Learning: Statistical mechanisms in language acquisition

Elizabeth Wonnacott

Abstract The grammatical structure of human languages is extremely complex, yet children master this complexity with apparent ease. One explanation is that we come to the task of acquisition equipped with knowledge about the possible grammatical structures of human languages — so-called "Universal Grammar". An alternative is that grammatical patterns are abstracted from the input via a process of identifying reoccurring patterns and using that information to form grammatical generalizations. This statistical learning hypothesis receives support from computational research, which has revealed that even low level statistics based on adjacent word co-occurrences yield grammatically relevant information. Moreover, even as adults, our knowledge and usage of grammatical patterns is often graded and probabilistic, and in ways which directly reflect the statistical makeup of the language we experience. The current chapter explores such evidence and concludes that statistical learning mechanism play a critical role in acquisition, whilst acknowledging holes in our current knowledge, particularly with respect to the learning of 'higher level' syntactic behaviours. Throughout, I emphasize that although a statistical approach is traditionally associated with a strongly empiricist position, specific accounts make specific claims about the nature of the learner, both in terms of learning mechanisms and the information that is primitive to the learning system. In particular, working models which construct grammatical generalizations often assume inbuilt semantic abstractions.

1 Introduction

Speaking at least one language is a ubiquitous human ability. Wherever humans are discovered, whatever else they are doing, they are talking. Conversing in our mother-

Elizabeth Wonnacott
Department of Psychology, University of Warwick, Coventry CV4 7AL, UK, e-mail: E.A.Wonnacott@warwick.ac.uk

tongue feels so effortless that it is rarely regarded as a skill or accomplishment, yet we know that this behaviour relies on a highly complex body of knowledge about the structure of that language. This knowledge is sometimes called 'grammatical', though it is important to realize that we are not talking about the sorts of prescriptive grammatical rules which (depending on the decade) we may have been taught at school. Rather, the type of grammatical knowledge which is the concern of Cognitive Science is what psychologists call *implicit* knowledge, i.e. knowledge which is subconscious and largely inaccessible to the speaker.

Consider an example at the level of *morphology* — the constraints governing how meaningful strings of sounds, *morphemes*, may combine to form words. An English speaker implicitly knows that the regular past tense ending, the one we write as 'ed', differs according to the sound of the final consonant of the verb stem: if the verb ends in a /t/ or /d/ sound, the past tense is pronounced as /əd/ (e.g. *loaded*); if it ends in a voiceless sound (i.e. one produced without the vocal chords vibrating) it is pronounced as /t/ (e.g. *liked*); if it ends in a voiced sound (i.e. one produced with the vocal chords vibrating) it is pronounced /d/ (e.g. *loved*). This knowledge goes beyond a memory for the forms of individual verbs, since we are able to produce appropriate past forms for new verbs — try using the verbs *wid*, *wuf* and *wug* in the past tense (Berko, 1958). As we will see, there is considerable debate about how such generalizations should be characterized. The point here is simply that such patterns must somehow be incorporated into our implicit knowledge of English "grammar".

For an example at the level of syntax (the constraints governing how words combine into higher level structures), consider the following English sentence:

(1) a. *Jack threw Henry the ball*

Our understanding of this sentence includes not only the meanings of the individual words within it, but also the semantic roles imposed by their structural positions: *Jack* was the agent of the throwing action; *the ball* was the transferred object; *Henry* was the recipient. The following sentences, though composed of different words, have the same formal structure and exemplify the same semantic relationships:

(1) b. *Oliver sent William the parcel*
c. *Poppy gave Charlie her book*
d. *Jasmine told Jessica the news*

(Note that, as in the last example, the 'transfer' action may be metaphorical rather than physical). This relationship between an abstract structure and a semantic event is known as a *construction*. As with morphology, a new words test can reveal our implicit understanding of this generalized knowledge: given an appropriate context, we can spontaneously produce and understand the construction with new verbs as in 1e (Gropen et al., 1989):

(1) e. *He gorped me the ball*

Again, this suggests that a mental grammar of English must contain knowledge of the relationship between a general *X Verb Y Z* pattern and the semantic information it

conveys. Interestingly, however, the grammar must also contain information which *prevents* us from applying the construction in certain circumstances, in particular there are a number of English verbs which *can't* be used in this construction, as in the ungrammatical, though perfectly comprehensible English sentences 1f–g:

(1) f. * *Oliver* **explained** *William the news*
g. * *Jack* **carried** *Henry the ball*

Thus an account of language acquisition has to explain *both* how we acquire the generalizations and the exceptions to those generalizations.

One final example will further illustrate the type of abstract structures which play a role in our use and understanding of language. Consider the sentence:

(2) *Put the block in the box on the table*

Your school-learned grammar might allow you to identify the following underlying linguistic categories or "parts of speech" (where 'Det' stands for Determiner and 'Prep' for Preposition).

Put	the	block	in	the	box	on	the	table
Verb	Det	Noun	Prep	Det	Noun	Prep	Det	Noun

These abstract categorizations also feature in our implicit knowledge. That is, we know which particular set of English words can fill, for example, the *Preposition* slots above. Note that this is not simply a question of knowing the word's meaning, since categorization is partially arbitrary (consider that the equivalent for a word which is a preposition in one language may be a verb in another and vice versa: for example in Chinese the instrument reading of the English preposition *with* — as in *eat with chopsticks* — is the verb[1] *yong*: DeLancey, 2005). Moreover, our grammatical understanding of this sentence goes beyond an unstructured string of categories. Rather, we recognize that substrings of words may be grouped, and that this grouping affects our interpretation of the sentence. For example, the above sentence may be described using two different structural organizations, which can be shown with schematic bracketing — where NP stands for "noun phrase" and PP "prepositional phrase":

Put $_{\text{NP}}$*[the block]* $_{\text{PP}}$*[in* $_{\text{NP}}$*[the box* $_{\text{PP}}$*[on* $_{\text{NP}}$*[the table]]]]*
Put $_{\text{NP}}$*[the block* $_{\text{PP}}$*[in* $_{\text{NP}}$*[the box]]]* $_{\text{PP}}$*[on* $_{\text{NP}}$ *[the table]]*

The first structure imposes an interpretation in which a block is placed in a box which is situated on a table. The second imposes an interpretation where the block was initially reposing in a box and is then is moved to the table. Note that the structures labeled "PP" and "NP" are embedded within each other — yielding what linguists refer to as *hierarchical phrase structure*. This organizing principle is central

[1] More accurately, this word is usually categorized as a "co-verb". Li & Thompson (1974, cited in DeLancey, 2005) argue that co-verbs are graded in how syntactically "verb like" they are.

to our understanding of syntactic phenomena. For example, in English the relationship between a *statement* and *yes–no question* is that the entire *NP* which is the *subject*[2] of the verb inverts with the auxiliary — as in the following examples:

[*The boy who is happy*]$_{\text{NP}}$ **is** *singing?*
Is *[the boy who is happy]*$_{\text{NP}}$ *singing?*

The purpose of the above examples was to give the English speaking reader an insight into his or her implicit knowledge of the language. This, of course, only scrapes the surface of the intricacies of English grammar, and similar complexity underlies all human languages.[3] The topic of this chapter is how structural patterns of different levels of abstraction are acquired by native speakers. Perhaps the most remarkable feature of this learning is that, in normal circumstances, it occurs in early childhood: a good bulk of the grammatical system is in place by the age of four, meaning that the average child is in some sense a competent grammarian before she can brush her own teeth. It is clear that this not a result of explicit teaching. Few parents or teachers are aware of the types of patterns discussed above — and I doubt that any would relish the prospect of explaining the relevant concepts to a young child. Of course adult speakers do have an *intuitive* knowledge of the grammatical patterns of their native language, and so will be aware when their children produce utterances which are un-adult-like. However, studies have repeatedly shown that children receive very little explicit correction for grammatical errors (Braine, 1971; Brown & Hanlon, 1970; Newport et al., 1977).

Somehow, then, small children extract grammatical patterns via exposure to the language they hear around them *without explicit instruction*. Moreover the outcome of learning is very consistent, i.e. native speakers largely agree in their grammatical intuitions.[4] This makes learning a native language quite different from some types of human learning, such as learning how to grow crops or do mathematics, but rather similar to others, such as learning to perceive scenes in terms of discrete objects with particular locations and depths. In contrast to visual learning, however, language learning is a species specific behavior. No other animal communication system even approaches human language in its complexity. Attempts to teach human language to other primates showed that these animals had little propensity to acquire the grammatical structure of human languages, despite intensive training

[2] "Subject-hood" is itself defined in terms of the position that the NP holds within the hierarchical structure.

[3] The 19th century assumption that non-Western languages are more grammatically primitive is long discredited. This is not to say that particular languages may lack particular grammatical devices. To take an extreme example, Pirahã, a language spoken by a tribe of around a hundred people in a remote area of the Amazon, has been reported (controversially — e.g. Nevins et al., 2009) to lack certain grammatical structures previously thought to be universal. Nevertheless, Everett points out that Pirahã employs a highly complex, intricate grammatical system: "No one should draw the conclusion from this paper that the Pirahã language is in any way 'primitive'. It has the most complex verbal morphology I am aware of and a strikingly complex prosodic system." (footnote in Everett, 2005).

[4] Languages may have different dialects, but there is internal agreement for speakers of that dialect.

regimes (Terrace et al., 1979; Seidenberg & Pettito, 1979). In contrast, there is evidence that children begin learning the patterns of their native language from the first months of life (Aslin et al., 1998) and spontaneously produce their own utterances from about one year of age. Strikingly, the latter has also been found to be true even for children who are not exposed to any language. This is seen in deaf children who do not have exposure to a signed language. Being deaf, they do not acquire spoken language, but instead create their own gestural communication systems, dubbed 'home-sign'. Although more simple than mature languages, these systems have been found to have several properties in common with other human languages (but lacking in other species' communication systems), including use of discrete symbols to indicate fixed meanings (i.e. words) and, as we shall see later, the use of certain grammatical devices (see Goldin-Meadow, 2003, for a review).

All of this indicates that children are born with a biological predisposition for language learning. But what is the nature of this predisposition? In the 1960s Chomsky famously proposed that it takes the form of an innate 'Universal Grammar' (henceforth UG), i.e. that children are born with innate knowledge about the possible grammatical organization and structure of human languages (Chomsky, 1965). This radical theory revolutionized the scientific study of language which, at the time, was primarily conducted according to the principles of Behaviorism, a paradigm which rejected a role for mental structures in psychological theory. Chomsky pointed out the inadequacy of this approach for understanding human language: any account of linguistic behaviour must allow for the mental structures which underlie the utterances we produce and understand. He also argued that the simple associative learning mechanisms of behaviourist learning theory were inadequate to account for the abstraction of the necessary linguistic structures. Thus innate UG was proposed to act as a 'blueprint' for acquisition. According to this account of learning, the child's task is not to create structure, but rather to identify which of a set of known structures match the sample of language she hears around her. The theory received apparent support from the fact that linguistic structures frequently recur across the languages of the world, even in 'unrelated' languages whose speakers have little or no contact (Greenberg, 1963). One explanation is that languages are constructed from a single grammatical template with parameters which can set differentially for different languages.

It would be hard to overstate the influence of the UG hypothesis in Linguistics and Cognitive Science: the existence of some form of UG became an underlying premise of the main stream *Generative Linguistics* paradigm in the 1960s, and remains so to this day (although it is explicitly rejected by other brands of Linguistics: Langlacker, 1987; Lakoff, 1987; Bybee, 1985). Nevertheless, the concept has been through many permutations over the years, even for researchers working within the Chomskian framework (for some current approaches see Chomsky, 1995; Hauser et al., 2002; Crain & Pietroski, 2006, and for a very different UG perspective, Jackendoff, 2002). Some researchers use the term UG more generally, to include whatever structures and processes, language specific or otherwise, we bring to the task of language learning (for example, see Goldin-Meadow, 2005). However the argument that all languages follow from, and are thus learnable from, an innate template of

specifically *grammatical* knowledge has become increasingly untenable. For example, one claim about UG (e.g. Pinker, 1984) has been that the categories which occur in the world's languages are drawn from a fixed set. However several researchers have argued that cross-linguistic evidence does not support this claim. Although certain categories can be identified across many languages (e.g. nouns and verbs, adjectives and prepositions/postpositions) this identification relies largely upon knowing the semantic properties of the words in the category. However, comparing *across* languages these classes may be syntactically quite different. For example, in some languages "verbs" (i.e. the class of words referring to actions) are marked for tense and action, but in other languages that property is associated with "nouns" (i.e. the class of words referring to things). In fact, Croft (2001) argues that categories across languages are so varied that they are essentially language specific (see also Evans & Levinson, 2009). An alternative explanation for the fact that linguistic categories which are very similar — both in terms of semantics and grammatical behavior — do frequently reoccur across languages is that they comprise a 'good solution' for building a communication system within the confines of human conceptual biases and broader cognition. They thus emerge in the process of language change (see Christiansen & Chater, 2008, for a general account of this type, and also Kirby and Oudeyer, this volume).

If children are not "pre-equipped" with grammatical knowledge, they must instead be endowed with learning mechanisms which abstract that information from their input. In recent years, many researchers have argued that this depends on a process of *statistical learning* (Elman, 1990; Newport & Aslin, 2000; Rumelhart & McClelland, 1986; Seidenberg, 1997), that is, an ability to identify reoccurring relationships between elements of the input, and make appropriate generalizations from probabilistic patterns. A growing body of evidence suggests that young children come to the task of learning with an ability to track probabilistic patterns. For example, Saffran et al. (1996) demonstrated that 8 month old infants are sensitive to syllable co-occurrence probabilities. Such information provides a useful cue for identifying word boundaries — an important part of acquisition since, in spoken language, unlike in written language, there are no 'gaps' between words. For example, in the sequence of syllables *pre-ty-ba-by* one cue to the fact that *pre* and *ty* form a 'unit', while *ty* and *ba* do not, is that across the whole language *pre* is followed *ty* about 80% of the time, but *ty* is followed by *ba* only about 0.03% of the time. Saffran et al.'s experiments demonstrated that 8-month-olds who were exposed to a stream of nonsense syllables could distinguish between those syllables which had frequently co-occurred in the string ("words") versus those which had infrequently co-occurred (i.e. "part-words" which crossed words boundaries)[5]. Computational

[5] A variety of techniques exist for assessing whether pre-verbal infants distinguish different types of stimuli. Saffran et al. used *preferential listening* where infants indicate their interest in some aural stimuli by looking at a light which they associate with that stimuli. Longer looking times are taken to indicate greater interest in the stimuli. Saffran et al. found that, after exposure to the nonsense syllable stream, infants showing longer looking times for *part-word* test items than for *word* test items (the stimuli were played repeatedly until the infant looked away from the light). The interpretation is that they found the part-words to be more *novel* and therefore more interesting.

work has revealed that very similar statistics are relevant to *grammatical* learning. For example, Mintz et al. (2002) conducted computational analyses over samples of speech (English) spoken to particular children (from the CHILDES database, see MacWhinney, 2000). Their analysis treated each utterance in the input set as a string of (meaningless) words and tracked how often particular words co-occurred adjacently. Importantly analyses were conducted over very large samples (15,000 – 20,000 utterances in each corpus[6]). Clustering techniques were then applied to this data and revealed that there was sufficient information to separate words into the English categories 'noun' and 'verb' with good accuracy. Finally, further evidence that language learning involves tracking co-occurrence statistics comes from the abundant evidence that such probabilistic knowledge plays a role in real time language understanding. For example, many studies have shown that when we encounter a verb we predict what type of construction is likely to follow that verb on the basis of our past experience. For example, English speakers expect the verb *find* to be used in a transitive construction with a direct object, which is the construction with which it is most likely to occur across the language. Our sensitivity to this probabilistic information shows up when we read a sentence in which this expectation is violated as in 3 where 'found' is followed by a sentence complement.

(3) *The chef found the recipe would require using fresh basil*

The reader's 'surprise' can be captured using various psycholinguistic techniques (such as monitoring hesitation in eye-movements at the word *would*). Importantly the same 'surprise' does *not* occur for verbs which are likely to be followed by a sentence complement (e.g. *claim*: Trueswell et al., 1993; see also Garnsey et al., 1997; Snedeker & Trueswell, 2004; Trueswell & Kim, 1998). The point is that if language *processing* relies on knowledge of statistical likelihoods, that same information must somehow be accumulated as a part of language *learning* (see also Wonnacott et al., 2008).

In the remainder of this chapter, I will consider the statistical learning hypothesis with respect to the acquisition of certain aspects of Morphology and Syntax. The aim is to illustrate domains in which a statistical learning approach has been applied and explore the strengths and weaknesses of current accounts. Two overarching themes emerge. The first is that both our knowledge of grammatical patterns, and the ways in which we use and process them reflect the probabilistic nature of the input to which we are exposed. The second is that a statistical account of language acquisition is far from a "blank slate" theory of learning.[7] In fact, any such account

[6] This under-estimates, rather than over-estimates, the quantity of language to which a child is likely to be exposed. Hart & Risely (1995) estimate that working class children hear an average of 6 million words per year.

[7] Both of these themes have been emphasized by other researchers. See Newport & Aslin (2000) for a statistical learning approach which strongly emphasizes the importance of innate constraints on learning. See Elman et al. (1996) for a connectionist approach to the issue of "innateness" in terms of the architectural make up of networks in different domains; See Seidenberg (1997) for a discussion of the relationship between statistical effects in language learning and language processing.

must specify both the sources of information that are accessible to the learner (i.e. the *primitives* to the learning system), and the ways in which these different sources are combined and evaluated to yield generalization.

2 Statistical Learning in Morphology

Words may have internal structure. That is, we can identify meaningful parts, known as *morphemes*, which re-occur across different words in systematic ways. In the introduction I used the example of the English *"ed"* past tense form to illustrate our implicit knowledge of morphology. Linguists have traditionally described such patterns using "rules" which capture the relationships between stem morphemes, inflectional (i.e. grammatical) morphemes and composed forms. For example, Figure 1 shows a formulation of the English past tense in terms of a conditionalized rule which takes the stem morpheme (*like*, *love*, *load* etc.) as 'input', concatenates it with an appropriate form of the inflectional morpheme (*/t/*, */d/* or */əd/*), in accordance with the phonology (i.e. sounds) of the stem, and provides the appropriate past tense form (*likeT*, *lovD*, *loadID* etc.) as 'output'. This rule not only provides a neat description of a widespread linguistic phenomenon (97% of English verbs form their past tense form in this manner), but also appears to capture actual human *behavior* — when given a new verb stem (e.g. *wug*, *wuf*, *wid*) as input, we can produce past tense output (*wugD*, *wufT*, *widID*) according to the pattern.

Similar "new words" tests will generally reveal a variety of such productive rules for a given language, not only for verbs but also for other parts of speech (e.g. plural marking and case marking on nouns, agreement features on adjectives, etc.[8]). From about 3 to 4 years of age, children have also been shown to productively inflect new words (Akhtar & Tomasello, 1997; Berko, 1958). This behavior indicates that the child has abstracted these regular morphological patterns from her experience of words in the input, and has somehow incorporated this knowledge as a productive part of the mental grammar. The question for theories of acquisition is *how* such learning occurs. One apparently straightforward account could be that that the child comes equipped with a learning algorithm which explicitly seeks out linguistic rules like those in Figure 1. This process must involve both identifying the various reoccurring morphemes across a language, and extrapolating and mentally representing the rules which combine these morphemes into complex forms. According to this story, a speaker has no need to store forms like *"walkT"* and *"likT"* in memory, only morphemes like *walk*, *like* and *T*, since once the rule has been acquired com-

[8] Languages may make much more extensive use of productive morphology than English. For example, in many languages (e.g. many of the Eskimoan languages) entire nouns may be attached to the verb-stem as dependent morphemes, rather than appearing as separate words within the sentence (a phenomenon known as "noun incorporation").

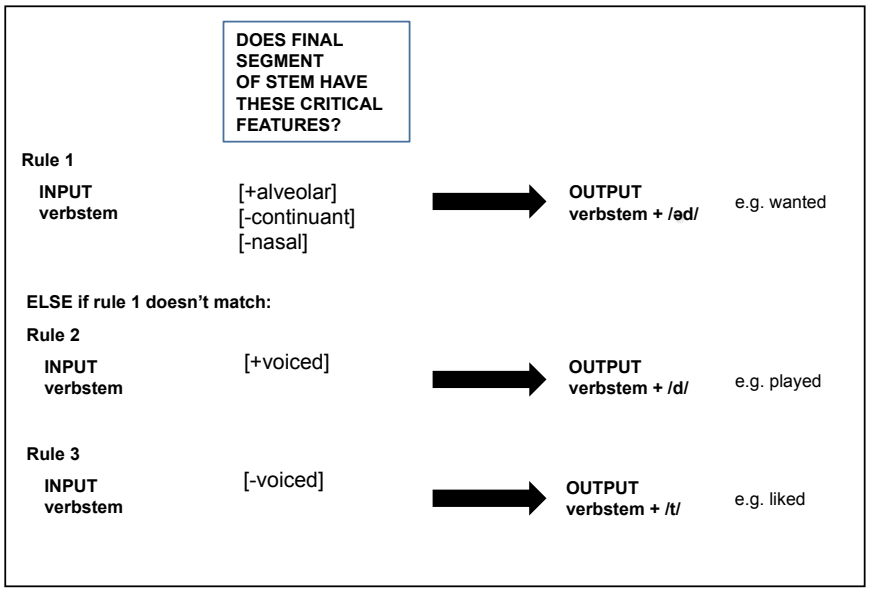

Fig. 1 Rule for forming the English regular past tense, taking the stem as input. The properties in square brackets are distinctive features (roughly following Chomsky & Halle, 1968) which pick out a set of speech sounds. [+alveolar] means "produced via contact between the tongue and the alveolar ridge", [-nasal] means "air escapes from the mouth" (and not the nose, which is the case for e.g. /n/) and [-continuant] means "production involves a complete closure completely blocking airflow". This combination picks out the alveolar oral stops (i.e. /t/ and /d/). [+voiced] means "produced with the vocal chord vibrating", and picks out all the voiced speech sounds (i.e. vowels and consonants such as /g/ and /v/). [-voiced] picks out the non-voiced consonants such as /k/ and /f/. The sub-rules must be ordered so that rules 2 and 3 only apply if the word does not end in /t/ or /d/.

plex forms can be routinely composed and decomposed 'on-line" during language production and comprehension (e.g. Prasada & Pinker, 1993)[9].

An obvious problem with the theory as described above is that there is no account of how we learn exceptional or *irregular* forms such as the past tense forms *went* and *ate*. Such forms appear to have to be stored as wholes, with some additional mechanism which 'blocks' the application of morphological rules where there is a stored exception (e.g. the stored form *went* blocks the formation of *goed*: Marcus et al., 1995). However a little further probing reveals that a system of wholly productive rules and unproductive exceptions is insufficient. For example, try applying the new words test to the stem *ping*. You may come up with *pingD*, in line with Figure 1, but you might also hit on *pang* (Bybee & Moder, 1983; Prasada & Pinker, 1993). This behavior is clearly related to the existence of verbs like *sing*, *ring*, *spring* with their past tense forms *sang*, *rang*, *sprang*. The critical point is that the underlying

[9] Some more recent versions of this theory allow that at least some regular forms also be stored as whole forms (Pinker, 1999; Pinker & Ullman, 2002).

pattern appears again to have some productivity, i.e. English speakers can access some generalized process which converts *ing* → *ang*. In fact, for the English past tense there are a number of such semi-productive patterns, so we might also get past tense forms *med* for the stem *meed* (in line with the *eed* → *ed* pattern in *bleed/bled*, *feed/fed* etc) and *prew* for *prow*, (in line with the *ow*→ *ew* pattern in *blow/blew*, *grow/grew* etc). Although some early theories (e.g. Chomsky & Halle, 1968; Halle & Mohanan, 1985) attempted to capture these types of regularities in terms of rules also (for example rules which altered the vowel in the stem in particular contexts), it turns out to be very difficult to identify a precise set of phonological patterns which trigger particular past tense forms (see Bybee & Slobin, 1982; Pinker, 1999).

More fundamentally, characterizing these semi-productive regularities in terms of clear cut rules ignores an important finding: the extent to which the patterns are extended to new words depends on our experience of how consistently they apply across the language. For example, the probability of producing *pang* as the past tense of *ping* will be affected by the number of verb stems which are phonologically similar to *ping* and have past tense forms similar to *pang*. This statistic is known as 'type frequency'. Exactly how similar the new verb is to familiar verbs which use the pattern is also important. The effects of these statistics may also be seen in the patterns of errors made by young children. The majority of morphological errors are so called overgeneralization errors which arise from over-applying the regular pattern (e.g. saying *goed*, *gived* etc), but other patterns may also be over-generalized. For example, incorrect forms such as *brung* and *brang* (Xu & Pinker, 1995; Bybee & Slobin, 1982) may result from relatively high frequency of the relevant patterns across past tense forms in the language. The frequency of individual verbs ('token' frequency) is also important — children are less likely to use the wrong pattern with a more frequent word (e.g. *sleeped* is a less common error than *weeped*).

Generalization errors are not restricted to child language but are also made by adults, particularly the overgeneralization of high frequency patterns to low frequency words (these are also the items that are most likely to change their morphological behavior over *generations* of speakers and understanding the interaction of type and token frequency is critical to understanding the process of language change: Bybee, 1985). In fact, such similarity-based *graded* productivity turns out to be rife in natural language morphology (see Bybee, 1995; Hay & Baayen, 2005). From the perspective of language acquisition, our theories must therefore include an explanation of how learners come to track different statistics, such as type and token frequencies and how these become integrated into the productive morphological system which emerges. One class of statistical learning system which has been extensively studied in this domain are *connectionist models*, also known as *neural networks* since their architecture is inspired by the fact that neural circuitry is comprised of networks of interconnected units (neurons) which learn by adjusting the connections between those units (synapses)[10]. These models are able to extract probabilistic patterns in the course of learning mappings between sets of input and output nodes. Such a model was first applied to morphological learning by

[10] Although connectionist models are neutrally inspired, there is no claim that they constitute a biologically plausible model of neural circuitry.

Rumelhart & McClelland (1986) who presented a landmark model of the acquisition of English past tense. This model (and many subsequent models, e.g. Plunkett & Marchman, 1991; Hare et al., 1995; Daugherty & Seidenberg, 1992)[11] learned to map a set of input nodes representing the sound patterns in the verb-stem to a set of output nodes representing the related past tense form. Different input/output nodes represent different aspects of the phonology of the stem or past tense form, meaning that representations were *distributed* (for example, the representations of *sing* and *ring* would have shared components, i.e. some subset of nodes would be activated for both). These models can be "trained" to map a set of stems to past tense forms (e.g. given *sing* they generate *sang*), importantly, without having any explicit rule formation process — after sufficient training the models may also generalise appropriately to new words, i.e. i.e. given *wug* generate *wugD*, given *ping* generate *pingD* or possibly *pang*. The models also make errors, particularly before they are fully trained, and, as for human children and adults, errors are affected by token and type frequency statistics. In the models, this is a direct consequence of the statistical nature of the learning process: as particular words are frequently encountered, the mappings between the stem phonology and past tense phonology are proportionally strengthened (token frequency) but, since words have *distributed representations*, aspects of those mappings which frequently re-occur across words are also strengthened (type frequency). Errors are therefore likely if a verb is low frequency but its stem is consistent with an alternative high-frequency pattern. These models thus capture the probabilistic effects of *phonological similarity*.

Some more recent connectionist models have reconstrued the learning problem so that rather than mapping directly between different phonological forms, the network's central 'task' is to map *phonological* and *semantic* representations, i.e. the sounds of words to their meanings (e.g. Joanisse & Seidenberg, 1999; Plaut & Gonnerman, 2000). So, for example, for past tense the phonological string *walkT* might map to a semantic representation including the information *WALK-PAST-THIRDPERSON* and the form *walkS* to a semantic representation including *WALK-PRESENT-THIRDPERSON*. Links between different forms of the same verb can thus emerge from shared semantic and shared phonological representations. Models of this ilk also have the potential to capture effects of *semantic* similarity (i.e. when clusters of words with similar meanings show similar morphological behavior, as is quite common in morphological systems across the world's languages: Wierzbicka, 1988). Moreover, mapping the sounds of words to their meanings provides a more natural model of the child's actual task during language acquisition.[12] Note that in this view "morphemes" such as */s/* = *THIRD-PERSON-PRESENT* are *emergent* rep-

[11] Later models had more complex architectures, including layers of hidden units between the input and output units, and used different learning algorithms.

[12] All connectionist models require an error signal to drive learning. The models learn by predicting outputs for given inputs (early on predictions are random guesses), receiving feedback as to what the correct response should be, and then updating the "weights" (which drive the predictions) accordingly. For models which map between phonology and semantics, we are envisioning a child who implicitly compares the sound she would have expected for a given meaning with the one she is hearing, and the meaning she would have expected for a given form with the one that is currently implied.

resentations which arise when pieces of form and meaning are *repeatedly* associated (e.g. Gonnerman et al., 2007).

In short, connectionist models have proved useful in elucidating the origins of graded productivity and the probabilistic usage of morphological patterns, for example, explaining effects of token and type frequency. However whether this type of account is appropriate remains highly controversial. In particular, there is concern that the generalizations which networks acquire only approximate bona fide "rules". One theory claims that these are necessary to account for *regular* morphological processes, although generalization seen with irregulars may be accounted for by the storage of those forms in a connectionist-style associative memory system (a so-called 'dual-route' account, Prasada & Pinker, 1993; Pinker & Ullman, 2002). Debate has therefore centered around whether regular and irregular forms show differences in processing. For example, some studies found effects of token frequency (e.g. Prasada et al., 1990) and phonological similarity (e.g. Prasada & Pinker, 1993) for irregulars but *not* for regulars. However both of these effects have since been found for regulars (token frequency: Schreuder et al., 1999; Hare et al., 2001; phonological similarity: Albright & Hayes, 2003). From a statistical learning perspective, the fact that graded, statistical effects are harder to detect with regulars, so that the patterns therefore appear more 'rule-like' in their application, results from the statistical properties of the input. Regular patterns generally have a much higher type frequency than the alternatives,[13] resulting in a strong drive to apply those patterns across the board. This tends to overwhelm any factors concerning the properties of particular words.

More recently, arguments have focused on *neurological* evidence which suggests that producing and comprehending regular and irregular forms involves different brain areas. This comes both from brain imaging studies (e.g. participants hear/read a word and we see which brain areas are activated; e.g. Joanisse & Seidenberg, 2005) and from studies of individuals who suffer damage to different areas of the brain (i.e. damage to one area of the brain may affect the production and comprehension of regulars, damage to another the processing of irregulars; Marslen-Wilson & Tyler, 2007). However, the interpretation of these differences remains controversial. One explanation is that they actually arise from differences in the extent to which producing and comprehending regular versus irregular forms relies on *semantic* and *phonological* representations. For example, the task of producing the idiosyncratic *took* when presented with *take* strongly relies on identifying the particular word. Thus semantic representations play a role (Joanisse & Seidenberg, 1999). Also, regular forms tend to be more phonologically complex than irregulars (for example they are more likely to end in consonant clusters as in the /spt/ in *claspT*: McClelland &

[13] Some researchers have argued that the most frequent form is not always the one that acts as the regular rule (e.g. Marcus et al., 1995). However in such cases the *variety* of types may be important. Plunkett & Nakisa (1997) demonstrate that a pattern which is not the most frequent can become the most productive in a connectionist model provided that the set of words to which the pattern applies are more dissimilar to each other than is the case for the sets of words associated with alternative patterns. Capturing such variability relies on the use of models with a more complex architecture, including a layer of hidden units between input and output mappings.

Patterson, 2003). Joanisse & Seidenberg (1999) showed that when a connectionist model with separate banks of semantic and phonological units had been trained up on the English past tense, the production of *irregulars* and *regulars* could be differentially affected by knocking out semantic and phonological areas respectively.

An alternative account of the neurological evidence is that producing and comprehending regularly inflected words does actually involve the assembly/decomposition of complex words from/to their component morphemes. Some direct evidence for decomposition has been presented for comprehension. Post et al. (2008) argue that any word, including a new word, which potentially matches the output of the schema in Figure 1, may be automatically decomposed. For example, the word *nomd* (presented aurally) is a potential past tense which can be formed from *nom+/d/* (try reading 'nommed' aloud). Evidence that such forms are actually decomposed comes from an experiment in which listeners had to say if pairs of words were the "same" or "different". Participants took longer to say that pairs like *nomd–nom* were different than pairs like *nomt–nom*. This is interpreted as evidence that *nomd* is decomposed into *nom+/d/* (note that *nomt* does not fit with the schema in Figure 1 since /t/ should follow a voiceless consonant and /m/ is voiced).[14] Thus, at least for comprehension, there is some evidence for the storage of separable morphemes (like /t/, /d/ and /əd/) and for the usage of "rules" to access these morphemes where appropriate. It remains to be seen whether connectionist-style models where rules and morphemes are *emergent* forms, with *graded* representations, can capture this type of data.

In summary, the statistical make up of the input language has important consequences for the emergent morphological system, and probabilistic patterns may be seen even for very rule-like systems. Any model of morphological learning must account for this, as well as accounting for situations in which processing is indeed very rule-like. Connectionist models have been important in elucidating the origins of statistical phenomena, but it remains to be seen whether they can account for the full range of behavioral evidence, or whether statistical learning systems with different architecture assumptions are necessary. One piece of evidence suggests there may be some further constraints or biases in the statistical learning system. This comes from the study of a child whose language input contained *inconsistent* patterns of morphological usage (Singleton & Newport, 2004). This child was deaf and his only input was the sign language used by his parents who were imperfect users of that language, having themselves not been exposed to the language until adulthood (this is typical of so called late-learners of a language; Newport, 1990). Surprisingly, the child's own language in many ways surpassed that of his parents. Most relevant here is that there were situations where the child's parents erratically used multiple complex morphological forms (a little like randomly using all three of, say, *sleepT*, *sleepD* and *slept*) but the child himself did not replicate this probabilistic usage and instead boosted the frequency of the most frequent form and

[14] The critical factor appears to be whether past-tense forms are potentially decomposable, rather than whether the relationship between stem and past tense is regular. For example, *slept* is traditionally irregular but is nevertheless decomposable into *slep* + /t/ (note that this fits Figure 1 as /p/ is voiceless) and it seems to be processed akin to regulars rather than irregulars (Joanisse & Seidenberg, 2005).

eliminated the others. Thus the child did not replicate the probabilistic patterns of the input, but in a sense 'sharpened' these patterns to make the system more consistent and rule-like (see also Newport & Aslin, 2000). To my knowledge, this type of language change within a single learner has not yet been addressed within the connectionist literature.

Whatever the adequacy of statistical models of morphology, one final point is worth emphasizing. As I said in the introduction, statistical models are far from blank slate learners. This is particularly clear when considering working computational models. Models instantiate hypotheses, not only as to the architecture of the learning system, but also as to the information available to that system. For example, models of English past tense may assume features such as 'voiced', 'voiceless' and 'alveolar' as primitives in the phonological representations of words. If the model maps phonology to semantics it will have (at least) PRESENT and PAST as primitives. This does not preclude an account in which these features are themselves learned (though this of course opens the question of how *that* learning occurs), but where the behavior of the model depends on a particular set of primitives this makes the strong claim that by the time that morphological learning occurs, such features are available as candidates for mapping.

3 Statistical Learning of Syntax

While *morphology* governs how words are formed from smaller meaningful parts (morphemes), *syntax* is the system which governs how those words combine to form phrases and sentences. It is syntax which provides the massive productivity and expressivity of human language. Following early (and extremely influential) arguments made in Chomsky (1957), many researchers rejected the notion that statistical learning mechanisms could appropriately abstract syntactic knowledge. For example, Pinker (1987) argued that a learning procedure which simply attended to how words are distributed within sentences could easily be led astray. As an example, he suggested that a distributional analysis of the sentences in 4a–c could lead to the incorrect generalization in 4d:

(4) a. *John ate fish.*
b. *John ate rabbits.*
c. *John can fish.*
d. **John can rabbits.*

More recently however, access to fast computing has allowed researchers to explore how a distributional learner would fare if given access to very large amounts of linguistic input. We saw in the introduction that distributional computational analyses which cluster words on the basis of adjacent co-occurrence statistics can distinguish English "nouns" and "verbs" with good accuracy, provided they are applied to a sizable corpus of sentences, as opposed to just three utterances (Mintz et al., 2002).

The potential error in 4d — which is the result of a mis-categorizing the word *rabbits* — is avoided because words like *fish* get clustered with both nouns and verbs and words like *rabbit* do not, since *rabbit* shares many more distributional characteristics with words used primarily as nouns (Mintz et al., 2002). Further research has shown that distributional statistics can divide words into a more comprehensive set of categories, and these correspond fairly well to the types of syntactic categories identified by linguists (adjectives, prepositions etc.; Mintz, 2003; see also Finch & Chater, 1994). An inherent advantage of a statistical approach is that it has the potential to capture the situation where category membership is graded rather than absolute, and where words appear to act like partial members of more than one category (e.g. the English words word *near* appears to be a blend between an adjective and a preposition, Manning & Schutze, 2001). Related statistical analyses may also capture some information about permissible and impermissible sequences of categories (Elman, 1990; Church, 1988).

Such computational research has played an important role in demonstrating that the input holds a good deal of information for a learner trying to build a syntactic system, providing that that learner is equipped with mechanisms which can tap into sequential patterns. On the other hand, we know that human syntactic knowledge cannot be captured by a grammar which generates unstructured sequences of categories. How far can we take a statistical approach to syntax learning? In the remainder of this section I consider this problem with respect to three topics: linking syntactic structure to abstract semantics, avoiding overgeneralization and acquiring hierarchical phrase structure.

3.1 Linking formal structure and meaning

The types of statistical analyses discussed may yield a formal system for generating possible word strings, but the strings themselves convey no further information. In contrast, the *raison d'etre* of natural language syntax is to provide a means of systematically encoding a structured *message*. For example, we have seen that the *X verb Y Z* structure (as in *Jack threw Henry the ball*), conveys a *transfer* event and further indicates the roles which the entities denoted by the noun phrases *X*, *Y* and *Z* play in that event (the so called "thematic" roles which linguists label *agent*, *recipient* and *theme*). This is an example of a *construction* i.e. a systematic mapping between a formal pattern (here the positions that words and phrases can occupy within the utterance) and a semantic pattern. Many researchers have focused on the acquisition of constructions, and particularly constructions centered around verbs (verb-argument structure constructions), as a core component of syntax acquisition (e.g. Tomasello, 2000; Gleitman et al., 2005).

As always for theories of language acquisition, theoretical debate concerning the learning of constructions has focused on whether the necessary structures can be gleaned via exposure to the input, or whether children come equipped with relevant innate knowledge. For example, Tomasello (2000) argues against a UG approach

on the basis of evidence that young children's grammars (before they reach about 3 years) are not 'adult-like'. In particular, he claims that for verb argument structure constructions such as the *X verb Y Z* structure, there is no evidence that children know the link between the formal structures and abstract thematic relations like *agent* and *theme*. This is because, unlike adults and older children, young children are unable to use the structure with new verbs, and their usage of the structures in everyday speech is generally limited to one or two verbs. On the basis of such data, Tomasello (2000) proposed the "verb island hypothesis": early on children use structures which revolve around specific verbs, such as *"X want Y"* where *X = person-that-wants*, *Y = thing-wanted*. According to this theory, abstract constructions only emerge once children have acquired multiple related verb-specific structures and notice the relationships across them. Interestingly, however, experiments using preferential looking[15] have revealed that children may have some knowledge of the relationship between word order and abstract thematic roles at a much earlier age than they are able to demonstrate in production. For example, one study found that 21-month-olds who heard a transitive sentence containing a new verb (such as *Rabbit is blicking Monkey*) whilst viewing two scenes with a novel action — one with correct noun assignments, one with the roles reversed (e.g. correct: *RABBIT–JUMPS-ON–MONKEY*, reversed: *MONKEY–JUMPS-ON–RABBIT*) — tended to look longer at the correct scene. Some researchers have argued that this early evidence of abstract knowledge indicates that learning is not entirely input-driven and that the child "contributes" some structure to the learning process (Gertner et al., 2006; see also Fisher, 2002).

What is the role of statistical learning in these accounts? In fact, the verb-island hypothesis relies on statistical learning mechanisms: the child must have an ability to form generalizations once a 'critical mass' of related structures has been accumulated, just as we saw that repeated patterns could lead to generalization in morphology. Here, however, the ability to generalize also relies on the child's ability to notice (subconsciously of course) the abstract semantic relationships which hold across sentences. That is, she must be able to identify that in *Henry kisses mummy* and *Poppy drinks milk* the roles played by *Henry–Poppy* and *mummy–milk* in the kissing and drinking events are analogous. One way that a statistical learning model might capture this type of learning is to include semantic representations alongside 'word string' representations of input sentences and some models have taken this approach (St. John & McClelland, 1990; Miikkulainen, 1996; Chang et al., 2006). One such model, presented by Chang et al. (2006), not only proved able to learn abstract constructions, but also captured some of the developmental data discussed above. The model included an SRN (Serial Recurrent Network) which is a type of connectionist statistical learning system which learns by sequentially predicting upcoming words and learning from incorrect predictions. This type of model has been shown to be able to abstract grammatically relevant information from word sequences (Elman, 1990). Critically, in the Chang et al. model, each sentence was also coupled with a structured semantic representation including — amongst other

[15] This is a standard methodology for assessing infant preferences for a particular visual stimuli.

things — thematic roles like *agent* and *theme*. This semantic information also fed into the prediction process. In line with the developmental data discussed above, the model showed evidence of verb-island effects early in learning. Specifically, it showed different degrees of accuracy when using the same structure with different verbs, and its ability to produce sentences with new verbs only gradually developed. However, results from preferential looking experiments were also replicated: given a sentence containing a new verb, the model could identify which of two semantic representations was correct *before* it would be able to correctly produce that sentence itself.[16] Eventually, like older children and adults, the model also passed the new verbs test in production, indicating that abstract structures had been learned.

The Chang et al. model provides a good illustration of how an input-driven, statistical learning explanation may still embody strong claims about what is 'built in' to the learner. On the one hand, the model is able to acquire abstract syntactic representations without access to innate *syntactic* knowledge of the type envisioned by some UG accounts (e.g. Pinker, 1989; Radford, 1988). On the other hand, it does assume that the learner has access to abstract *semantic* structures, which in the model are given rather than learned. In this way it is in line with some approaches which emphasize the structure innately contributed by the learner (Gertner et al., 2006; Goldin-Meadow, 2003). In particular, the model comes pre-equipped with abstract representations like *agent*, *recipient* and *theme*. Its behaviour thus demonstrates that input-driven 'verb-island' effects still arise, given the task of matching such representations up to particular words in the input.

Assuming that thematic roles are innate is a strong hypothesis, but it is supported by some independent evidence. This comes from the study of *home-sign* systems — the self-created language systems of deaf children, to which I eluded in the introduction. These language systems — which are formed without linguistic input — have repeatedly been found to encode a fixed sets of abstract thematic roles (including *agent*, *recipient* and *theme*, Goldin-Meadow, 2003). This is consistent with a hypothesis in which such representations are inbuilt. It also demonstrates an unlearned bias to desire to communicate *this specific type of information* — a bias also inherent in the Chang et al. (2006) model.

In other ways, the Chang et al model is relatively unconstrained. For example, the link between word order and thematic role emerges during learning, given the serial processing nature of the SRN architecture and the semantic representations. This contrasts with approaches which assume innate links between thematic role and word-order (and also between thematic-role and morphological case-marking — the other cross-linguistically common device; Pinker, 1984; Jackendoff, 2002). It has been shown that the model can equally learn languages in which the same information is marked via morphology or some combination of morphology and word order (Chang, 2009).[17]

---

[16] The data from the Gertner et al. (2006) experiments were not specifically modelled in Chang et al. (2006) but the result is generally consistent with the model's account.

[17] Ultimately we need an account of language learning and language change which explains why word order and case marking are so prevalent as means of encoding thematic information. However from the perspective of learning, the account must also be sufficiently flexible to explain the

In short, there is evidence that a statistical learning system can acquire essential links between *syntax* and *semantics*, provided it has access to structured semantic representations over which it can generalize. Of course current models are far from acquiring the full range of constructions for any language. Nevertheless, this type of modeling work is likely to play a central role in future research into syntactic learning.

3.2 Avoiding Overgeneralization

A classic criticism of input-driven theories of acquisition is that unconstrained learning may lead to an over-generalized grammar (e.g. Baker, 1979; Pinker, 1989). This can again be illustrated with respect to verb-argument structures. We saw in the introduction that not all combinations of verbs and argument structures are grammatical in the adult grammar, even where that combination would seem semantically plausible (sentences 1f and 1g above). However, once they are able to generalize constructions to new verbs, children may start to spontaneously use known verbs in constructions in which they have not encountered them. This may result in overgeneralization as in 5a and 5b (from Gropen et al., 1989):

(5) a. *Carry me the ball.*
b. *Don't say me that!*

The theoretical problem is how the child eventually learns that such combinations of verbs and structures are incorrect, given that they cannot rely on overt correction from caregivers.[18] In other words, if children are able to generalize verbs to new structures, but they don't get corrected when they use them with incorrect structures, how do they eventually learn that this generalization is actually ungrammatical?

This is the classic problem of *no negative evidence* and it applies whenever there is a plausible but ultimately incorrect linguistic generalization. One possible solution to the paradox is that apparent "exceptions" to generalizations are not arbitrary but are in fact conditioned. Overgeneralization will therefore cease once the child has identified the correct conditioning factors, perhaps with the help of innate knowledge of what such factors might be. For example Pinker (1989) proposes that the argument structures of verbs is in fact conditioned by subtle semantic factors which are not apparent at first glance. However attempts to come up with sets of ab-

---

learning of other additional or alternative devices. For example, sign languages may also employ the modality specific device of directing signs with the signing space (e.g. moving a GIVE gesture towards a particular person to indicate that they are the recipient).

[18] Since I have found that people outside of this discipline (particularly middle class academically minded parents, accustomed to explicitly correcting their children's grammar), have difficulty accepting this point, it is worth highlighting. To further see that parental correction does not account for our knowledge of verb syntax, consider that many of the verbs which are ungrammatical in this construction are Latinate verbs (e.g. *donate*). It seems unlikely that such verbs are widely used (and therefore corrected) in childhood, yet we all known their syntactic restrictions.

solute conditions have generally been found to be unsuccessful (Bowerman, 1988; Braine & Brooks, 1995; Goldberg, 1995) and arbitrary exceptions remain.

Although the problem of no-negative evidence is often presented as evidence against input-driven accounts, many researchers have argued that the solution may lie in the statistical nature of language learning and usage. The first step is to relax the criteria on what is learned. If the end of state of learning is a grammar which can determine "grammaticality" with an absolute yes-no judgment, there is indeed a learnability paradox. If instead we permit a grammar which allows *varying degrees of certainty*, "grammaticality" may be determined via probabilistic inference. In fact, at least for verb-argument structure constraints, there is evidence that judgments are graded in just this way. For example, Theakston (2004) asked both children and adults to rate "ungrammatical" sentences in which verbs occurred in the wrong structures. She found that such sentences received higher ratings when the verbs were of low frequency. For example, children judged *"He arrived me to school"* to be better than *"He came me to school"* (*come* occurs with higher frequency than *arrive*). Even adults, who of course have more familiarity with all verbs than children, nevertheless gave higher ratings when the verbs were very low frequency (for example preferring *"He vanished the rabbit"* to *"He disappeared the rabbit"*). In other words, the more a particular verb has been encountered in a particular set of structures, the less likely speakers are to extend that verb to a new structure (Braine & Brooks, 1995). As we saw in the introduction, the idea that we track how often different verbs occur in different structures is further supported by data from sentence processing.

From this statistical perspective, determining 'grammaticality' is a question of weighing up the evidence in the input. It is logically true that a child can never know that the verb *come* may not one day show up in a transitive structure — but their wide experience of that verb appearing in other structures *but not the transitive* can make them pretty certain. With less frequent, or entirely novel, verbs, it makes sense to assume that more general patterns may apply. Note that this tendency for more overgeneralization with low frequency items is exactly what we saw with morphology, pointing to common statistical inference processes. Partial conditioning factors such as verb semantics can now be considered cues which play a role in this statistical inference. For example, if other verbs similar to verb X occur in structure Y, that provides evidence that verb X may also do so. There is evidence that children and adults are influenced by both semantic and phonological similarity in just this way (Braine & Brooks, 1995; Ambridge et al., 2008; Brooks & Tomasello, 1999; Gropen et al., 1989). Finally statistical patterns at a 'higher level' may also play a role. If there is evidence that a construction is very "open", learners are more likely to generalize using that construction than if there is evidence that the construction's usage is restricted to particular words (Goldberg, 2005; Wonnacott et al., 2008).

The picture that emerges is one in which multiple sources of information can influence a judgment of grammaticality. The problem of how to evaluate and combine probabilistic cues is in fact well known in cognitive science more generally. For example, it is seen in the problem of combining visual cues to give percepts of depth and localization. Recent approaches to cognition have emphasized the use

of *Bayesian statistical inference* to estimate cue reliability from correlations in the input (Jacobs, 2002; Chater et al., 2006). This type of statistical inference is to some extent implicit in the connectionist approach discussed previously, however Bayesian models differ in making the formation and evaluation of hypotheses explicit. Both connectionist and Bayesian approaches have been applied to the problem of constraining overgeneralization (Connectionist: Allen & Seidenberg, 1999; Chang, 2002; Bayesian: Perfors et al., 2010; Dowman, 2000; Onnis et al., 2002)

An important question for future research is whether there are further constraints on the process of restricting generalization. For example, if a verb does not appear in structure X, does its frequent appearance in *any* other structure count as evidence against its future appearance in X? Some researchers have argued not (e.g. Goldberg, 2005). Identifying such constraints may be important in understanding how native language learners end up with a set of grammatical intuitions that are so similar.

3.3 Hierarchical Phrase Structure

Syntactic analyses of very different and unrelated languages have repeated revealed that sentences are composed from phrases, which may themselves be composed from smaller phrases, and so on. This is *hierarchical phrase structure*. We saw examples in the introduction with the two structures underlying the ambiguous *"Put the box on the block on the table"*. Any theory of language acquisition must account for how children are able to acquire grammars which generate these types of structures. Within the Chomskyan tradition, this principle of syntactic organization constitutes part of Universal Grammar. In other words, children are supposed to come to acquisition assuming that utterances are composed of phrases and expecting syntactic relationships to operate over phrases rather than single word categories (the principle of *structure dependence,* Chomsky, 1968). However this approach has assumed that constituency is universal, and this is controversial. Evans & Levinson (2009) argue that there are many languages which show no evidence of constituency since they have "free" word order and words which are semantically grouped are not necessarily contiguous within a sentence. It is interesting that these languages do nevertheless have a means of encoding a hierarchical message: elements of distinct levels of structure may be grouped using multiple levels of morphological case marking (i.e. word endings). This suggests the possibility that it may be the structured nature of conceptual representations which is "universal" to human language[19], rather than a particular means of encoding that information.

Nevertheless, the ability to learn hierarchical phrasal structures poses an important challenge for statistical models of acquisition (Chomsky, 1957). Within the connectionist tradition there has been an attempt to demonstrate that models can capture behaviours which appear to rely on phrasal constituency. For example, the ability to

[19] Although Everett (2005) controversially claims that Pirahã lacks the ability to encode *recursion*, a particular type of hierarchical structure whereby the same phrase may be embedded within a phrase of the same type.

compose indefinitely long complex noun phrases means that agreement relations may hold over several words — as in the subject-verb agreement in 6 (note that only the highest levels of phrasal structure are shown):

(6) [*The* **boy** *[who chases dogs which chase cats]]*$_{\text{NP}}$ **runs** *fast*

Elman (1993) probed whether an SRN was sensitive to such long distance dependencies. As discussed above, SRNs learn by predicting upcoming words in a sentence (input units represent the current word, output units represent the next word — and the difference between this prediction and what the next word turns out to be provides the error signal which drives learning). SRNs also have *hidden units* between the input and output which feed into the prediction, and, critically, a set of 'context' units, which carry a copy of the previous state of the hidden units. Since these serve as additional inputs to the hidden units, the current activation of these units is affected by both their current and previous activation, which was in turn affected by the previous activation, and so forth. Thus, although the predictions of SRN models are most strongly dependent upon the previous word, there is also a rapidly diminishing memory for the earlier sentence context. Elman showed that an SRN which was trained on sentences from a pseudo-English grammar learned to reject sentences like **The* **boys** *who chase dogs which chase cats* **runs** *fast*, demonstrating that it was able to hold information about agreement over long distances (in fact the network only succeeded when it was first trained on simple sentences such as *The boy runs* and *The boys run*; however, this is controversial since it was not replicated in a later study, Rohde & Plaut, 2003). Another study (Lewis & Elman, 2001) showed that an SRN could learn that question forms such as 7a were acceptable whilst forms such as 7b were not.

(7) a. Is the man who is coming here?
b. * Is the man who coming is here?

This ability appears to rely on an ability to recognize the noun phrase in 7c:

(7) c. [The man who is coming]$_{\text{NP}}$ is here.

Does the SRN succeed in these tasks by learning something about hierarchical structure? It is certainly clear that the type of structure that is acquired is not equivalent to that which can be implemented in a symbolic processor. For example, a symbolic system has no limits on the depth of embedding which it can process. In contrast, processing in the SRN may rapidly breakdown, particularly given a certain type of embedding known as "center embedding" (Christiansen & Chater, 1999). However, this is not necessarily a shortcoming of the models, since human processing may also break down in these circumstances (try figuring out *"A man that a woman that a child that a cat that I heard saw knows loves beer"*.)

Nevertheless, it is not clear that SRNs do extract phrasal structure. Steedman (2002) argues that the SRN models approximate the class of Finite State Markov Machines. This means that they treat essentially represent sentences as an unstructured string of categories. The form of a word may thus depend upon the previous set

of words up to some length (a so called *"n-gram"* — although unlike fixed n-gram models, SRNs can potentially learn what length n-gram is most relevant). Such a system may prove able to track fairly long distance relationships, but will never be able to represent structures of the type necessary to disambiguate such sentences as *Put the box on the table on the shelf*. In addition, the interpretation that the network presented by Lewis & Elman had learned something about an NP constituent is challenged by recent work showing that the relevant sentences can be differentiated by a learner sensitive only to relationships between adjacent words (Reali & Christiansen, 2005). This same statistic *cannot* deal with equivalent question formation in other languages (Kam et al., 2008) or with a variety of other linguistic phenomena which rely on internal sentence structure.

One obvious limitation of the studies discussed above is that the models were asked to learn syntactic patterns without access to semantic structure. Yet phrase structure is a means of representing conceptual groupings — for example an entire noun phrase serves to picks out a particular entity or set of entities. Still it is interesting that connectionist models which do attempt to link syntactic and semantic structure have tended to employ additional specialized mechanisms for dealing with the encapsulated interpretation of embedded structures (Miikkulainen, 1996; Pollack, 1988; though see Bryant & Miikkulainen, 2001). These systems still differ from symbolic systems in showing plausible memory degradation for center embedding, as discussed above. However to date such work has only dealt with fairly basic linguistic phenomena. It remains to be seen how statistical approaches will scale up to deal with the full complexity of natural language syntax, and the types of learning architectures necessary to capture these behaviors.

4 Concluding Remarks

In this chapter, I have presented evidence that statistical learning processes play an important role in language acquisition. We have seen that statistical models are necessary to explain graded, probabilistic effects in morphological systems (even those that appear very rule-like), although it is currently unclear whether and how current models will scale up to capture all of the human data. We also saw evidence that a statistical system can learn abstract relationships between syntactic form and semantic structure (at least if given access to the requisite semantic representations). Further, an ability to track and evaluate probabilistic evidence may explain how learners avoid rampant overgeneralization and converge on highly similar grammatical intuitions. However, it is important to emphasize that we are still a long way from possessing a full account of statistical grammar learning. In particular, accounts of many 'higher level' syntactic behaviors are lacking, particularly those which require access to hierarchical structure.

Another phenomenon which statistical learning theory must address, and one which I have neglected in this chapter, is the fact that acquisition is generally more successful when it begins in early childhood. This has been shown to be the case

even when controlling for years of exposure and external factors such as 'motivation' (Johnson & Newport, 1989; Newport, 1990). These studies reveal that although the ability to learn language is not entirely lost, the grammatical system acquired by late-learners is characterized by grammatical inconsistency and probabilistic use of incorrect forms. This suggests that there may be important differences in the statistical learning process that takes place at different ages. Newport (1990) suggests that these stem from constraints placed on the system by children's limited memory capacity, which restrict the input to the statistical learning system in the early stages of learning (Hudson Kam & Newport, 2009; Elman, 1993 — though see Rohde & Plaut, 2003). Another possibility is that there may be differences in the way child and adult learners weigh and combine different probabilistic sources. Exploring these possibilities may further illuminate the mechanisms of native language learning and why it is so consistently successful.

Despite holes in our current knowledge, it seems clear that statistical learning mechanisms play a critical role in human language. Since I began this chapter by emphasizing our biological "predisposition" for language, it is worth considering again how this approach fits into the long-standing nature-nature controversy. Traditionally, statistical learning has been associated with an empiricist approach to language which deemphasizes the contribution of the learner. In contrast, I have emphasized that a full statistical learning account must specify (a) what statistical computations the system can calculate, and how information is integrated (b) the nature of the representations (formal and semantic) over which these analyses occur. In fact, statistical learning theories, and in particular working computational models, actually force us to make quite precise claims about the type of information that is primitive to the learning system. Of course it is always possible that representations which are primitive in one linguistic domain may ultimately be derived from lower-level primitives — but this leads to testable hypotheses about *that* learning process and how the derived representations feed into higher level processes.

Another contentious issue within the acquisition literature is the extent to which language learning rests on language-specific versus domain-general processes (see also Muller, this volume). I think that this division arises primarily from reactions for and against the "strong" UG view, which certainly makes claims about linguistic specificity. However, if our goal is to understand the cognitive processes and structures which allow human language, the focus on whether these are shared by other cognitive systems appears less important. For example, we have seen that children's self-created gestural systems communicate "thematic role" information. Many researchers have pointed out that such conceptual information might not be specific to the linguistic system (e.g. McClelland & Bybee, 2007). In fact, Goldin-Meadow (2005) does not dispute this point. Nevertheless, as she points out, the fact that the children communicate this particular set of conceptual structures, and that these also show up across human languages, is surely important in understanding what children bring to language learning. Similarly, the types of hierarchical relations seen in human language may also be evident in other cognitive systems such as motor planning (Rosenbaum et al., 1983). However, recognizing that human language (and no

other communication system) conveys messages which are hierarchically structured is surely critical to understanding the nature of our endowment for language.

Other more general cognitive developments are undoubtedly vital for language learning. Another topic which I have neglected in this chapter is the contribution of more general social cognition. Tomasello in particular has argued that human language learning rests on more general social adaptations, and reports various ways in which human social interactions differ from those of primates (Tomasello et al., 2005). In particular, he has emphasized the human ability to comprehend *intention*, which is critical in inferring the message conveyed by a linguistic utterance.

In short, it is likely that our 'specialization' for language relies on a variety of different cognitive abilities. Each of these may also play a role in other cognitive behaviors, and be shared to some extent by other species. Our goal is to understanding how these come together to give us the — uniquely human — Language Phenomenon.

Acknowledgements Many thanks to the following people for helpful discussions and/or comments on earlier drafts of the chapter: Adele Goldberg, Franklin Chang, Joanne Taylor, Jennifer Thomson and Edward Longhurst.

References

Akhtar, N. & Tomasello, M. (1997). Young children's productivity with word or-der and verb morphology. *Developmental Psychology*, 33, 952–965.

Albright, A. & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90, 119–161.

Allen, J. & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *Emergentist Approaches to Language: Proceedings of the 28th Carnegie symposium on cognition*, (pp. 115–151). Hillsdale, NJ: Lawrence Earlbaum Associates.

Ambridge, B., Pine, J. M., Rowland, C. F., & Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgments of argument-structure overgeneralization errors. *Cognition*, 106, 87–129.

Aslin, R. N., Jusczyk, P., & Pisoni, D. B. (1998). Speech and auditory processing during infancy: constraints on and precursors to language. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology: cognition, perception, and language*, (pp. 147–254). New York, NY: Wiley.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10, 533–581.

Berko, J. (1958). The child's learning of english morphology. *Word*, 14, 150–177.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In J. A. Hawkins (Ed.), *Explaining Language Universals*, (pp. 73–101). New York, NY: Basil Blackwell.

Braine, M. D. S. (1971). On two types of models of the internalization of grammars. In D. I. Slobin (Ed.), *The ontogenesis of grammar: A theoretical symposium*, (pp. 153–186). New York, NY: Academic Press.

Braine, M. D. S. & Brooks, P. J. (1995). Verb argument strucure and the problem of avoiding an overgeneral grammar. In M. Tomasello & W. E. Merriman (Eds.), *Beyond names for things: Young children's acquisition of verbs*, (pp. 352–376). Hillsdale, NJ: Lawrence Earlbaum Associates.

Brooks, P. J. & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology*, 35, 29–44.

Brown, R. & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language*, (pp. 11–53). New York, NY: Wiley.

Bryant, B. D. & Miikkulainen, R. (2001). From word stream to gestalt: A direct semantic parse for complex sentences. Technical Report TR-AI98-274, Department of Computer Sciences, The University of Texas at Austin.

Bybee, J. L. (1985). *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins.

Bybee, J. L. (1995). Diachronic and typological properties of morphology and their implications for representation. In L. Feldman (Ed.), *Morphological aspects of language processing*, (pp. 225–246). Hillsdale, NJ: Lawrence Earlbaum Associates.

Bybee, J. L. & Moder, C. L. (1983). Morphological classes as natural categories. *Language*, 59, 251–270.

Bybee, J. L. & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language*, 58, 265–289.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, 26, 609–651.

Chang, F. (2009). Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61, 374–397.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, 113, 234–272.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287–291.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1968). *Language and Mind*. New York, NY: Harcourt, Brace and World.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper and Row.

Christiansen, M. H. & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23, 157–205.

Christiansen, M. H. & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489–509.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, (pp. 136–143). ACL.

Crain, S. & Pietroski, P. (2006). Is generative grammar deceptively simple or simply deceptive? *Lingua*, 116, 64–68.

Croft, W. (2001). *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford: Oxford University Press.

Daugherty, K. & Seidenberg, M. S. (1992). Rules or connections? The past tense re-visited. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, (pp. 259–264). Hillsdale, NJ: Lawrence Earlbaum Associates.

DeLancey, S. (2005). Adpositions as a non-universal category. In Z. Frajzyngier, L. Hodges, & D. S. Rood (Eds.), *Linguistic Diversity and Language Theories*, (pp. 185–202). Amsterdam: John Benjamins.

Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Evans, N. & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32, 429–492.

Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology*, 46, 621–646.

Finch, S. & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. In A. R. . K. Eiselt (Ed.), *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, (pp. 301–306). Hillsdale, NJ: Lawrence Earlbaum Associates.

Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello (2000). *Cognition*, 82, 259–278.

Garnsey, S. M., Pearlmutter, N. J., Meyers, E., & Lotocky, M. A. (1997). The contribution of verb bias to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37, 58–93.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, 17, 648–691.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, 1(1), 23–64.

Goldberg, A. E. (1995). *A construction grammar approach to argument structure*. Chicago, IL: The University of Chicago Press.

Goldberg, A. E. (2005). *Constructions at work: the nature of generalization in language*. Oxford: Oxford University Press.

Goldin-Meadow, S. (2003). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. New York, NY: Psychology Press.

Goldin-Meadow, S. (2005). What language creation in the manual modality tells us about the foundations of language. *The Linguistic Review*, 22, 199–225.

Gonnerman, L. M., Seidenberg, M. S., & Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General*, 136, 323–345.

Greenberg, J. H. (Ed.) (1963). *Universals of Language*. Cambridge, MA: MIT Press.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language*, 65, 203–255.

Halle, M. & Mohanan, K. P. (1985). Segmental phonology of modern English. *Linguistic Inquiry*, 16, 57–116.

Hare, M., Elman, J. L., & Daugherty, K. G. (1995). Default generalization in connectionist networks. *Language and Cognitive Processes*, 10, 601–630.

Hare, M. L., Ford, M., & Marslen-Wilson, W. D. (2001). Ambiguity and frequency effects in regular verb inflection. In J. Bybee & P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure*, volume 45 of *Typographical Studies in Language*, (pp. 181–200). Amsterdam: John Benjamins.

Hart, B. & Risely, J. (1995). *Meaningful Differences in the Everyday Experience of Young Children*. Baltimore, MD: Brookes Publishing Co.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it and how did it evolve? *Science*, 298, 1569–1579.

Hay, J. B. & Baayen, R. H. (2005). Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences*, 9, 342–348.

Hudson Kam, C. L. & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59, 30–66.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, 6, 345–350.

Joanisse, M. F. & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, USA*, 96, 7592–7597.

Joanisse, M. F. & Seidenberg, M. S. (2005). Imaging the past: Neural activation in frontal and temporal regions during regular and irregular past tense processing. *Cognitive, Affective, and Behavioral Neurosciences*, 5, 282–296.

Johnson, J. S. & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology*, 21, 60–99.

Kam, X.-N., Stoyneshka, I., Tornyova, L., Fodor, J., & Sakas, W. (2008). Bigrams and the richness of the stimulus. *Cognitive Science*, 32, 771–787.

Lakoff, G. (1987). *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago, IL: University of Chicago Press.

Langlacker, R. W. (1987). *Foundations of cognitive grammar: Volume 1, Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Lewis, J. D. & Elman, J. L. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development*.

Li, C. & Thompson, S. A. (1974). Coverbs in Mandarin Chinese: Verbs or Prepositions? *Journal of Chinese Linguistics*, 2, 257–278.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates, 3rd edition.

Manning, C. D. & Schutze, H. (2001). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, 29, 189–256.

Marslen-Wilson, W. D. & Tyler, L. (2007). Morphology, language and the brain: the decompositional substrate for language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362, 823–836.

McClelland, J. L. & Bybee, J. (2007). Gradience of gradience: A reply to Jackendoff. *The Linguistic Review*, 24, 437–455.

McClelland, J. L. & Patterson, K. (2003). Differentiation and integration in human language: A reply to Marslen-Wilson and Tyler. *Trends in Cognitive Sciences*, 7, 63–64.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science*, 20, 47–73.

Mintz, T., Newport, E. L., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26, 393–424.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90, 91–117.

Nevins, I., Pesetsky, D., & Rodrigues, C. (2009). Pirahã exceptionality: a reassessment. *Language*, 85, 355–404.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, 14, 11–28.

Newport, E. L. & Aslin, R. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

Newport, E. L., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. Snow & C. A. Ferguson (Eds.), *Talking to Children: Language Input and Acquisition*. Cambridge: Cambridge University Press.

Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overregularizations in language acquisition? In *Proceedings of the 24th Conference of the Cognitive Science Society*, (pp. 720–725). Mahwah, NJ: Lawrence Earlbaum Associates.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37, 607–642.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure*. Learning, development and conceptual change. Cambridge, MA: MIT Press.

Pinker, S. (1999). *Words and Rules*. New York, NY: Basic Books.

Pinker, S. & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6, 456–463.

Plaut, D. C. & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, 15, 445–485.

Plunkett, K. & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition*, 38, 43–102.

Plunkett, K. & Nakisa, R. C. (1997). A connectionist model of the Arabic plural system. *Language and Cognitive Processes*, 12, 807–836.

Pollack, J. B. (1988). Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society*, (pp. 33–39). Hillsdale, NJ: Erlbaum.

Post, B., Marslen-Wilson, W. D., Randall, B., & Tyler, L. K. (2008). The processing of English regular inflections: Phonological cues to morphological structure. *Cognition*, 109, 1–17.

Prasada, S. & Pinker, S. (1993). Generalizations of regular and irregular morphology. *Language and Cognitive Processes*, 8, 1–56.

Prasada, S., Pinker, S., & Snyder, W. (1990). Some evidence that irregular forms are retrieved from memory but regular forms are rule generated. Paper presented at the Psychonomic Society meeting.

Radford, A. (1988). *Transformational grammar: a first course*. Cambridge: Cambridge University Press.

Reali, F. & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cognitive Science*, 29, 1007–1028.

Rohde, D. L. T. & Plaut, D. C. (2003). Less is less in language acquisition. In P. Quinlan (Ed.), *Connectionist modeling of cognitive development*, (pp. 189–231). Hove: Psychology Press.

Rosenbaum, D. A., Kenny, S., & Derr, M. A. (1983). Hierarchical control of rapid movement sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 86–102.

Rumelhart, D. & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, volume 2. Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.

Schreuder, R., de Jong, N., Krott, A., & Baayen, H. (1999). Rules and rote: Beyond the linguistic either-or fallacy. *Behavioral and Brain Sciences*, 22, 1038–1039.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275, 1599–1604.

Seidenberg, M. S. & Pettito, L. A. (1979). Signing behavior in apes: A critical review. *Cognition*, 7, 177–215.

Singleton, J. L. & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, 49, 370–407.

Snedeker, J. & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49, 238–299.

St. John, M. F. & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217–257.

Steedman, M. (2002). Connectionist and symbolic representations of language. In *Encyclopedia of Cognitive Science*. Macmillan.

Terrace, H. S., Petitto, L. A., Sanders, R. J., & Bever, T. (1979). Can an ape create a sentence? *Science*, 206, 891–902.

Theakston, A. L. (2004). The role of entrenchment in constraining children's verb argument structure overgeneralisations: a grammaticality judgment study. *Cognitive Development*, 19, 15–34.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74, 209–253.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675–735.

Trueswell, J. C. & Kim, A. E. (1998). How to prune a garden-path by nipping it in the bud: Fast-priming of verb argument structures. *Journal of Memory and Language*, 39, 102–123.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19, 528–553.

Wierzbicka, A. (1988). *The semantics of grammar*. Amsterdam: John Benjamins.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, 56, 165–209.

Xu, F. & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, 22, 531–556.