Chapter 16

Language processing and language learning

Nick Chater, Andy Perfors & Steven T. Piantadosi

A main objective of this book has been to illustrate how probabilistic inference over rich structured representations provides a powerful machinery for modelling human intelligence. We have seen that structured representations from graphical models to logic can help represent knowledge and categories, encode the perceptual world, or reason about naïve physics or other minds. Yet the domain in which structured representations are most transparently relevant to cognition is, of course, human language. In this chapter, we apply some of the foundational ideas developed in this book to understand the cognitive processes underpinning language processing.

Probabilistic ideas have often been overlooked or even actively pushed aside in the study of language (e.g., Chomsky, 1969, see Norvig, 2012). One reason for this is that it is sometimes been assumed that a probabilistic approach to language can only work if language has a very simple structure, corresponding to statistics over pairs or triples of phonemes or words (for discussion, see Jurafsky & Martin, 2008), or through learning associations between words or distributional patterns linking words and their contexts (Landauer & Dumais, 1997; Redington, Chater, & Finch, 1998). Indeed, probabilistic approaches to language, and by extension connectionist approaches, have sometimes been viewed as something close to a covert return to behaviorism (Fodor & Pylyshyn, 1988). But we have seen that this reaction against probabilistic ideas as incompatible with structured symbolic representations is out of date (Chater & Manning, 2006). Indeed, understanding how language is processed and learned crucial requires the integration of both structured representations and probabilistic methods.

Probability enters crucially into the cognitive science of language in two ways. First, we have the problem of interpreting language—i.e., creating a rich representation of the phonemes, words, syntactic structure and, crucially, the meaning, or linguistic input from a noisy and highly ambiguous stream of speech (or the similar ambiguous stream of visual input in the context of sign language). The problem of inferring the most likely structure from a noisy input is, of course, a paradigm example of probabilistic inference—and as we have seen throughout this book, the standard Bayesian approach to this type of problem is to attempt to invert a generative model of the language. Thus, to work out the most likely analysis of the speech input requires inverting a model for generating, or synthesising, this speech input. This general line of thinking has a long history in the psychology of language, tracing back to the analysis-by-synthesis models of speech perception developed at the Haskins laboratory in the 1950s and 1960s (e.g., Halle & Stevens, 1962), and is, of course, in line with the Bayesian viewpoint on cognition explored throughout this book. We shall see that many of the computational problems associated with symbolic approaches to language processing, such as the spectacular ambiguity of natural language, are greatly eased when multiple levels of probabilistic constraints can be applied to prune the vast number of possible readings of sentence, or indeed, an acoustic wave form. We note that there is an increasing body of experimental evidence across a range of linguistic levels which fits well with the probabilistic framework.

The second way in which probability enters the story concerns how the generative model of language is *learned*. Here, the objective is not merely to infer the structure of the speech or other linguistic input in real time. It is instead to infer a model of the entire language itself, and to thus be able to use this model in using language—to correctly produce language, to understand novel sentences that may never have been heard before, and to distinguish between sentences which are grammatically acceptable and those that are not. This model is learned through experience, particularly crucially, in the first years of life. Learning language is, then, yet another problem of probabilistic inference, where now the aim is not to infer the structure of a particular speech signal, given a generative model of the language, but to infer the generative model itself, given a history of linguistic (and potentially other) input. Bayesian models of learning natural language grammars also throw a radically new perspective on nativist arguments by (Chomsky, 1980; Pinker, 1994) and many others. These authors have argued for the necessity of an innate universal grammar due to the supposed impossibility of learning an infinite language from the finite, and indeed somewhat noisy, sample of assurances available to children. The claim that learning a language without a great deal of innate information is impossible is back up by **poverty of the stimulus** arguments: that the child has too little, and too poor-quality, data for learning to be possible (Chomsky,

1986).

As in the case of understanding individual utterances, purely symbolic models of language acquisition have no principled way of prioritizing between the vast range of possible models of the language that will fit with a particular body of linguistic data (and indeed, they also tend to fare badly when dealing with noisy data). The Bayesian approach addresses this problem, by focusing on grammars that are a priori more plausible[1] but which also fit well with the observed data.

Mathematical arguments show that this approach can work in principle, to overcome the apparent "logical" problems of language acquisition, without building in strong language-specific prior information (e.g., Chater, Clark, Goldsmith, & Perfors, 2015; Chater & Vitányi, 2007). Here we will describe recent computational work that has demonstrated that Bayesian learning models can acquire a wide range of aspects of natural language from language corpora, without needing to build language-specific prior information. Later in this chapter, we review work that focuses on the acquisition of the key grammatical patterns hypothesized by Chomsky and others as central to language, using probabilistic inference over programs. This work uses **adaptor grammars** and similar approaches, which allow the possibility of gradually constructing an increasingly complex grammar as more linguistic data is encountered, using the principles of nonparametric Bayesian modeling introduced in Chapter 9). This approach is readily compatible with item- or construction-based models of language that are currently prevalent in linguistics and language acquisition research. We also consider how Bayesian learning over symbolic structures relates to the astonishingly high-levels of performance in principle natural language tasks by very large deep neural networks trained on vast linguistic corpora. Nonetheless, we argue that such **large language models** do not yet provide a plausible cognitive model of human language processing, for a number of reasons, including the lack of a natural interface with meaning and pragmatic use of language (Chater, 2023).[2]

16.1 Language Processing

According to conventional approaches in linguistics and psycholinguistics, language is governed by many layers of representations, which can be ordered in increasing levels of abstraction (see Figure 16.1). When interpreting speech, for example, we might begin with an acoustic representation of the input arriving at the ear (which might be something akin to a Fourier analysis, picking out the spectral power at each frequency). The acoustic input arriving at the ear will, of course, contain not merely the speech we are attempting to understand, but background noises of all kind—ranging from the chatter of other speakers, to background music and the rumble of traffic. One immediate challenge is, then, to split out the acoustic signal associated with speech. Another challenge is to infer the phonemes in the signal (irrespective of the enormous variations between accents, individual speakers, acoustic environments (down a phone line, in an echoey swimming pool, etc), patterns of intonation and many more variations). Then there is the task of inferring how the stream of phonemes splits into words (and morphemes, such as verb endings, case markings and so on), and how these words cohere to convey meaningful phrases and sentences and more abstract of levels of meaning concerning how what is being said fits into the rest of the conversation, or how it relates to current perceptual input or to background knowledge. Each of these steps is formidably challenging and potentially interdependent (so that information about meaning may, for example, help us decode noisy or corrupted speech Mattys, Davis, Bradlow, & Scott, 2012).

In view of our focus on probabilistic models of structured representations, we focus here on language interpretation at or above the level of the word. Language scientists typically assume the patterns in

[1] One natural a priori bias is in favor of simpler grammars, as we will discuss further in the final substantive chapter of this book

[2] Nonetheless, Contreras Kallens, Kristensen-McLachlan, and Christiansen (2023) convincingly argues that the spectacular success of large language models does severely undermine the credibility of in-principle arguments that language learning is impossible without substantial innate grammatical knowledge.

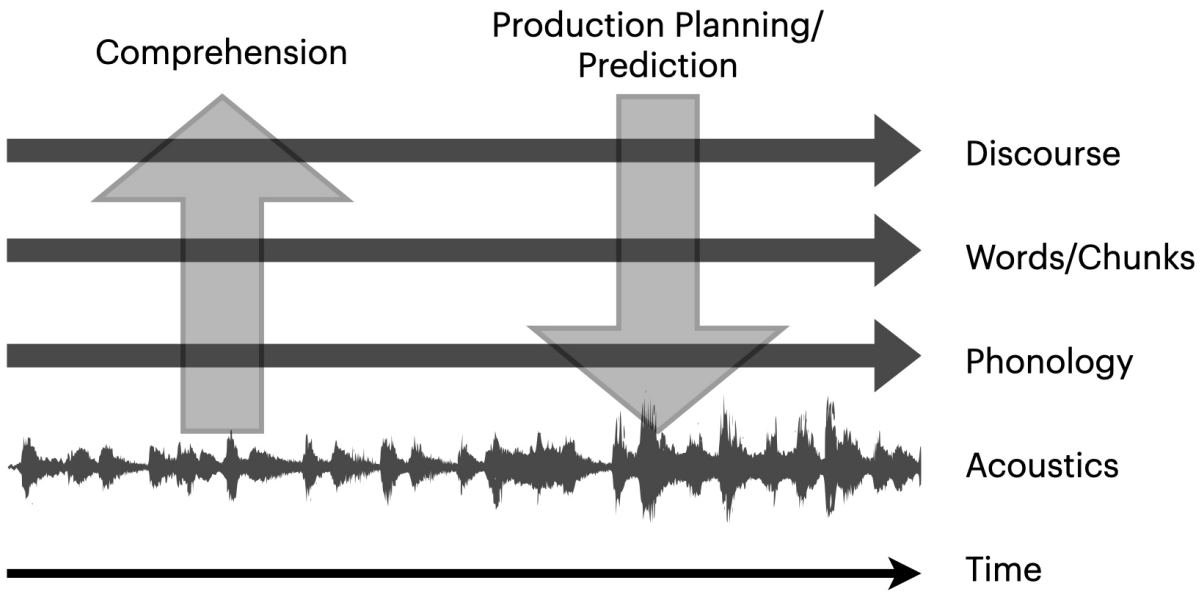


Figure 16.1: Computing levels of linguistic representation in real-time language understanding and production. In language understanding, the brain is assumed to begin with acoustic representations and to successively compute more abstract and temporally extended representations (from phonemes, to words, all the way up to representations of the entire discourse—theorists differ concerning the number and nature of the specific representations). In language production, the process is reversed, beginning with abstract representations of meaning and generating an acoustic signal. Yet the computational mechanisms underpinning understanding and production may be tightly coupled (Pickering & Garrod, 2013): when we are listening to speech, we are engaging in top-down processing to (re-)construct what a person is saying to us in real time. This "analysis-by-synthesis" perspective on the perception of speech is naturally aligned with the Bayesian approach to interpreting sensory input more broadly (Yuille & Kersten, 2006). Figure redrawn and adapted from Christiansen and Chater (2016b).

language above the level of the word (or more strictly, the morpheme, to include meaningful sound units, such as tenses and cases) is internally represented as a grammar: that is, a system of symbolic rules that can generate the sentences in a language, and which provide an analysis of sentence structure which provides a framework for semantic analysis (i.e., the analysis of the meaning of a sentence).

16.1.1 The challenge of ambiguity

The process of grammatical analysis is fraught with difficulties. One set of problems concerns the spectacular ambiguity of individual words. You most likely interpreted the word *set*, in the last sentence without pause or difficulty. But, taken in isolation, it has at least thirteen different possible meanings assuming *set* is functioning as a noun, and even more verb and adjective meanings, according to a psycholinguistic database of meanings like WordNet (Miller, 1995). We have the set lived in by a badger, a set of tennis, we get set, read set texts, we can set our heart on something, set off on a journey, wonder whether our ice has set, be set up, and many, many more—where the same word has many syntactic functions, as well as different meanings. This lexical syntactic and semantic ambiguity likely helps to make language efficient by allowing us to leave out information which is redundant with the context (Piantadosi, Tily, & Gibson,

2012) but in turn it makes language processing and understanding more complex because comprehenders need to sort out which meaning was intended. If we created distinct words for every possible meaning of *set*, for example, these words would necessarily be, on average, longer and more complex—there are, after all, only a limited number of short words to go around. But the same point applies to resolving ambiguity in any way—it takes extra linguistic material to remove ambiguity, which slows down communication. The cognitive system must find a balance between using snappy but highly ambiguous language, which needs to be disambiguated by inferences from context, and less ambiguous but more ponderous language. It turns out that the humans prefer a surprisingly high level of ambiguity, with which our brain copes remarkably well.

The challenges increase when we turn to syntactic ambiguity, where a single sequence of words can have multiple possible sentence structures, and often these alternative structures have distinct meanings. For example, "I tripped the clown with the skateboard" could mean that I *used* the skateboard to trip the clown, or could mean that the clown *had* the skateboard.

Psycholinguists—who use primarily experimental methods to study how the brain represents and processes language—have studied even more diabolical examples, such as **garden path** sentences. To choose a famous example from (Bever, 1970), consider the sentence fragment *The horse raced....* This is naturally (and, one might think, inevitably) viewed as a sentence in which *the horse* as the subject of the verb *raced*; but if the sentence turns out to be *the horse raced past the barn fell*, this assumption is revealed to be incorrect, and the language processor will be confounded. To understand the sentence, we have to reanalyze the beginning and resolve the ambiguity of *The horse raced...* another way. The only way to make sense of the structure of this sentence is to see it as a contraction of *the horse (that was raced past the barn) fell*—a structure analogous to *The picture painted by the artist fell.* The horse turns out not to be the subject of the verb *raced* after all, but of the verb *fell*—and we are picking out the particular horse *that was raced past the barn*, by person or persons unknown. This example shows that even a simple string of three words *the horse raced...* can turn out to be unexpectedly locally syntactically ambiguous—until we see the rest of the sentence, the language process often cannot know for sure that it has the right structure.

The language processor does, though, tend to jump to conclusions. Indeed, it turns out that the brain typically uses all the information it can to resolve syntactic, and other, ambiguities as quickly as possible. But this approach to ambiguity resolution will sometimes lead to trouble if the brain's first guess is incorrect. Thus, once the language processor has jumped to the conclusion that *the horse* is the subject of the past tense verb *raced*, the arrival of the next word *fell* leads the language processing system run aground: a recalculation is required.

This type of garden path phenomenon is no mere curiosity. Indeed, one of the remarkable discoveries of symbolic computational linguistics has been that lexical ambiguity, and local syntactic ambiguity, is everywhere—so that the space of possible 'readings' of the parts of a sentence that the brain has to choose between typically grows exponentially with sentence length. It is easy to imagine we can dismiss such phenomena by the riposte that *in context* and with the right intonation (if the sentence is spoken, not written), it will almost always be "obvious" which structure is the right one. This is quite right (Piantadosi et al., 2012; Miller, 1951)—but it raises the scientific questions targeted by linguistics and psycholinguistics of *how* the language processing system is able to succeed in using context to resolve such ambiguities, and to do so in real time.

One simple strategy is for the language processor to have a bias in favor of common syntactic strutures. And indeed, people have been shown to use cues like the baseline frequency of different readings in order to make the best guess about the appropriate structure. But lots of other factors matter too, such as the specific words involved, prior context, stress, intonation, and many more. The integration of different types of probabilistic cues to provide the best overall interpretation is precisely where a probabilistic approach to inference is especially helpful. Jurafsky (1996) outlines a pioneering model of how to frame and solve this type of problem in probabilistic terms, accounting for a wide range of psycholinguistic

phenomena (see also Jurafsky, 2003).[3] This model ranks possible ambiguous meanings and syntactic constructions by their conditional probability, pruning "low-ranked" options using a popular heuristic breadth-first search algorithm called beam-search. The pruning of unpromising interpretations in parsing explains why, for example, the reading of *the horse raced past the barn* as meaning *the horse that had been raced by the barn* has been rejected as highly unlikely, before the arrival of *fell*. So when *fell* is encountered, the language processor becomes stuck, and finds it difficult to recover to make sense of the sentence.

16.1.2 Probabilistic parsing

Let's look at the problem of assigning syntactic structures to sentences—the problem known as "parsing"—more formally from a probabilistic point of view. Probabilistic parsing involves estimating the probability of different parse trees t (or whatever grammatical formalism we favor, which might take the form of dependency diagrams, or attribute value matrices, among many others), given a sequence of words s.[4] Suppose we have a probabilistic model P_m of the language. Then, using the normal Bayesian formula, we have:

$$P_m(t|s) = \frac{P_m(s|t)P_m(t)}{\sum_{t'} P_m(s|t')P_m(t')} \tag{16.1}$$

So we need a prior $P_m(t)$ over tree structures t; and a way of working out the conditional probability $P_m(s|t)$ of a sentence s, conditional on a particular t. But, to work out the denominator, we face the usual problem of summing over a potentially very large set of possible trees—which can be computationally costly.

The prior $P_m(t)$ can, perhaps most naturally, be determined by the complexity of the parse tree.[5] Specifically, if the shortest code that could express the parse tree is $length(t)$, then one natural prior is proportional to $2^{-\text{length}(t)}$. But the task of finding the *shortest possible* code for a parse tree (or almost any other representation) will not in general be computable—the space of possible codes is too difficult to search. But a crude heuristic is to use instead the length of the parse tree when expressed a standard form, and a simple coding language (or even more crudely, simply to count the number of, e.g., grammatical rules invoked). The more crucial question concerns working out the conditional probability of the sentence given the tree. If the language model has a simple form, such as a stochastic phrase structure grammar, then this may be fairly straightforward.

Figure 16.2a shows a simple grammar fragment, with **phrase structure** rules and rules for converting **syntactic categories** into specific words. Each of these rules can be associated with a probability. So, for example, generating any sentence begins with an S symbol which in this simple grammar generates NP VP with probability 1. The NP then generates either V NP with probability .75 or a V NP PP (i.e., adding a prepositional phrase) with probability .25. A crucial simplifying assumption here is that the rules operate independently—each NP has the same probability of converting into an NP VP or a V NP PP structure, for example, whatever its role in the rest of the sentence. The process of applying rules continues until we have a string of words, e.g., *the girl saw the boy with the telescope*. This whole process provides a generative probabilistic model P_m for syntactic structures and thereby over strings of words (the parse tree t specifies a string of words s, so that $P_m(s|t) = 1$). Thus, the probability $P_m(t, s)$ of a particular string of words s with a parse tree t is simply the product of the probabilities of the different rules in the parse tree. Then the total probability of a string of words, $P_m(s)$, is just the sum

[3] Many influential early psycholinguistic models deliberately ignored probabilistic cues and focus instead on attempting to prune syntactic ambiguities based on principles based on syntactic structure alone, with principles such as "minimal attachment" and "late closure" (e.g., Frazier & Fodor, 1978). This perspective was partly driven by the idea that syntactic processes in language should be independent of other linguistic levels (e.g., Fodor & Garrett, 1974).

[4] We'll abstract away from details of the speech input henceforth, and consider s to be a string a words—but this is an oversimplification. Intonation, in particular, provides useful guidance about the syntactic structure of a sentence.

[5] We will return to these issue in Chapter 20, on algorithmic probability and related ideas

of these probabilities for the different trees, t' that yield that string of words: $\sum_{t'} P_m(t', s)$ (this is just a rearrangement of the denominator in the Bayesian equation above).

Figure 16.2b-c illustrate how two different syntactic trees can generate the same string of words—generating a syntactic (and here also semantic) ambiguity. The structures differ regarding whether *with the telescope* is a prepositional phrase modifying how the action of seeing was achieved: *the girl* [*saw*] [*the boy*] [*with the telescope*] or whether this phrase picks out a particular boy who was seen *the girl* [*saw*] [*the boy with the telescope*].

When faced with a string of words which is ambiguous, the usual Bayesian procedure will prefer the parse tree with the highest probability. For example, if Bayesian inference is approximated by sampling, it may be biased towards choosing high-probability trees, in proporition to their probability. As can be seen in Figure 16.2, these preferences can sometimes be determined locally, by considering only the relevant parts of the parse tree that differ. From a probabilistic viewpoint, which structure is preferred depends on the specific probabilities in play, in contrast with previous structural models of parsing where the shape of syntactic tree is decisive. Experiments have indicated that parsing preferences do seem to follow probabilistic, rather than a purely structural, principles across a number of languages (Levy, 2008; Desmet & Gibson, 2003; Desmet, De Baecke, Drieghe, Brysbaert, & Vonk, 2006).

Note, though, that capturing such psychological data requires a much richer probabilistic model than specified here, in which the resolution of syntactic ambiguities can be influenced by specific lexical items, and the hearer's recent experience and general background knowledge (Traxler, 2014). Thus, for example, if the hearer is listening to a report of evidence that boys have been stealing equipment from the observatory then the much more likely interpretation is that the boy, not the girl, has the telescope. Similarly, the parallel sentence *the girl saw the boy with the microscope*, while equally syntactically ambiguous, will not typically be interpreted as imply the girl is looking through the microscope; on the other hand, as part of a story in which some children have been magically shrunken to a minuscule size, then this interpretation will suddenly become more probable. This last example illustrates that the probabilistic approach to language focuses on the probability that particular strings of words will be said (and the underlying meanings and syntactic structures that might underlie people's generation of these strings), not the probability that these sentences are true. After all, the language processor has no difficulty understanding fairy tales, by assuming that each new sentence is a plausible continuation of the story, even though the probability of the events being true may be close to zero.

To calculate what is a plausible sentence in a particular communicative context, and given a complex and noisy speech input, can depend, in principle, on information of any and every sort—what is plausible depends on the speech signal itself, the immediate environment, prior linguistic context, the hearer's model of the mind of the other person, their general knowledge about the world (and hence, indirectly, what it is reasonable to say about that world) and so on.

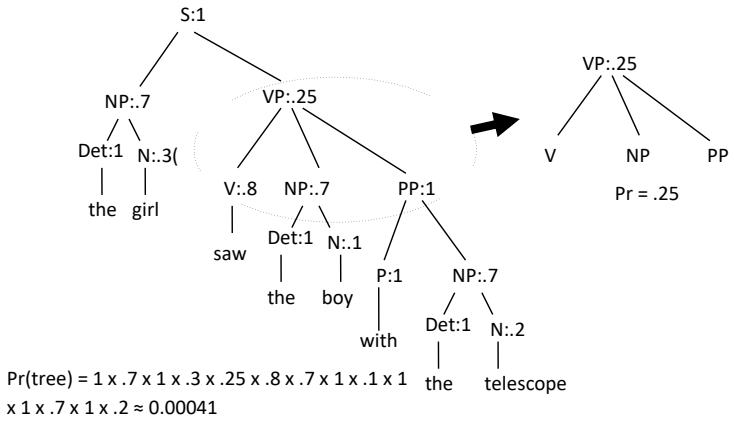
This full complexity is surely too great to be mentally represented in a probabilistic model of the language; and, in any case, even moderately complex probabilistic models are too complex for exact calculation, as problem exacerbated by the time-pressure under which language processing operates. And, indeed, the language processor does often settle for what seem to be "good enough" parses, using incomplete probabilistic analysis, although these analyses may turn out to be incorrect. Thus, for example, people typically, but wrongly, interpret *while Anna dressed the baby spit up on the bed* as implying that Anna dressed the baby (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Ferreira, Christianson, & Hollingworth, 2001). Indeed, people's interpretations of anomalous sentences can be well-explained by Bayesian decoding that takes into account both plausibility, as well as the probability of different kinds of mistakes or mis-hearings in the sequence of words (Gibson, Bergen, & Piantadosi, 2013).

The calculations in Figure 16.2 may suggest that the language processor considers the probability of different readings of a sentence its entirety—but this would require waiting for the end of the sentence before a probabilistic analysis. This would make communication painfully slow; but more importantly,

(a)

S → NP VP		(1)	V → saw	(.8)	N → cat	(.1)
VP → V NP		(.75)	V → prodded	(.2)	Det → the	(1)
VP → V NP	PP	(.25)	N → telescope	(.2)	P → with	(1)
NP → Det N		(.7)	N → stick	(.3)		
NP → NP PP		(.3)	N → girl	(.3)		
PP → P NP		(1)	N → boy	(.1)		

(b)



(c)

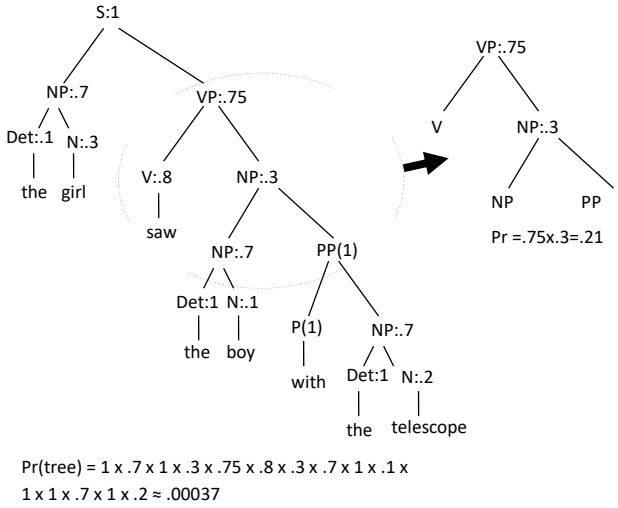


Figure 16.2: Ambiguity and phrase structure. (a) A fragment of a stochastic context-free phrase structure grammar, which can generate a simple sentence in two ways (b), (c). A Bayesian parser should prefer trees with higher probability. Focusing on the conditional probability of the sentence, given the tree, note that the two options (b) and (c) differ only regarding whether the prepositional phrase *with a telescope* attaches to the verb (modifying how the seeing is done) or the object noun phrase *the boy* (i.e., it is the boy with the telescope who is seen). The parts of the tree that differ are highlighted to the right. Here, the flatter tree structure invokes one less grammatical rule and assigns the word string a higher conditional probability, and hence should be preferred. Figure adapted from Chater and Manning (2006).

would run into fundamental limitations of human memory, which requires linguistic information to be chunked and recoded as soon as it is received, because otherwise it will immediately be overwritten by the onrushing torrent of speech (Christiansen & Chater, 2016b). So the language processor has to make probabilistic guesses in-the-moment, in the light of whatever information is available; and, as fresh words arrive, these guesses will sometimes prove to be incorrect (as we saw with *the horse raced past the barn*

fell above).

From this point of view, we can think of language processing as requiring continual anticipation of what is likely to come next, suggesting a tight coupling between language understanding and language production (Pickering & Garrod, 2013). This viewpoint captures the fact that we are often able to finish off one another's sentences, and that turn-taking handovers in fluent dialogue are astonishingly fast, implying that we often know what people will say before they have finished saying it, and are preparing our reply (Levinson, 2016). Moreover, a wide range of psycholinguist experiments and models have indicated a powerful role for predictive processes (Levy, 2008; Lowder, Choi, Ferreira, & Henderson, 2018). Indeed, it is natural to think of language understanding as analogous to word-by-word language production. This viewpoint fits, of course, with the broader analysis-by-synthesis perspective on language understanding, mentioned earlier, and aligns accounts of language processing with Bayesian accounts of perception (Yuille & Kersten, 2006).

16.1.3 Rational speech acts: Inferring what people are doing with language

Language processing must, of course, go beyond the analysis of individual lexical items, and syntactic and semantic structure, to work out what message the speaker is attempting to convey—and, again, to do this in real-time. This "pragmatic" interpretation of linguistic utterances is extremely complex and will involve knowledge of language, social conventions, the nature of the conversational interaction, and arbitrary background knowledge about the world. While a full discussion is far beyond our scope here, note that one approach is to view the problem of pragmatic inference as one of inferring people's intentions from what they say, just as we attempt to infer people's actions from what they do: using an inverse planning approach, that we discussed in the previous chapter. The idea is that in communication both parties assume that communicative signals are chosen to convey the meaning of interest as efficiently as possible. So, when a person says *some of my fish are black* we tend to infer that they aren't all black—because otherwise, the person could communicated more precisely (and at no extra communicative cost) that *all my fish are black*. Equally, if we see an escaped cow running down the street, we remark *look at that!* without having to give any further details, because it is clear that the escaped cow is the most unexpected aspect of the scene (indeed, saying *look at the escaped cow running down the street!*) would seem bizarrely prolix. But if you happened to want to draw attention to some other aspect of the scene, you would clearly have to be much more precise. The pragmatic principles underlying these types of inference can be formulated in Bayesian terms, in the **rational speech act** framework(Frank & Goodman, 2012; Goodman & Frank, 2016; Goodman & Stuhlmüller, 2013).

The rational speech act approach assumes that both speaker and hearer have common knowledge of literal meanings of their linguistic terms.[6] In particular, we can begin the analysis with the notion of a *literal listener*, who is presumed to use possible utterances from a speaker, combined with background knowledge, to make inferences about the state of the world. The literal listener, by virtue of being literal, makes no assumptions about the speaker being, for example, as helpful or informative as possible. But, in reality, of course, the speaker will choose messages in a thoughtful way. Thus, an utterance such as *I have a dog* means, to the literal listener, that the speaker has at least one dog. But in most contexts, if the speaker had two dogs, it would be more informative and helpful to say *I have two dogs*–this would provide the literal listener with more precise information. More generally, the speaker is assumed to choose what to say in order to maximize some utility function.[7] Knowing that the speaker will do this, the real listener (rather than the hypothetical literal listener) will assume that if the speaker had more

[6]The very idea of literal meanings is somewhat controversial in philosophy and the language sciences. Some theorists, for example, argue that context-specific meanings are primary, and abstract literal meaning that applies across contexts is at best a useful approximation (see, e.g., Christiansen & Chater, 2022). But even so, assuming literal meanings may provide a useful assumption.

[7]The speaker is assumed to choose probabilistically, use a soft-max function, rather than deterministically choosing to maximizing utility—otherwise the probabilities in the calculations all become jammed at 0 or 1.

than one dog, she would probably have said so; and hence that it is likely, though not certain, that the speaker has exactly one dog. This pattern of reasoning is not specific to language or communication—the listener is inverting a model of the speaker's actions, to infer the speaker's intentions. Now we can take a further step: knowing that the listener will make such inferences, the speaker can deliberately choose what she says, by inverting the model of the (non-literal) listeners inferences. So, for example, the speaker might not choose to say *I have one dog* precisely because the speakers knows that the more informal and slightly easier-to-say *a dog* will be interpreted as implying a single dog in any case. Thus, the listener is inverting the speaker's model, which will itself involve inverting the model of the literal listener. In principle, this hierarchy might continue further, though many iterations may be both unnecessary and cognitively infeasible.

The rational speech act approach has been applied to capturing a variety of what are known as conversational implicatures of language–i.e., inferences that go beyond the literal meaning of what is said in highly predictable ways (Grice, 1975). Thus, following the logic described above, the approach can explain why *some of the people enjoyed the party* seems to preclude the speaker knowing that *all of the people enjoyed the party* (or the speaker would have chosen to provide this more informative message). Related, though distinct, reasoning can help explain other non-literal uses of language, such as hyperbole, as when a person describes the weather as *boiling* or objects that there are *millions of reasons* why a project won't work (Kao, Wu, Bergen, & Goodman, 2014); or why unusual ways of expressing a meaning tend to suggest that this meaning applies in some unusual way, so that *Maria has the ability to finish the homework* seems to have very different import than the plain *Maria can finish the homework* (Bergen, Levy, & Goodman, 2016). The model captures the following line of reasoning: if there is nothing unusual about the situation being described, then the simplest way of expressing it would have been chosen. But it wasn't—so the speaker will intend, and the hearer infer, that the case is not usual. Mostly like, of course, the unusual aspect of the claim is that while John has the ability to finish his homework later, the usual implication that he is likely to do so is blocked.

There are, of course, other unusual situations that might be relevant instead. For example, suppose that the homework consists of a very hard set of math problems, and Maria is the class's star math student. So one teacher might remark to the another: "Maria has the ability to finish the homework, but none of the other students have the slightest chance." Here the *can* is avoided to stress Maria's specific aptitude for math. The open-ended nature of pragmatic reasoning, and its dependence on background knowledge of all kinds, makes it challenging to model. It also, of course, suggests that such inference will need to be seen as continuous with commonsense reasoning about the physical and social worlds, which we have argued should be formulated in Bayesian terms in this book.

We have so far assumed the existence of a literal listener as a "base case" from which successive layers of recursive inference can arise. If this assumption is removed, then the above Bayesian reasoning strategy based on inverse planning runs the risk of circularity. Indeed, this type of case arises when people use communicative signals that have no conventional meaning (whether or not those conventions are linguistic, or concern, say, facial expressions or gestures). For example, suppose that a professor ostentatiously uses a rival's book to stabilize a wobbly table (where other books and papers would work equally well; and perhaps the wobbly table is not a problem in any case). This action may be intended, and interpreted, as expressing contempt for the book (or perhaps even its author)—and might successfully convey this dismaying message to the audience present (particularly if they know about the rivalry). All may agree that such an action conveys a sense that the book is most useful as a physical object, rather than deserving to be read. But there is, of course, no such prior convention (and, of course, no literal meaning). The general problem of inferring meanings from signals in cases like this raises new and perhaps unexpected challenges. As Clark (1996) points out, communication is a joint activity—the parties have to have common understanding of what signals convey which messages.[8] And finding this understanding

[8]For experiments illustrating the sophistication of human joint reasoning with simple communicative set-ups (see, e.g., Galantucci, 2005; Misyak, Noguchi, & Chater, 2016).

requires that each can align with the mind of the other. Thus, a traditional Bayesian mind-reading approach to this problem can appear to lead to regress (Chater, Zeitoun, & Melkonyan, 2022). The receiver tries to guess what the sender intends the signal to convey. But in choosing her signal, the sender should therefore try to second-guess what the receiver will infer (so that she infers whatever the sender intended). But now the receiver has to "third-guess" what that second-guess might be, and so on, indefinitely. Following Clark (1996), one way to proceed is to assume that both parties should aim not to read-the-minds of the other, but rather to jointly infer the most appropriate signal-meaning mappings, based on their common ground (Chater & Misyak, 2021). This type of problem is particularly pressing, if we suspect that viewing communication as involving joint action and joint reasoning applies even to linguistic communication, perhaps because of a doubts that literal meaning is well-defined (Clark, 1996; Christiansen & Chater, 2022). Capturing this type of reasoning in a Bayesian framework is an interesting challenge for future research (for related work, see Stacy et al., 2021; Wang et al., 2021)).

16.2 Language acquisition

Children's astonishing learning abilities are nowhere better exemplified than in language: in just a few years babies transform themselves from helpless, nonverbal blobs to linguistic whizzes with a vast vocabulary, mastery of complex abstract syntax, and even the ability to indulge in wordplay and sarcasm. How do they do this? This is a vast question, impossible to do justice in one small part of one small chapter. Here we address three topics in language acquisition (learning how to recognize phonemes, how to segment speech into words, and learning grammar), providing a quick glimpse at how Bayesian approaches have been useful for shedding light on enduring questions within these areas.[9] The goal is to highlight the scientific insight that can result from the ability of Bayesian models to clarify *what can be learned* and *from what input* and (most importantly) *why*.

16.2.1 Phoneme learning

Phoneme learning refers to the process of acquiring the speech sounds specific to the language one speaks. One of the most interesting issues in this area is the **perceptual magnet effect**, which occurs after successful learning: discriminability between vowels is reduced near prototypical vowel sounds (this phenomenon is also discussed in Chapter 4). The pattern underlying this shrunken perceptual space is marked by decreased distance between items within a phonetic category and increased distance between items across categories: as an example, all /i/ exemplars sound more similar than their raw acoustic representations suggest, while /a/ and /i/ exemplars sound more dissimilar.

The perceptual magnet effect has been thoroughly empirically studied. For a long time most computational models either implicitly assumed that it was a categorisation effect parallel to categorical perception (Iverson & Kuhl, 1995) or focused on purely process-level accounts of how the effect might be implemented (Vallabha & McClelland, 2007). But key questions remained unanswered: *W*hy should prototypes exert a pull on nearby speech sounds? And why should a learner shrink phonetic space in particular directions as they learn phonetic categories? Feldman, Griffiths, and Morgan (2009) presented a Bayesian analysis that answered these questions by approaching speech perception as a problem of Bayesian statistical inference. They asked to what extent learners perform this task optimally given certain assumptions about their hypotheses, likelihoods, and priors. The model assumes that vowel sounds are generated from phonetic categories by sampling them from a target and adding noise; the learner must then work backwards from the exemplars of sounds they hear to infer the nature of the most probable target production. The model predicts that with experience, learners will realise that sounds near the

[9] One obvious omission here is learning the means of individual words, which we omit because many of the relevant issues are discussed in Chapters 3 and 8.

centre of categories are more frequent; they will then compensate for the noisy speech signal by biasing perception to the centre of the category. The model captures this intuition mathematically and also explains a range of empirical effects while making predictions about others. For instance, it predicts that category variance and degree of noise in the environment should affect the strength of the perceptual magnet. For more details on the model, see Chapter 4.

Extensions of this approach maintain the idea of speech perception as optimal statistical inference, and use Bayesian models to investigate how additional knowledge or assumptions can help. They indicate that **categorical perception of consonants** can indeed be accounted for within the same framework by assigning consonants less variability (compared with the variation due to noise) than vowels (Kronrod, Coppess, & Feldman, 2016). Other work indicates that learning to segment words at the same time as phonetic category learning can make both tasks more tractable, since word-level information is useful for disambiguating English vowel categories (Feldman, Goldwater, Griffiths, & Morgan, 2013). This may help to explain how tasks that individually might seem too difficult for an infant can jointly constrain and simplify each other (rather as the interweaving of answers in a crossword makes solving the individual clues easier, rather than more, difficult, as the answers mutually constrain each other).

Other work, similar in approach, can help clarify at what level the perceptual reorganisation underlying the perceptual magnet effect occurs (Kuhl, 2004). Are people more sensitive to overall *d*imensions that encompass many potential phonetic contrasts, like general voice-onset time distinction applied to many phonemes? Or do they become differentially sensitive only to specific contrasts, like the /b/-/p/ distinction? Results suggest that perceptual reorganisation involves making inferences about general dimensions, and that this constitutes a kind of hierarchical learning of the sort captured by a hierarchical Bayesian model of the kind discussed in detail in Chapter 8 (Pajak & Levy, 2014). One consequence of this is that second-language sound categories may be filtered through the prior learning shaped by a person's native-language phonetic inventory (e.g., Strange, 2011). This might, of course, lead to a systematic distortion of the phonetic categories in the second language, as is evident in the distinctive "accents" of second language speakers with different first languages.

Still other work has adapted this approach to the question of how speakers adapt to the phonetic variability in the world. Within a language, dialects vary systematically; and even within a speech community, individual speakers vary markedly from each other, such that one person's /b/ might sound like another person's /v/. Kleinschmidt and Jaeger (2015) developed a Bayesian model that views adaptation and learning as parts of the same process – inferring the correct generative model for the current speaker – but operating over different time scales. This model accounts for phenomena as disparate as perceptual recalibration, selective adaptation, and generalization across groups of speakers. It has also been extended to capture aspects of patterns by which phonology changes over successive generations of language users, whereby the phonetic cues that are more likely to be reduced over time are those that carry less information (Hall, Hume, Jaeger, & Wedel, 2018).

16.2.2 Word segmentation

The problem of **word segmentation** is the problem of identifying a lexicon by segmenting words out of continuous streams of speech of the sort that learners hear. There has been a long-tradition of research exploring the possibility that people solve this issue, at least in part, by learning the transitional probabilities (TPs) between phonemes or syllables. Specifically, it is assumed that low TPs are an indication of a word boundary (Saffran, Aslin, & Newport, 1996). Learning based on TPs is well-studied experimentally, but this work has mostly been limited to exploring how well people use TPs, sometimes in combination with other information, to segment artificial languages. These artificial languages are often learned in less than an hour. By contrast, modelling is valuable for exploring how different assumptions, and the utility of different kinds of information, scales with extremely large amounts of data – an amount comparable to what people hear over multiple years of life.

Consider one of the most influential Bayesian models of word segmentation (Goldwater, Griffiths, & Johnson, 2009), built on an earlier model developed by Brent (1999). This model, given continuous speech input, was able to infer a vocabulary whose size did not need to be pre-specified (thanks to a prior that assigned a positive probability to all vocabulary sizes, though with a strong bias towards small vocabularies, based on the ideas from nonparametric Bayesian statistics introduced in Chapter 9), and was originally used to explore the impact of different assumptions learners might make about how words are generated. One question was whether words are presumed to be generated independently, or are thought to be predictive of other words in the sentence. The model showed that assuming that each word is independent from all others means that under-segmentation errors are more likely (e.g., thinking *thedoggie* is one word rather than two), while assuming that each word constrains its neighbors reduces such errors considerably. Intuitively, this follows because the model can allow that the frequent co-occurrence of *the* and *doggie* can be explained as a result of a statistical connection between these words, rather than the assumption that they must form a single word—the independence model only has the latter option.

Frank, Goldwater, Griffiths, and Tenenbaum (2010) compared a variety of segmentation models and found that a Bayesian model was better able to capture human performance on an artificial segmentation task with varying word and sentence lengths, exposure times, and total vocabulary size than simpler models. Interestingly, however, all models did poorly unless they were modified to take human memory limitations into account. Other related models have been used to explore different ways to implement and test such limitations (Borschinger & Johnson, 2011; Phillips & Pearl, 2015). Still other questions have been addressed with Bayesian models of word segmentation. One is how (and whether) different kinds of information besides TPs are useful in helping learn with learning word segmentation – information such as stress, phonotactics, or referential information (Doyle & Levy, 2013; Borschinger & Johnson, 2014). These models have also been used as a basis for explanation or comparison while investigating how segmentation performance is affected by the distribution of words (Kurumada, Meylan, & Frank, 2013) or the size of the input (Borschinger, Demuth, & Johnson, 2012). Overall, while it is possible to defined moderately successful algorithms of segmentation by looking directly at TPs, viewing the problem of segmentation in the framework of Bayesian inference allows a range of productive extensions to be defined and explored.

16.2.3 Abstract linguistic structure

One of the most important topics in linguistics is the abstract structure of language, including questions about **morphology** (roughly, the system which composes the form and meaning of a word from its part, including markers for tense, case, or plurals and of course word-stems themselves) and **syntax** (roughly, the principles that determines allowable arrangements of words and morphemes to compose phrases and whole sentences). Developing full Bayesian models for learning these patterns is difficult because the space of possible patterns is very large. Nonetheless, such models have proven useful for investigating the learning of specific aspects of linguistic structure, or casting new light on the apparently severe problems inherent in learning abstract linguistic patterns from partial and noisy linguistic data.

Consider the **no-negative-evidence problem** (Baker, 1979), which centres around the issue of how language learners can make correctly pick up on the right linguistic generalizations (when the are often so many possibilities) in the absence of evidence about which ones are *incorrect*. Thus, the child hears positive examples of what can be said (e.g., by hearing the speakers around her)–but does not seem to have access to negative examples of what cannot be said. This problem arises, for example, in the problem of learning **verb-argument constructions**. Such constructions correspond to the set of possible arguments each verb can take, and are highly specific to individual verbs. Crucially, they vary considerably and are hard to predict based on underlying features like phonology or meaning. In English a verb like *give* can occur in two constructions, one taking a direct object dative (as in "he gave her the package") and one

indicating the recipient using a prepositional phrase (as in "he gave the package to her"). Other verbs, like *donate*, occur on only one of those constructions: in most dialects it is ungrammatical to say "he donated her the package." This is particularly puzzling, given that the meaning of *give* and *donate* are otherwise extremely similar. The puzzle is that the child doesn't make the apparently reasonable assumption that "he donated her the package" is likely to be acceptable. Of course, she never hears it—but then again this will be true of an almost limitless number of perfectly viable sentences. One might suspect that part of the story is that child might get negative feedback from care-givers, either through direct correction or indirectly through incomprehension. But many language acquisition researchers, following (Brown & Hanlon, 1970) have assumed that such feedback is rarely available and/or not sufficient even when it is provided (although see Chouinard & Clark, 2003; Hirsh-Pasek, Treiman, & Schneiderman, 1984).

In any case, let us suppose that negative evidence is at least not required for successful language acquistion. Without such evidence, how might a learner figure out which constructions go with which verbs without being told when she incorrectly overgeneralizes? Bayesian modeling suggests an answer, at least in principle. Just as hypotheses in any domain that more tightly predict the data should be preferred, so too should a learner (given enough data) be able to learn that some verbs take different constructions than others. Indeed, take the parallel with science: all we can ever observe are things that *can occur*. But we develop principles (including, for example, the conservation of energy, or the second law of thermodynamics) which powerful precisely because they specific what can't occur (e.g., no energy mysteriously coming into being out of nowhere; no spontaneous heat flow from cool bodies to hot bodies, and so on). So we can't decisively rule out the possibilities that these principles don't hold, and perhaps merely appear to hold by coincidence so that we might expect to see violations at any moment. But, for the Bayesian, the coincidence story is not impossible, just spectacularly unlikely. This general perspective is part of the motivation for the Bayesian approaches in the philosophy of science, (e.g., Earman, 1992; Horwich, 1982). Can it work in the more confined domain of learning verb argument structure and other puzzling aspects of language learning?

This approach is embodied by multiple specific computational models that differ dramatically in their representational assumptions, ranging from language-specific and intricate (Alishahi & Stevenson, 2008) to domain-general (Hsu & Griffiths, 2009; Perfors, Tenenbaum, & Wonnacott, 2010). Overall, such models demonstrate that it is the abstract behaviour of the probabilistic reasoning instantiated within the likelihood that drives the effect, while a prior preference for simplicity prevents overfitting.[10] Bayesian models have also been applied to extensions of the purely syntactic problem of learning verb constructions, such as the question of how the learner links the syntactic argument positions of a verb with the thematic roles specified by its semantics (Pearl & Sprouse, 2019).

The tradeoff between likelihood (which favours models that tightly fit the input data) and prior (which favours simpler models with shorter representations) is evident in many Bayesian models of morphology and grammar learning. Work by Perfors, Tenenbaum, and Regier (2011) demonstrates that a Bayesian learner given typical child-directed input can infer that grammars with hierarchical phrase structure provide a better representation for that data than grammars without it. This emerges out of a general prior favouring simplicity, since grammars with hierarchical phrase structure can capture typical English input more parsimoniously while still fitting the data well. Other research shows that human learning of artificial grammars can be captured by models that implement this kind of likelihood-prior tradeoff (Frank & Tenenbaum, 2011), and that children's early utterances can be captured by item-based grammars (Bannard, Lieven, & Tomasello, 2009). Indeed, Bayesian models have been applied to many classic learnability problems, from syntactic island effects (Pearl & Sprouse, 2013) to interpreting the anaphoric *one* in English (Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009) to linking logical and syntactic forms (Abend, Kwiatkowski, Smith, Goldwater, & Steedman, 2017). In some cases, it is use to the question purely in terms of simplicity (we'll explore the deep connection between Bayesian and simplicity-

[10] See also Chater et al. (2015) and Pearl (2021) for more general overviews of how Bayesian models can be applied to syntactic questions in language acquisition.

based ideas in Chapter 21). Specifically, where specific hypotheses about the abstract struture of language are being compared, we can estimate whether the additional complexity in formulating the hypothesis (e.g., specifying that *give* and *donate* have different argument structures) pays off providing a more precise encoding of the observing linguistic data (i.e., a higher likelihood, and a shorter code), to provide an overall shorter, simpler description (Hsu & Chater, 2010; Hsu, Chater, & Vitányi, 2011). These various types of analysis shows that the amount of linguistic data that must be encountered by a language learner (typically, a child) in order to overcome the problem of no negative evidence. Each case of the problem can be dealt with a separate analysis—giving different verdicts for different linguistic phenomena. It turns out, for example, that a few years of language input easily suffices to distinguish the arguments structures for *give* and *donate*. Other aspects of language, such as Chomsky's observation that, in many dialects at least, there are linguistically subtle restrictions on when we can, and cannot, contract *want to* to *wanna* in casual speech, seem to require infeasibly large amounts of data.[11]

Finally, as in other areas, Bayesian models of abstract linguistic structure have been useful in investigating to what extent synergies between different kinds of knowledge might lead to improved learning. For instance, Johnson, Demuth, and Frank (2012) shows that a learner who acquires collocational structure (roughly which words go next to each other) at the same time as attempting to link words to their referents and following social cues does better than a learner without these cues. This is because the tasks mutually constrain each other, as in solving a crossword, so that a learner that is sensitive to all can make more headway than a learner who tries to learn only one area at a time.

Many of the models that investigate the role of multiple sources of information use an approach called the adaptor grammar framework (Johnson, Griffiths, & Goldwater, 2007). It posits multiple layers of representation, some which capture the frequency of observed items (e.g., sentences, rules, or constructions), others which capture which linguistic patterns are permissible. Such models have been extended to incorporate contextual information (Synnaeve, Dautriche, Börschinger, Johnson, & Dupoux, 2014) and function word learning (Johnson et al., 2007) into models of word segmentation. Adaptor grammars have also been applied to problems like native language identification (Wong, Dras, & Johnson, 2012) and discourse structure (Luong, Frank, & Johnson, 2013). A similar framework is called fragment grammars (O'Donnell, 2015), which have been used to capture aspects of phonotactic learning (Futrelle, Albright, & O'Donnell, 2013), sentence processing (Luong, O'Donnell, & Goodman, 2015), and argument structure (Bergen, Gibson, & O'Donnell, 2013).

Overall, then, from the most basic of learning tasks to the most complex – from learning phoneme categories to making abstract grammatical generalizations – it is evident that children accomplish a great deal of linguistic learning in the first years of life. Bayesian models of language acquisition have been especially useful in defining the terms of that learning: clarifying how the assumptions and capacities of the learner lead to different outcomes. We now turn to another, still more abstract, aspect of the challenge of learning a language, which has been highly influential in shaping nativist thinking about language acquisition.

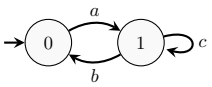
[11] So, for example, we can contract *who do you want to take to the party* to *who do you wanna take to the party*; but it seems distinctly odd to contract *who do you want to take Aishah to the party?* with **who do you wannna take Aishah to the party?*, using the conventional linguists asterisk to denote ungrammaticality. A standard linguistic analysis focuses on the restriction that there is a "hidden" gap *want _ to* in this case, where the name of a person, say, *Sarah* might be inserted. And the restriction appears to be that contractions can't occur across gaps. This type of constraint could arise from innate aspects of language processing, or constraints required to construct the best model of the systematic patterns across the rest of the language (i.e., just as we can figure out the answer to one crossword clue more easily by crosschecking with the answers to other clues).

16.3 Ascending the Chomsky hierarchy

One of the most important developments in the history of both linguistics and computer science was the discovery the **Chomsky hierarchy** of formal languages. In this way of thinking, a **formal language** is a technical term referring to a set of strings (Chomsky, 1957): for example, the set of strings $\{a, baa, aba\}$ is an example of a language; so is the set of English words containing the sequence *ing*; so is the set of all binary sequences that never contain 11 (e.g. $\{0, 1, 01, 00, 10, 001, \ldots\}$); and so is the set of sequences of words which are considered grammatical sentences in a natural language such as Tagalog, Swahili or Basque. Any such set may be finite or infinite, but primarily sets are distinguished by the computational resources that are required to generate or recognize them. This is why this area is considered to be foundational not only in linguistics, but also in the theory of computation (Hopcroft, Motwani, & Ullman, 2001), which seeks to characterize what kinds of problems computers are able to solve using different resources.

16.3.1 Finite and regular languages

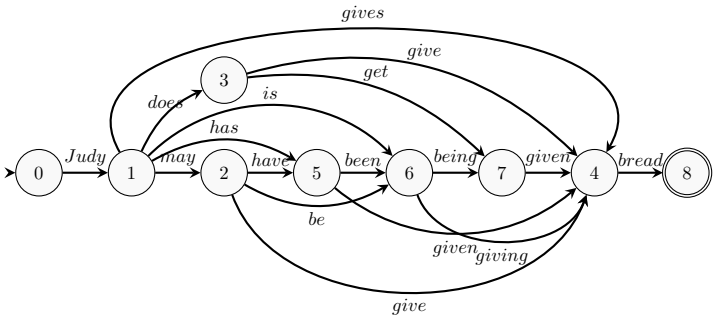
Perhaps the simplest kinds of languages there are are **finite languages**. These are sets which contain a finite number of elements and often are described just through listing the elements. The set of English words containing *ing* is one such example because there are a finite number of English words, so that there can only be finitely many containing *ing*. What is interesting, and perhaps not obvious at first, though, is that we are able to define *in*finite sets of strings by using finite resources. **Regular languages** (also called **finite-state languages**), can be generated by a computational device that contains a finite set of states. Strings in the language correspond to "walks" between these states. A simple example is,



This machine has two states, 0 and 1. When this device generates strings, one must start at the start state 0 (the starting state has an incoming arrow on the left) and then follow edges, while emitting the symbols ("a", "b", "c") that label those edges. For example, one could emit the string *acccb*, starting in state 0, looping three times in state 1, and then transitioning back to state 0. As should be clear from this example, there is no upper bound on the length of strings this machine can produce. Moreover, all strings follow a specific pattern—and that pattern is determined by the nodes, edges, and labels of the machine. For example, this machine could never emit a string with two consecutive *a*s or consecutive *b*s, because each *a* must be followed by either a *b* or a *c* according to the edges shown.

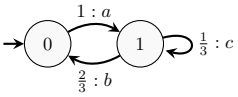
We note that formal presentations of finite state machines include several additional technical components, for example labeling of some states as "accepting" states (that must be the last state visited for a string to be valid), transitions between states that don't emit characters, as well as other variations like labeling states rather than edges with symbols (Hopcroft et al., 2001). Regular languages—the sets of strings generated by these machines—are often written in a notation known as **regular expressions**, which describe the strings generated. For example, if we require strings end with the machine in the 0 state, we could write the expression $(ac^*b)^*$ where x^* means that x (either a symbol or a sequence of symbols) can be repeated zero or more times. Thus, we always emit an a, any number of cs, and a b; and then we can do that whole thing any number of times.

Finite state machines can have any finite number of states. Here is a an example of the English system of auxiliary verbs, adapted from (Berwick & Pilato, 1987):



Walks on this graph provide acceptable strings of English, where for example we can say "Judy does get given bread" but not "Judy does have give bread."

Additionally, a common variant of finite state machines introduces *probabilities* on transitions (e.g. all outgoing edges must sum to 1), so that the machine generates a distribution on strings rather than a set. This is useful because it allows machines to generate probabilistic predictions about upcoming symbols.

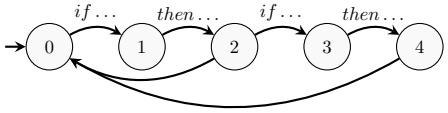


Here, we have annotated probabilities on each transition. Now, when we see an "a", it will be followed by a b on 2/3 of the times, and a c otherwise. Moreover, the number of cs we see in a row will be geometrically distributed since state 1 will have a 1/3 chance of generating "c" and returning to state 1. Such probabilistic machines – which we discussed in Chapter 5 in the guise of **hidden Markov models** – are amenable to a wide range of efficient algorithms for learning and inferring states from strings (Manning & Schütze, 1999). In turn, these devices which specify particular kinds of probability distributions on strings are widely applied throughout language technology, from speech recognition to statistical language analysis.

It's interesting to consider the range of possible uses for finite state machines: can *any* pattern on strings be captured with such a device? You might suspect the answer is yes because a finite state machine can have any (finite) number of states, so if we need more, we can always add them. However, Chomsky (1956) showed that there are intuitively simple languages that generate sentences that cannot be captured by *any* finite state machine (Chomsky, 1957). One example is the language $\{ab, aabb, aaabbb, aaaabbbb, \ldots\}$ consisting of n "a" symbols followed by exactly n "b" symbols, sometimes written as $a^n b^n$.

The key insight is that a finite machine with say k states can only "remember" k different values. The machine's knowledge about what characters it has produced, or can produce next, is entirely determined by the state that it is in, and there are only finite many states. However, $a^n b^n$ requires you to "remember" an unbounded number of values because n can be arbitrarily high. To put this a bit more formally, note that a path of length $n > k$ on a finite state machine *must* be in some state twice since each character requires us to visit a state. But if a path is in a state twice, that shows there must be a loop (a state where we can follow a sequence of states to get back to where we started). So, we could, in principle, follow that loop as many times as we want, generating acceptable strings with as many repetitions of that subsequence loop as we would like. Thus, if $n > k$, accepting the string $a^n b^n$ *also* means that the machine accepts strings with *other* numbers of as (since that sequence of as must be in some state twice, so we can follow that loop again if we like, as many times as we want). This shows that no finite state machine can accept exactly the language $a^n b^n$. This style of argument is called "pumping" and can be found in e.g. Hopcroft et al. (2001).

Chomsky (1956) argued that some structures in language, like *if-then* structures follow this pattern, where in English each "if" (think a) must be followed by exactly one "then" (think b). Accordingly, for example, we can say "**If if** coffee is good for you, **then** you'll drink it, **then** you are making sensible choices" (though, as the reader will be well aware, such sentences can be extremely difficult to understand and are, needless to say, vanishing rare). There is a sense in which it makes sense to imagine that *but for memory limitations* we could embed sentences indefinitely in English, yielding something like an $(\text{if})^n(\text{then})^n$ language, and thus English cannot be captured by a finite state machine. On the other hand, people *do* reliably break down after a few of these embeddings (e.g. $n \leq 2$) (Gibson & Thomas, 1999) which would indicate that these structures are possible to process with a finite number of states – for example,



There is much discord between psycholinguistic and linguistic theories whether we should formulate scientific accounts of language that capture what people actually do (break down for $n > 2$) or what people could "in principle" do (e.g. arbitrarily many embeddings). Advocates of the latter viewpoint typically believe that an abstract knowledge of the language can be separated from the limited and error-prone mental processing operations that deploy that knowledge to help understand and produce sentences–and it is this abstract knowledge that should be the focus of interest. By contrast, others doubt that such a separation is possible (or perhaps that the abstract structure of a language is a theoretical convenience for linguists, but has no corresponding mental representation)—and hence prioritize modeling what people actually say and understand.

16.3.2 Context-free grammars and beyond

One alternative to finite-state machines, outlined by Chomsky, was to consider systems of **rewrite rules**, a **grammar**, that work in a fundamentally different way from finite-state machines. Rewrite rules start with a symbol, typically S, and then specify rules for how S may be replaced by other symbols; and then rules for how those symbols may be replaced, as we saw in Figure 16.2. For example,

$$S \to aSb$$
$$S \to ab$$

says that any symbol S on the left can be replaced by either aSb (creating a new S) or ab (yielding no additional Ses). If we follow this we might get

$$S \to aSb \to aaSbb \to aaabbb$$

where we followed the first rule twice and then the second rule. You can see that this example generates exactly the language $a^n b^n$, which shows immediately that this system of rewrite rules is more powerful than finite state machines: it can model a language that no finite state machine can. As we note above, this kind of grammar is called a **context-free grammar**: the context-freeness arises because when you replace a symbol like S, it doesn't matter what other characters (context) are before and after it.

In fact, context-free grammars are *strictly* more powerful than finite state machines, meaning that any language you can describe with a finite state machine you are also able to describe with a system of rules. The proof of this idea is simple: you imagine that the symbols that get rewritten are the states

of the machine, and there is one rewrite rule for each edge, which gives the next state. For example, the finite state machine we started with might be,

$$0 \to a1$$
$$1 \to c1$$
$$1 \to b0$$

where following the rules of the grammar is identical to walking around states of the machine, with the current state represented as the last character (we may also need a rule like $0 \to \epsilon$, saying that the computation can end by yielding an empty string ϵ in state 0). Thus, finite-state languages are a subset of context-free languages (although the containment changes somewhat when probabilities are introduced).

Importantly, we may also form a **probabilistic context-free grammar** (as above) which assigns probabilities to each expansion, like

$$S \to aSb \qquad p = 0.4$$
$$S \to ab \qquad p = 0.6$$

This in turn specifies a distribution on strings which have a context-free structure (similar to those we saw in Figure 16.2). Systems with dozens or even hundreds of rules have often been used in natural language processing tasks, where the probabilities and the rules are fit to natural language usage.

Because such grammars can do everything a finite state machine can do, people talk about the Chomsky hierarchy of different kinds of computational devices: finite languages are a subset of finite-state languages, which are a subset of context-free languages. Further mathematical development uncovers other classes of languages. For example, $a^n b^n c^n$ can be shown to be not generated by any context-free grammar, but it can be generated by a **context-sensitive grammar** where the allowed rewrite rules depend on what other characters it is near. Other formal languages are not describable with even context-sensitive grammar, and eventually one ends up with computational systems that posses the full power of Turing machines. One such formalism focuses on allowing **transformations** of what has been produced by a phrase structure grammar, by moving and manipulating branches of the trees in systematic ways.

A considerable amount of work has tried to determine where natural language falls in the hierarchy and document structures that require different kinds of computational processes (Jäger & Rogers, 2012). Transformations were initially widely used in formal models of natural languages in linguistics (Chomsky, 1957), but their theoretical role has gradually reduced either to a single transformation (Chomsky, 1995) or none (Pollard & Sag, 1994; Steedman, 2000). It now is typically thought that language requires more than context-free power, but somewhat less than context-sensitive (e.g., Joshi, 1985)), and the reasons why our communication system should occupy that place in the hierarchy remain unclear. However, the characterization of human language in computational terms may depend on fairly philosophical question of how to handle people's finite memories. Should a theory of language capture what humans might do with unbounded memory (e.g. process (if)^n(then)^n for any n?). Or should our theory attempt to explain what people actually can do with their limited cognitive systems? After all, people's memory is finite and so the set of strings we can process is finite. But, at the same time, the regularities we can process seem well approximated by computational devices like grammars that seem very naturally unbounded, including hierarchical patterns in music, planning, problem-solving, and elsewhere.

16.3.3 Learnability and the Chomsky hierarchy

This brief overview of language structures and computation has highlighted some of the key conceptual tools mathematicians, computer scientists, and linguists have developed for thinking about patterns in sets of strings. Given these patterns, the natural question about human nature is whether or not the computations underlying natural language must be "built in" in some sense, or whether they could be

learned. Can a probabilistic approach really meet the challenge of learning the types of grammatical structures found in natural language, from reasonable small amounts of data, without prior assumptions so strong that they amount to a built-in universal grammar? Many theorists working on language acquisition have assumed not, and indeed of then the argument for linguistic nativism relies on related mathematical results.

Gold (1967) mathematically studies a learning situation where a learner sees strings from a formal language and must use the examples to infer the complete (usually, infinite) language (see Johnson, 2004). For example a learner might see the strings $\{aaaabbb, ab\}$ and infer $a^n b^n$. But Gold shows that even among regular languages, not all string sets could be discovered by learners who observe examples from the set of strings. That is, no matter *what* the learner does with data, there will be languages that they cannot learn. This has often been taken to indicate that children's learning space must be severely constrained by, for example, an innate grammar (Carnie, 2013). As a result, Gold's proof spurred development of complex theories of language learning under Gold-style assumptions (Wexler & Culicover, 1983), which often required learners to have a highly constrained set of hypotheses and transition between them in a particular order when data was observed.

However, for reasons of mathematical tractability, Gold's setup required assuming that parents (teachers) could provide maximally unhelpful examples to learners—i.e. it studies the learning situation in the *worst* case. The worst case is one where parents actively try to mislead children about the rules of language, and thus is likely not relevant to actual language acquisition. In milder formal setting, where learners observe strings sampled probabilistically from a target grammar, it can be shown that learners can learn the correct languages out of the space of all computations in theory (Chater & Vitányi, 2007). This work draws on prior idea in general inductive inference (Solomonoff, 1964), where essentially learners try to find concise programs to describe data they see. It can be shown that learns who do that will make optimal inferences about the structures in the world (Hutter, 2005). This work also helps address the no negative evidence problem that we outlined earlier in the chapter—at least with sufficient data and computational resources, it is possible to learn language from positive evidence alone.

Yang and Piantadosi (2022) implemented these ideas in a Bayesian inference model which observed strings and inferred programs that generate strings. They showed for example, that most of the simple formal languages used in linguistics and experimental learning studies could be discovered with a domain-general inference scheme that does Markov-Chain Monte-Carlo over programs (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Earlier work by, e.g., Elman (1990) and Christiansen and Chater (1999) using neural networks indicated that this is possible in principle, but that generalization is very limited, and no explicit representation of the grammatical regularities is created.

In Yang and Piantadosi (2022), learners were assumed to have access to a family of simple, domain-general computational primitives, like the ability to pair tokens together in a list, call probabilistic coin flips, recurse, etc. The model took some observed strings and inferred what program was likely to have generated the strings. For example, when given a few strings from $a^n b^n$, the model learns the program

$$F(x) := append(a, pair(if(flip(1/3), x, F(\epsilon)), b)).$$

This program will combine an a at the beginning of a string with a b at the end, and flip a coin to decide whether or not to recurse in the middle. Thus, it implements the probabilistic $S \to aSb$ grammar described above. Note that the probabilistic operation $flip$ is very important for this model because it allows the program to generate a distribution of outputs. In this probabilistic setting, straightforward Bayesian inference can compute the probability of a program h given data d as $P(h \mid d) \propto P(h)P(d \mid h)$, where $P(d \mid h)$ is the program H's probability of generating the data. As Yang and Piantadosi (2022) show, the model is able to construct programs that implement finite-state machines, context-free grammars, context-sensitive grammars and beyond. It constructs these programs as ways of explaining the data it observes, much like a scientist would formulate a computational theory to explain data they they see.

The question of whether or not the learned grammars are innate for the model is somewhat subtle. There is a sense in which the learned computational devices are innate for this model because they are constructed from built-in primitives and built-in rules of combination; but there is also an important sense in which the computational devices are constructed. In this model, *every possible* computation can be represented, so the model is in some sense maximally unconstrained in what it can learn—as Yang and Piantadosi (2022) argue, this builds in the *least* amount of information into the learning model. Perhaps analogously, when one opens a word processor, it is possible to write any book; and it takes real work to construct the best one.

Critically, the model does require children to have the ability to form program-like representations and evaluate them as probabilistic theories of data. Given the range of domains in which children learn new algorithmic structures (Rule, Tenenbaum, & Piantadosi, 2020)—from social rules, to arithmetic, to games—it is plausible that children deploy these general-purpose capabilities in language acquisition in order to represent the computational structures required for language.

16.4 Have deep neural networks solved the problem of processing and learning language?

The Bayesian approach to the acquisition of language aims to learn a model of the language from linguistic experience, and, as we have seen, a model of the language has typically been viewed as specified by abstract formal rules. We have seen that Yang and Piantadosi's work shows that patterns governed by these highly abstract rules can be learned from surprisingly small amounts of data using a general purpose probabilistic programming language–and we will see other examples of learning using probabilistic programs in Chapter 19. The high level of abstraction embodied in the rules of grammar outlined by Chomsky and learned successfully for a range of artificial languages without any built-in language-specific knowledge does not represent the only type of regularity in human language, however. Indeed, linguists and language acquisition researchers have increasingly been focusing on **construction grammars** (e.g., Goldberg, 2006; Tomasello, 2009) which capture the observation that language consists of a mix of regularity, subregularity, and outright exceptions, which seem better captured by a more flexible linguistic formalism (e.g., Dunn, 2017), consisting of constructions: linguistic patterns (at a range of level of abstraction) paired with their meanings. A long-term challenge for Bayesian models of grammar learning is to learn the patterns of highly abstract and very specific "constructions" that comprise a language, from naturally occurring corpora from that language. The promising results obtained so far indicate that this may be possible, without requiring the strong language-specific innate constraints that many authors have assumed to be required—i.e., without recourse to any innate universal grammar (as proposed by, e.g., Chomsky, 1980; Pinker, 1994; Crain & Lillo-Martin, 1999).

Recent developments in neural network models of language processing might, though, appear to suggest that the very attempt to build a specific model of language and language acquisition is unnecessary. In 2020, the results from the Large Language Model (LLM) GPT-3, a giant neural network with an extraordinary 175 billion trainable parameters (Brown et al., 2020), astonished many in both the academic community and the media (at the time of writing, GPT-4 is fast evolving, alongside many competing large language models, with some remarkable results, as we'll mention below—but let us focus on GPT-3 for the moment). GPT-3 was trained to predict the next items in a corpus, based on previous items—and the corpus was almost the entire contents of the World Wide Web, totally around a trillion words. The computer power required was correspondingly vast. GPT-3 had, of course, no symbolic representational language (whether a conventional grammar or logical formulae). Nonetheless, it is able to generate surprisingly convincing language, in any apparently highly flexible way. For example, seeded with the author Jerome K Jerome, the title "The importance of being on Twitter," and the first word "It," GPT-3 provided remarkably convincing opening paragraphs before drifting into incoherence (Klingeman, 2020).

Similarly, seeded with the identity of the philosopher David Chalmers and a series of philosophical questions on consciousness (Shevlin, 2020), it provided brief responses giving a tolerable account of Chalmers's views.

While impressive, a closer analysis of GPT-3's performance highlights that is better viewed as building a model of the language to which it has been exposed rather than a model of the world that the language describes. GPT-3 has been trained to map text onto text; and it does this in an astonishingly sophisticated way. But is it really learning about the world, or the nature of the mapping between language and the world, or indeed, the nature of communication itself. With GPT-3, caution seems appropriate. For example, it might seem that GPT-3 may provide insights into how children are able to learn the complex mixture of rules, sub-rules and exceptions that make up a human language. But this appearance may be somewhat misleading. From the point of view of construction grammar, learning a language involves acquiring collection of pairs of linguistic regularities and their meanings. But GPT-3 appears to have no representation of meaning, and no representation of syntactic structure onto which meaning may naturally be mapped (though see Pavlick, in press).

So despite its remarkable ability to write stories and answer questions, GPT-3 may have no real understanding of any of the language that it produces. Lacker (2020) nicely illustrates that while GPT-3 can answer an impressive range of general knowledge queries (it has almost the entire contents of the Web at its disposal, after all), it generates complete bizarre answers when asked bizarre questions: *Q: How many eyes does my foot have? A: Your foot has two eyes.* And it happily gives nonsense answers when asked nonsense questions: *Q: How many rainbows does it take to jump from Hawaii to seventeen? A: It takes two rainbows to jump from Hawaii to seventeen.* GPT-3 is flummoxed here, presumably, because these bizarre strings of words don't occur on the Web. This suggests that the vastness of the corpus used by GPT-3 is probably crucial to its success—it generalizes successfully on where its data is rich. Human children learn language from exposure only to tens of millions of words, along with the pragmatic and physical contexts in which those words are spoken—rich symbolic representations of the structure of language, and the structure of the world, may be critical to generalizing so effectively.[12]

But what about successors to GTP-3? How will this picture change as large language models become increasingly sophisticated? At the time of writing, later iterations of GPT, including the rapidly evolving GPT-4, have been trained (with the assistance of human users) to block many nonsensical responses and more generally to tune out unacceptable outputs, which are otherwise difficult to avoid give that the training corpus is the largely unfiltered contents of the internet (Ouyang et al., 2022). Moreover, the performance of LLMs continue to develop rapidly, including building surprisingly good links betwen language and visual images, and showing what some have described as "sparks" of general intelligence (Bubeck et al., 2023), including high-end human-level performance in IQ tests, and academic and professional exams.

How far such models will progress over the coming years is, of course, difficult to predict—but the kinds or errors GPT-3 makes should at least give us pause. They suggest that large language models may have a strategy for mimicking intelligent behavior by mining vast quantities of data which may rely on a shallower analysis of language and the world that we, as humans, might imagine.[13] On the other hand, it is possible that further iterations of large language models are leading those models to a fundamentally different, and less shallow, mode of operation, which do indeed involve building rich and flexible representations of the world, at least to some degree. Indeed, it is also possible that matching human intelligence requires some combination of two different types of process: general but fairly shallow representations of language and the world (through the analysis of vast quantities of linguistic and sensory input), alongside limited capacity mental operations over explicit symbolic representations, which may

[12]For further discussion, see the Epilogue in Christiansen and Chater (2022), Pavlick (in press), and Schultz and Frank (2023).

[13]Although, of course, some aspects of human cognition may themselves operate on shallower representations of the world that we imagine (Chater, 2018).

be required in deliberative reasoning and planning, especially in highly novel contexts.

In the context of language research, we can see Yang and Piantadosi's work, and LLMs such as GPT-3 and GPT-4, as representing the opposite ends of a continuum of approaches to learning language. Yang and Piantadosi use Bayesian inference over rich general-purpose logical representations, and obtain high levels of generalizations from relatively small training sets, albeit in simple artificial languages. LLMs involve training very large neural networks using non-symbolic representations of linguistic input, generalizing much less from the inputs it is given, but working successfully with a sufficiently large training corpus that only modest generalization is often sufficient. It is likely that future cognitive models of grammar learning may require combining insights from both approaches (McCoy & Griffiths, 2023). Moreover, we suspect that successful models of language acquisition will need to capture the fact that use language effectively involves the skilful use of communicative signals, including the repertoire of complex symbols comprising human languages, in social interactions (e.g., Christiansen & Chater, 2022). In short, models of language acquisition will need to see the learning of grammar as part of a wider process of learning to communicate and socially interact (e.g., Chater, McCauley, & Christiansen, 2016; Clark, 2009, 1996).

16.5 Future directions

The last few decades have seen remarkable strides in reverse engineering the human language acquisition and processing system, in parallel with a rapidly changing understanding of the nature of language itself. But, of course, as in other areas of cognition, the astonishing performance of human brain far outstrips any computational model that has yet been devised, whether from the perspective of Bayesian models, deep neural networks, or classical symbolic conversational models, or any combination of these. AI models of various kinds may perhaps be acquiring some "sparks" of general intelligence—but they still appear far from the rich, creative and flexible performance achieved by the human mind.

One theme that we see is likely to play an increasing role in future work is understanding the rich pragmatic inferences that allow even the simplest communicative signals (including pointing, gestures and facial expressions, aside from full-blown language) successfully to convey information on a highly flexible way, depending on recent discourse context, past interactions of the speaker and listener, the current goals and environment, and background knowledge of every kind (Galantucci, 2005; Sperber, 1985). More broadly, seeing language not just as a complex system of interlocking patterns (principles of phonology, morphology, prosody, syntax and so on), but as a set of tools for guiding human interaction, may be productive. This will require, of course, a much better understanding of the interface between language, social interaction, and thought.

A second, and related, theme is the interactive, conversational nature of most language (Clark, 1996; Pickering & Garrod, 2021). Researchers have often treated languages monologue first, and dialogue second; but recent developments in the language sciences suggests that the reverse may be the case. Indeed, this perspective may also be crucial for acquisition: children appear primarily to learn language through interactions with caregivers and other children rather than mere exposure to linguistic input (e.g., from the TV or radio). Language acquisition is, primarily, the ability to acquire a skill which allows children to join in with the many and varied conversational games played by those around them. Reverse engineering the nature of this skill, and how it can be learned from realistic amounts of linguistic interaction (measured in the tens of millions of words, rather than the trillion or so used by typically LLMs) is clearly a huge challenge. As we saw in Chapter 1, a crucial question for cognitive science is how the human brain is able to create so much from so little.

A third area of future development may, we suggest, be to understand how language has itself been shaped by its function in social interactions and underlying computational machinery recruited by the brain to achieve that function (Christiansen & Chater, 2016a; Kirby & Tamariz, 2021). Like any cultural

product, human languages will be strongly shaped by the nature of biases of our cognitive systems (and, of course, the perceptual and motor machinery through which speech and signs are generated and perceived) – a Bayesian analysis of mechanisms of cultural transmission that could have this kind of effect (Griffiths & Kalish, 2007) appears in Chapter 11. Reverse-engineering the language system successfully therefore should cast light on how a child's language changes through development, languages gradually change over time (through, for example, process of grammaticalization), and how over the long term, how language has evolved. Language is shaped to fit with human processing and learning biases (rather than being a purely abstract mathematical system), and this may greatly ease the problem of acquisition. The guesses that the child makes that structured language will tend to be correct precisely because the language has been shaped by similar guesses by past learners (Chater & Christiansen, 2010; Zuidema, 2002).

Finally, reverse-engineering the language system successfully requires explaining how the computational processes underlying language are rooted in neural hardware. A Bayesian perspective on language acquisition, and the success of computational models which learn linguistic structure from experience, has suggested that poverty of the stimulus arguments for the necessity of an innate universal grammar may be unpersuasive. But the problem of explaining how neural machinery is recruited to the challenge of acquiring the skill of conversational interaction with others, from experience, and language is processed with such nuance and subtlety in real time remains formidable. More broadly, though, while the challenges of understanding human language remain substantial, progress in computational modeling, together with important developments in linguistics and cognitive neuroscience, suggest that real, and rapid, advances may be possible.

16.6 Conclusion

Language is perhaps humanity's most remarkable, and far reaching, collective invention. It underpins our ability to formulate, transmit and record knowledge; to work together over long periods on complex joint projects; to create and enforce legal and moral codes; to invent ideologies, religions and scientific theories; and to create social, economic and technological worlds of astonishing flexibility and richness. Yet the fact that language is a structured symbolic system has sometimes led researchers to conclude to probabilistic ideas can be of no more than marginal interest in understanding language. But throughout this book, we've seen that symbolic and probabilistic ideas are better viewed as complementary rather than competing. The study of language illustrates how productive this complementarity can be: showing that the processing of language at every scale, from recognizing a word, to parsing a sentence, to acquiring a language, requires rich probabilistic inference over sophisticated structured linguistic representations.

References

Abend, O., Kwiatkowski, T., Smith, N., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Alishahi, A., & Stevenson, S. (2008). A computational model for early argument structure acquisition. *Cognitive Science*, *32*(5), 789-834.

Baker, C. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533–581.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284-17289.

Bergen, L., Gibson, E., & O'Donnell, T. (2013). Arguments and modifiers from the learner's perspective. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics.*

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Berwick, R. C., & Pilato, S. (1987). Learning syntax by automata induction. *Machine learning*, *2*(1), 9–38.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language.* Wiley.

Borschinger, B., Demuth, K., & Johnson, M. (2012). Studying the effect of input size for Bayesian word segmentation on the Providence corpus. In *Proceedings of COLING* (p. 325-340).

Borschinger, B., & Johnson, M. (2011). A particle filter algorithm for Bayesian word segmentation. In *Proceedings of the Australasian Language Technology Association Workshop* (p. 10-18).

Borschinger, B., & Johnson, M. (2014). Exploring the role of stress in Bayesian word segmentaiton using adaptor grammars. In S. Riezler (Ed.), *Transactions of the Association for Computational Linguistics* (Vol. 2, p. 93-104).

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33* (pp. 1877–1901).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al.. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Carnie, A. (2013). *Syntax: A generative introduction.* Wiley.

Chater, N. (2018). *The mind is flat: the illusion of mental depth and the improvised mind.* Penguin.

Chater, N. (2023). How could we make a social robot? a virtual bargaining approach. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220040.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*(7), 1131–1157.

Chater, N., Clark, A., Goldsmith, J., & Perfors, A. (2015). *Empiricism and language learnability.* Oxford University Press.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 335-344.

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language*, *89*, 244–254.

Chater, N., & Misyak, J. (2021). Spontaneous communicative conventions through virtual bargaining. In S. Muggleton & N. Chater (Eds.), *Human-like machine intelligence* (p. 52-67). Oxford University Press.

Chater, N., & Vitányi, P. M. (2007). 'Ideal learning' of natural language: Positive results about learning from positive evidence. *Journal of Mathematical psychology*, *51*(3), 135–163.

Chater, N., Zeitoun, H., & Melkonyan, T. (2022). The paradox of social interaction: Shared intentionality, we-reasoning and virtual bargaining. *Psychological Review*.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*(3), 113–124.

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

Chomsky, N. (1969). Quine's empirical assumptions. In *Words and objections* (pp. 53–68). Springer.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(1), 1–15.

Chomsky, N. (1986). *Language and problems of knowledge: The Managua lectures.* MIT Press.

Chomsky, N. (1995). *The minimalist program.* MIT Press.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of child language*, *30*(3), 637–669.

Christiansen, M. H., & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing.* MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Christiansen, M. H., & Chater, N. (2022). *The language game: How improvisation created language and changed the world.* Hachette.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368–407.

Clark, E. V. (2009). *First language acquisition.* Cambridge University Press.

Clark, H. H. (1996). *Using language.* Cambridge University Press.

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large language models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), e13256.

Crain, S., & Lillo-Martin, D. C. (1999). *An introduction to linguistic theory and language acquisition.* Oxford University Press.

Desmet, T., De Baecke, C., Drieghe, D., Brysbaert, M., & Vonk, W. (2006). Relative clause attachment in dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes*, *21*(4), 453–485.

Desmet, T., & Gibson, E. (2003). Disambiguation preferences and corpus frequencies in noun phrase conjunction. *Journal of Memory and Language*, *49*(3), 353–374.

Doyle, G., & Levy, R. (2013). Combining multiple information types in Bayesian word segmentation. In *Proceedings of NAACL-HIT* (p. 117-126).

Dunn, J. (2017). Computational learning of construction grammars. *Language and cognition*, *9*(2), 254–292.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory.* MIT Press.

Feldman, N., Goldwater, S., Griffiths, T. L., & Morgan, J. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752-782.

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, *30*(1), 3–20.

Fodor, B. T. G., J. A., & Garrett, M. F. (1974). *The psychology of language: An introduction to psycholinguistics and generative grammar.* McGraw-Hill.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science*, *33*, 287–300.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107-125.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360-371.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

Futrelle, R., Albright, P., A an Graff, & O'Donnell, T. (2013). A generative model of phonotactics. In *Transactions of the Association for Computational Linguistics.*

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14*(3), 225–248.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.* Oxford University Press.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Grice, H. P. (1975). Logic and conversation. In D. Davidson & G. Harman (Eds.), *The logic of grammar* (p. 64–75). Dickenson.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Hall, K., Hume, E., Jaeger, F., & Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, *4*.

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory*, *8*(2), 155–159.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown & Hanlon revisited: Mothers' sensitivity to ungrammatical forms. *Journal of child language*, *11*(1), 81–88.

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2001). Introduction to automata theory, languages, and computation. *Acm Sigact News*, *32*(1), 60–65.

Horwich, P. (1982). *Probability and evidence.* Cambridge: Cambridge University Press.

Hsu, A., & Griffiths, T. L. (2009). Differential use of implicit negative evidence in generative and discriminative language learning. In *Advances in Neural Information Processing Systems 22* (p. 754-762).

Hsu, A. S., & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science*, *34*(6), 972–1016.

Hsu, A. S., Chater, N., & Vitányi, P. M. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*, *120*(3), 380–390.

Hutter, M. (2005). *Universal artificial intelligence.* Springer.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, *97*(1), 553-562.

Jäger, G., & Rogers, J. (2012). Formal language theory: Refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1956–1970.

Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, *71*(4), 571–592.

Johnson, M., Demuth, K., & Frank, M. C. (2012). Exploiting social information in grounded language learning via grammatical reductions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.*

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19.*

Joshi, A. K. (1985). Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing* (pp. 206–250). Cambridge University Press.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics.* MIT Press.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Prentice Hall.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kirby, S., & Tamariz, M. (2021). Cumulative cultural evolution, population structure, and the origin of combinatoriality in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*, 20200319.

Kleinschmidt, D., & Jaeger, F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Klingeman, M. (2020). *An imaginary Jerome K. Jerome writes about Twitter.* (https://twitter.com/quasimondo/status/1284509525500989445)

Kronrod, Y., Coppess, E., & Feldman, N. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin and Review*, *23*(6), 1681–1712.

Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831-843.

Kurumada, C., Meylan, S., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*, 439-453.

Lacker, K. (2020). *Giving GPT-3 a Turing Test.*

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*, 1166–1183.

Luong, M.-T., Frank, M. C., & Johnson, M. (2013). Parsing entire discourses as very long strings: Capturing topic continuity in grounded language learning. In *Transactions of the Association for Computational Linguistics* (p. 315-326).

Luong, M.-T., O'Donnell, T., & Goodman, N. (2015). Evaluating models of computation and storage in human sentence processing. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 14-21.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978.

McCoy, R. T., & Griffiths, T. L. (2023). Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *arXiv preprint arXiv:2305.14701.*

Miller, G. A. (1951). *Language and communication.* McGraw-Hill.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science*, *27*(12), 1550–1561.

Norvig, P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance*, *9*(4), 30–33.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage.* MIT Press.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pajak, B., & Levy, R. (2014). The role of abstraction in non-native speech perception. *Journal of Phonetics*, *46*, 147–160.

Pavlick, E. (in press). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A.*

Pearl, L. (2021). Modeling syntactic acquisition. In J. Sprouse (Ed.), *Oxford handbook of experimental syntax.* Oxford University Press.

Pearl, L., & Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, *20*(1), 23–68.

Pearl, L., & Sprouse, J. (2019). Comparing solutions to the linking problem using an integrated quantitative framework of language acquisition. *Language*, *95*(4), 583–611.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306-338.

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.

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*, 1-31.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329–347.

Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction.* Cambridge University Press.

Pinker, S. (1994). *The language instinct: How the mind creates language.* Harper Collins.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar.* University of Chicago Press.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915.

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

Schultz, T., & Frank, M. C. (2023). Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Shevlin, H. (2020). *A #gpt3 interview about AI with an imaginary David Chalmers.* (`https://twitter.com/dioscuri/status/1285385825971245058?lang=en`)

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control*, *7*(1), 1–22.

Sperber, D. (1985). Anthropology and psychology: Towards an epidemiology of representations. *Man*, *20*, 73-89.

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., Kleiman-Weiner, M., & Gao, T. (2021). Modeling communication to coordinate perspectives in cooperation. *arXiv preprint, arXiv:2106.02164*.

Steedman, M. (2000). *The syntactic process.* MIT Press.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466.

Synnaeve, G., Dautriche, I., Börschinger, B., Johnson, M., & Dupoux, E. (2014). Unsupervised word segmentation in context. In *Proceedings of COLING* (p. 2326-2334).

Tomasello, M. (2009). *Constructing a language.* Harvard University Press.

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences*, *18*(11), 605–611.

Vallabha, G., & McClelland, J. (2007). Success and failture of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, and Behavioral Neuroscience*, *7*, 53-73.

Wang, R., Wu, S., Evans, J., Tenenbaum, J., Parkes, D., & Kleiman-Weiner, M. (2021). Too many cooks: Coordinating multi-agent collaboration through inverse planning. In S. Muggleton & N. Chater (Eds.), *Human-like machine intelligence* (p. 152-170). Oxford University Press.

Wexler, K., & Culicover, P. (1983). *Formal principles of language acquisition.* MIT Press.

Wong, S.-M., Dras, M., & Johnson, M. (2012). Exploring adaptor grammars for native language identification. In *Proceedings of EMNLP/CoNLL* (p. 699-709).

Yang, Y., & Piantadosi, S. T. (2022). One model for the learning of language. *Proceedings of the National Academy of Sciences*.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zuidema, W. (2002). How the poverty of the stimulus solves the poverty of the stimulus. In *Advances in Neural Information Processing Systems 15.*