziggy Marley and the Melody	23
989,579	6	zinc oxide or titanium dioxide	26
989,586	6	Zukerman joins us now to	20
989,587	6	Zulu nationalist Inkatha Freedom Party	34

[1] Word frequencies www.wordfrequency.info/100 k_samples.asp (Retrieved November 28, 2014)

[2] 1 million most frequent 2-, 3-, 4- and 5-grams in the largest publicly available, genre-balanced corpus of English – the 450-million-word Corpus of Contemporary American English (COCA) www.ngrams.info/ Retrieved November 28, 2014

bottom of ranks of the approximately 100,000 most-frequent words in the corpus. Indeed, there is a tendency for old age, small size, versatility of meaning and a multiplicity of permutational associations all to be directly correlated with the highest frequency of usage here. These are the words which take the most pages of explanation of their many meanings and functions in major dictionaries (e.g., Simpson and Weiner 1989) and grammars (e.g., Biber et al. 1999). These are the words which enter the majority of different colligational permutations with other words, as well as the frequencies with which the permutations occur. These are the words which, if lost as a whole, would cause the relatively greatest cost of redesigning and retooling the grammar of English.

Note that the use of orthography blunts the shortening effects at highest frequencies, where words like *the* and *and*, which have three orthographic segments, are spoken as fewer phonemes, *the* as two and *and* being usually produced with only one. Jurafsky et al. (2001) used a phonetically hand-transcribed subset of 38,000 tokens from the Switchboard corpus to gauge the role of frequency, measuring word length on an acoustic representation of small subsets of words. They show that function words that are more predictable are shorter and more likely to have reduced vowels, supporting a probabilistic reduction hypothesis whereby the conditional probability of

the target word, given the preceding word and given the following one, plays a role on both duration and deletion.

Table 6.2 also illustrates the top six as well as a sample of lowest 1,000,000th-order 2-, 3-, 4-, and 5-grams (the lower-frequency examples are also from the end of the alphabet, since the lists are sequenced first by frequency, then by alphabetical order). The higher-frequency phrases are much shorter than the lower-frequency ones, and they tend to serve distinct grammatical or discourse functions. So too, the higher-frequency phrases illustrate the dynamics of chunking and contraction in process, with the multiple exemplars of *I don't* and *you've got*, as described in Bybee (2006). *I* is by far the most frequent pronominal subject of *don't* (210,940) and *you* is the most frequent pronominal subject of *'ve got* (*you* 25,765, *I* 17,535 as I search in COCA now).

Zipf's theoretical and empirical influences are very much in evidence in present-day research (e.g. Ferrer i Cancho and Solé 2003; Kello et al. 2010; Wiechmann et al. 2013; Williams et al. 2014; Thurner et al. 2015; Corral et al. 2015; Ellis et al. 2016).

6.7.5 Grammaticalization

Bybee (2010, 1998, this volume, 2003) and Bybee and Hopper (2001) have developed a model of grammaticization as the process of automatization of frequently occurring sequences of linguistic elements. With repetition, sequences of units that were previously independent come to be processed as a single unit or chunk. This repackaging has two consequences: the identity of the component units is gradually lost, and the whole chunk begins to reduce in form. As described above, these basic principles of automatization apply to all kinds of motor activities: playing a musical instrument, cooking or playing an Olympic sport. They also apply to grammaticization. A phrase such as *(I'm) going to (verb)* which has been frequently used over the last couple of centuries, has been repackaged as a single processing unit. The identity of the component parts is lost (children are often surprised to see that *gonna* is actually spelled *going to*), and the form is substantially reduced.

Thus, in Bybee's model, frequency and chunking are driving forces of language change: (1) frequency of use leads to weakening of semantic force by habituation; (2) phonological changes of reduction and chunking/fusion of grammaticizing constructions are conditioned by their high frequency; (3) increased frequency conditions a greater autonomy for a construction, which means that the individual components of the

construction (such as *go, to* or *-ing* in the example of *be going to*) weaken or lose their association with other instances of the same item (as the phrase reduces to *gonna*); (4) the loss of semantic transparency accompanying the rift between the components of the grammaticizing construction and their lexical congeners allows the use of the phrase in new contexts with new pragmatic associations, leading to semantic change; and (5) autonomy of a frequent phrase makes it more entrenched in the language and often conditions the preservation of otherwise-obsolete morphosyntactic characteristics (Bybee 2003).

6.7.6 Other Domains

Section 6.4.4 described how productivity of phonological, morphological and syntactic patterns is a function of type rather than token frequency (Bybee and Hopper 2001), whereas high token frequency promotes the entrenchment or conservation of irregular forms and idioms. The irregular forms only survive because they are high frequency.

For type and token frequency, and the effects of friends and enemies in the dynamics of productivity of patterns in language evolution, Lieberman, Michel, Jackson, Tang and Nowak (2007) studied the regularization of English verbs over the past 1,200 years. English's proto-Germanic ancestor used an elaborate system of productive conjugations to signify past tense, whereas Modern English makes much more productive use of the dental suffix, '-ed'. Lieberman et al. chart the emergence of this linguistic rule amidst the evolutionary decay of its exceptions. By tracking inflectional changes to 177 Old English irregular verbs, of which 145 remained irregular in Middle English and 98 are still irregular today, they showed how the rate of regularization depends on the frequency of word usage. The half-life of an irregular verb scales as the square root of its usage frequency: a verb that is 100 times less frequent regularizes 10 times as fast.

There is a rich literature on frequency effects in the chunking of compound morphology as well. Baayen et al. (2010) analyzed the processing times of English and Dutch compounds in word naming, lexical decision and eye-tracking as a function of the compound token frequency, head and modifier token frequency and head and modifier compound family sizes (type frequencies) in the reading of English and Dutch compounds to show effects of these frequency measures independently as well as in many complex dynamic interactions.

Constructions are nested and overlap at various levels (morphology within lexis within grammar, hierarchical semantic organizations, etc.).

Sequential elements can be memorized multiple times as wholes at these different levels. So there is no one direction of growth, but rather continuing interplay between modalities, between top-down and bottom-up processes and between memorized structures and more-open constructions. Constructions develop hierarchically by repeated cycles of differentiation and integration. This is why we need to go beyond univariate statistics, beyond multivariate statistics still, toward computational modeling (richly informed by corpus data), and why there is sense in viewing language as a complex adaptive system (Beckner et al. 2009).

As usage frequencies affect processing, so they affect language change. In the orthography, lower-frequency compounds are transcribed as two words, whereas higher-frequency compounds become individual lexical entities in their own right (compare *goat herd, pig man, shepherd, cowboy*). These are results of associative learning too. According to the *Shorter Oxford English Dictionary*, the word *pineapple* (from *pine* + *apple*) was originally used in late Middle English to refer to the reproductive organs of conifer trees (now known as *pine cones*). When European explorers discovered the tropical fruit *Ananas comosus* in the Americas, because they looked like (what we now call) pine cones, they named them *pine-apples* (first referenced in 1664). Zipf would appreciate the dynamics of the formal-semantic balance through which has evolved, in contemporary English, pineapple coming in chunks.

Notes

1. Not only the computations, but the graphs were drawn by hand too.
2. www.ngrams.info/ Retrieved November 28, 2014.