Computational Modeling as a Methodology for Studying Human Language Learning

Thierry Poibeau, Aline Villavicencio, Anna Korhonen and Afra Alishahi*

1 Overview

The nature and amount of information needed for learning a natural language, and the underlying mechanisms involved in this process, are the subject of much debate: how is the knowledge of language represented in the human brain? Is it possible to learn a language from usage data only, or is some sort of innate knowledge and/or bias needed to boost the process? Are different aspects of language learned in order? These are topics of interest to (psycho)linguists who study human language acquisition, as well as to computational linguists who develop the knowledge sources necessary for large-scale natural language processing systems. Children are the ultimate subjects of any study of language learnability. They learn language with ease, in a short period of time and their acquired knowledge of language is flexible and robust.

Human language acquisition has been studied for centuries, but using computational modeling for such studies is a relatively recent trend. However, computational approaches to language learning have become increasingly popular, mainly due to advances in developing machine learning techniques, and the availability of large collections of experimental data on child language learning and child-adult interaction. Many of the existing computational models attempt to study the complex task of learning a language under cognitively plausible criteria (such as memory and processing limitations that humans face), and to explain the developmental stages observed in children. By simulating the process of child language learning, computational models can show us which linguistic representations are learnable from the input that children have access to in a reasonable amount of time, and which mechanisms yield the same patterns of behaviour that children exhibit during this process. In doing so, computational modeling provides insight into the plausible

* Excerpts of this chapter have been published in Alishahi, A. (2010), *Computational Modeling of Human Language Acquisition*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

mechanisms involved in human language acquisition, and inspires the development of better language models and techniques.

The aim of this volume is to present a cross-section of recent research on the topic that draw on the relevance of computational techniques for understanding human language learning. These studies are inherently multidisciplinary, influenced by knowledge from fields such as Linguistics, Psycholinguistics and Biology, and the overview chapter starts with a discussion of some of the challenges faced, such as learnability constraints, data availability and cognitive plausibility. The strategies that have been adopted to deal with these problems build on recent advances in areas such as Natural Language Processing, Machine Learning, Artificial Intelligence and Complex Networks, as will be discussed in details in the chapters that compose this collection. Given the complex facets of language that need to be acquired, these investigations differ in terms of the particular language learning task that they target, and the overview chapter finishes with a contextualization of these contributions.

1.1 Theoretical accounts of language modularity and learnability

The study of human language acquisition pursues two important goals: first, to identify the processes and mechanisms involved in learning a language; and second, to detect common behavioural patterns in children during the course of language learning.

Languages are complex systems and learning one consists of many different aspects. Infants learn how to segment the speech signal that they receive as input, and they recognize the boundaries that distinguish each word in a sentence. They learn the phonology of their language, or the auditory building blocks which form an utterance and the allowable combinations which form individual words. They assign a meaning to each word form by detecting the referent object or concept that the word refers to. They learn the regulations that govern form, such as how to change the singular form of a noun into a plural form, or the present tense of a verb into the past tense. They learn how to put words together to construct a well-formed utterance for expressing their intention. They learn how to interpret the relational meaning that each sentence represents and how to link different sentences together. On top of all these, they learn how to bring their knowledge of concept relations, context, social conventions and visual clues into this interpretation process.

A central question in the study of language is how different aspects of linguistic knowledge are acquired, organized and processed by the speakers of a language. The useful boundaries that break the language faculty into separate "modules" such as word segmentation, phonology, morphology, syntax, semantics and pragmatics, have been historically imposed to facilitate the study of each of these aspects in isolation. However, later psycholinguistic studies on language acquisition and processing suggest that the information relevant to these modules is not acquired in a temporally linear order, and that there is close interaction between these modules during both the acquisition and processing of language. In addition, many of the

formalisms and processing techniques that have been proposed to handle a specific aspect may not be suitable for another.

The language modularity argument is part of a larger debate on the architecture of the brain, or the "modularity of mind." Proposals advocating a highly modular view rely extensively on the studies of Specific Language Impairments (SLI) which imply the isolation of language from other cognitive processes (e.g., 36), whereas a highly interactive views refer to more recent studies on the interaction of language and other modalities such as vision or gesture at the process level (see Visual World Paradigm, (61)).

The modularity debate has been highly interleaved with the issue of nativism or language innateness. On the topic of language, the main point of interest has been whether humans are equipped with a highly sophisticated module for learning and using natural languages, consisting of task-specific procedures and representations, the "language faculty" (9; 11). As complicated as it seems to master a language, children all around the world do it seemingly effortlessly and in a short period of time. They start uttering their first words around age one. By the time they are three to four-years old, they can use many words in various constructions, and can communicate fluently with other speakers of their native language. The efficiency with which children acquire language has raised speculations about whether they are born with some sort of innate knowledge which assists them in this process.

Human beings have an unparalleled skill for learning and using structurally complex languages for communication, and the learnability of natural languages has been one of the most controversial and widely discussed topics in the study of languages. The possibility of a genetic component that accounts for this unique ability of humans has been raised, but the extent and exact manifestation of this component is not clear. For instance, it has been argued that general learning and problem solving mechanisms are not enough to explain humans highly complex communication skills, and some innate knowledge is also needed to account for their exceptional linguistic skills (13; 26). This hypothesis, known as the Innateness Hypothesis, states that human beings have an innate predisposition for learning languages, a task or domain specific knowledge, defined by their genetic code, and without having access to such innately specified linguistic knowledge a child cannot learn a language. Indeed, a nativist view of language learning states that natural languages are not learnable from the linguistic data that is typically available to children (Primary Linguistic Data, or PLD). The main argument in support of this view is the Argument from the Poverty of the Stimulus (APS; 9), according to which PLD is both quantitatively and qualitatively too impoverished to allow for the acquisition of a natural language in its full structural complexity in a short period of time.

Of particular relevance to this discussion is the mathematical work of (28), who proved that a language learner cannot converge on the correct grammar from an infinitely large corpus without having access to substantial negative evidence. However, direct negative evidence (or corrective feedback from adult speakers of language) has been shown not to be a reliable source of information in child-directed

data (45; 44).[2] These findings have been viewed as compatible with nativist proposals for language acquisition such as that of a Universal Grammar (UG) (12), proposing that each infant is born with a innately specified representation of a grammar which determines the universal structure of a natural language. This universal grammar would be augmented by a set of parameters, which have to be adjusted over time to account for the particular language a child is exposed to.

In response to the nativist view of language learning, alternative representations of linguistic knowledge have been proposed, and various statistical mechanisms have been developed for learning these representations from usage data. A more empiricist view of language learning argues that a child does not have any innate prior knowledge about languages, and that languages can be learned using only general cognitive abilities which also underly other tasks (e.g. imitation, categorization and generalization (63; 64)) when these are applied to the sensory input to which a child is exposed. In an extreme version of empiricism, a child is like a *tabula rasa*, or a blank slate, when born, and all its language capabilities are learned from scratch from the environment (54).

Analyses of large collections of data on child-parent interactions have raised questions about the inadequacy of PLD (54; 35), arguing that child-directed data provides rich statistical cues about the abstract structures and regularities of language. In addition, recent psycholinguistic findings which hint at a 'bottom-up' process of child language acquisition have also questioned the top-down, parameter-setting approach advocated by the nativists. Advocates of this alternative view of language learning, also referred to as the *usage-based*, claim that children do not possess highly detailed linguistic knowledge at birth; instead they learn a language from the usage data they receive during the course of learning. Usage-based theories of language acquisition are motivated by experimental studies on language comprehension and generation in young children that suggest that children build their linguistic knowledge around individual items (38; 39; 3; 1; 63). This view asserts that young children initially learn verbs and their arguments as lexical constructions and on an item-by-item basis, and only later begin to generalize the patterns they have learned from one verb to another. However, the details of the acquisition of these constructions and the constraints that govern their use are not clearly specified. Explicit models must be explored, both of the underlying mechanisms of learning these regularities, and of the use of the acquired knowledge.

In sum while nativism emphasises the role of *nature* as providing the required equipment, empiricism emphasises the role of *nurture* assuming that the environment is rich enough to provide a child with all the necessary evidence for language acquisition. Different proposals vary in terms of the extent in which they rely on language specific mechanisms and on general-purpose skills.

[2] On the other hand, it has been suggested that the language learner can estimate the "typical" rate of generalization for each syntactic form, whose distribution serves as "indirect" negative evidence (42; 14)

1.2 Investigations of linguistic hypotheses

One fundamental difficulty in research on language acquisition is that due to its characteristics it has to rely on indirect observation about the target processes. Apart from ethical considerations, the lack of non-invasive technology that is able to capture acquisition in action over time means that researchers can only assess different hypotheses indirectly, e.g. through diaries of child language, corpora of child-directed speech, or psycholinguistic data. As an strategy for probing human behaviour when learning and processing language, psycholinguistics provides a variety of experimental methodologies for studying specific behavioural patterns in controlled settings. Evidence concerning what humans (and children in particular) know about language and how they use it can be obtained using a variety of experimental techniques. Behavioural methods of studying language can be divided into two main groups: *offline* techniques, which aim at evaluating subjects' interpretation of a written or uttered sentence *after* the sentence is processed; and *online* techniques, which monitor the process of analyzing linguistic input *while* receiving the stimuli.

In offline studies, child language processing is examined in an experimental setup using interactive methods in the form of *act-out scenarios* (when the experimenter describes an event and asks the child to act it out using a set of toys and objects), or *elicitation tasks* (when the child is persuaded to describe an event or action in the form of a natural language sentence). Preferential looking studies are another experimental approach conducted mostly on young children, where their preferences for certain objects or scene depictions is monitored while presenting them with linguistic stimuli.

In online methodologies, a variety of techniques are used (mostly on adult subjects) for identifying processing difficulties. A common technique in this category is measuring *reading times*. Many factors can affect reading times, therefore psycholinguistic studies use stimuli which are different in one aspect and similar in the others, and measure the reading time of each group of stimuli. Another technique that can be used on children as well as adult subjects is *eye-tracking*, where eye movements and fixations are spatially and temporally recorded while the subjects read a sentence on the screen. Using this technique, several reading time measures can be computed to evaluate processing difficulties at different points in the sentence. Also, anticipatory eye-movements can be analyzed to infer interpretations. Eye-tracking techniques have been employed in the *Visual World Paradigm* (61), where subjects' eye movements to visual stimuli are monitored as they listen to an unfolding utterance. Using this paradigm, the construction of online interpretation of a sentence and its mapping to the objects in the visual environment in real time can be studied.

More recently, *neuroscientific* methods have also been used for studying the processing of language in the brain. The most common approach is to measure *event-related potentials* (ERP) via electroencephalography (EEG): a stimulus is presented to the subject, while ERPs are measured through electrodes positioned on the scalp. Robust patterns have been observed in the change of ERPs as a response to linguis-

tic stimuli. For example, when presented with a sentence with a semantic anomaly (e.g., *I like my coffee with cream and dog*), a negative deflection is usually observed 400 milliseconds after the presentation of the stimuli. However, it is difficult to isolate the brain response to a particular stimulus, and it has been a challenge to derive a detailed account of language processing from such data. Functional Magnetic Resonance Imaging (fMRI) is another technique for measuring neural activity in the brain as a response to stimuli. Unlike EEG, fMRI cannot be used as an online measure, but it has higher spatial resolution and provides more accurate and reliable results.

In the majority of experimental studies of language, one aspect or property of the task or stimuli is selected and manipulated while other factors are held constant, and the effect of the manipulated condition is investigated among a large group of subjects. This approach allows researchers to isolate different language-related factors, and examine the significance of the impact that each factor might have on processing linguistic data. In such set-ups, it is only possible to manipulate the properties of the input data and the task in hand, and the learning or processing mechanisms that the subjects use for performing the task remain out of reach. Moreover, each subject has a history of learning and processing language which cannot be controlled or changed by the experimenter: all there is to control is a time-limited experimental session. Artificial languages are used to overcome any interference that the subjects' previous language-related experience might have on the outcome of the experiment. But the amount of the artificial input data that each subject can receive and process in these settings is very limited. These shortcomings call for an alternative approach for investigating the hypotheses regarding the acquisition and processing of natural languages.

2 Computational models of language learning

Over the past decades, computational modeling has been used extensively as a powerful tool for in-depth investigation of existing theories of language acquisition and processing, and for proposing plausible learning mechanisms that can explain various observed patterns in child experimental data. The use of computational tools for studying language dates back to the onset of Artificial Intelligence. Early models mostly used logic rules for defining natural language grammars, and inference engines for learning those rules from input data. Over the last twenty years a rapid progress in the development of statistical machine learning techniques has resulted in the emergence of a wider range of computational models that are much more powerful and robust than their predecessors. As a result, computational modeling is now one of the main methodologies in the study of human cognitive processes, and in particular language.

Using computational tools for studying language requires a detailed specification of the properties of the input data that the language learner receives, and the

mechanisms that are used for processing the data. This transparency offers many methodological advantages, such as:

- **Explicit definition of assumptions:** when implementing a computational model, every assumption, bias or constraint about the characteristics of the input data and the learning mechanism has to be specified. This property distinguishes a computational model from a linguistic theory, which normally deals with higher-level routines and does not delve into details, a fact that makes such theories hard to evaluate.
- **Control over input data:** unlike an experimental study on a human subject, the researcher has full control over all the input data that the model receives in its life time. This property allows for a precise analysis of the impact of the input on the behaviour of the model.
- **Control over experimental variables:** when running simulations of a model, the impact of every factor in the input or the learning process can be directly studied in the output (i.e., the behaviour) of the model. Therefore, various aspects of the learning mechanism can be modified and the behavioural patterns that these changes yield can be studied.
- **Choice of learning mechanisms:** the performance of two different mechanisms on the same data set can be compared against each other, something that is almost impossible to achieve in an experimental study on children.
- **Access to predictions of the model:** because of the convenience and the flexibility that computational modeling offers, novel situations or combinations of data can be simulated and their effect on the model can be investigated. This approach can lead to novel predictions about learning conditions which have not been previously studied.

Despite these advantages, computational modeling should not be viewed as a substitute but rather as a complement for theoretical or empirical studies of language. One should be cautious when interpreting the outcome of a computational model: if carefully designed and evaluated, computational models can show what type of linguistic knowledge is learnable from what input data. Also, they can demonstrate that certain learning mechanisms result in behavioural patterns that are more in line with those of children. In other words, computational modeling can give us insight on which representations and processes are more plausible in light of the experimental findings on child language acquisition. They can be viewed as the testing grounds for different theories and can provide information about the conditions under which these would succeed in a given task. However, even the most successful computational models can hardly prove that humans exploit a certain strategy or technique when learning a language. Cognitive scientists can use the outcome of computational modeling as evidence on what is possible and what is plausible, and verify the suggestions and predictions made by models through further experimental and neurological studies.

2.1 What to expect from a model

Traditionally, linguistic studies of language have been focused on representational frameworks which can precisely and parsimoniously formalize a natural language according to how adult speakers of that language use it. In this approach, the focus is on the end product of the acquisition process, and not on the process itself. On the other hand, psycholinguistic studies mainly emphasize the process of learning and using a language rather than the language itself (15). This dual approach is also reflected in the modeling of language acquisition. One modeling strategy is to demonstrate the feasibility of extracting an optimal structure from a given linguistic input (e.g., a grammar from a text corpus, or a phonetic or lexical segmentation from a large stream of speech signals), aiming at compatibility of the results with a target. An alternative strategy is to focus on developmental compatibility and replicate the stages that children go through while learning a specific aspect of language, such as vocabulary growth in word learning or the U-shaped generalization curve in the acquisition of verb argument structure. Therefore, given the priorities of each of these strategies it is important to evaluate a model in the context that it is developed in, and with respect to the goals that it is aiming at.

Another critical point when assessing a model is to identify the fundamental assumptions that the model is based on. When developing a model for computational simulation of a process, all the details of the process must be implemented, and no trivial aspect of the representational framework or the procedure can be left unspecified. However, many of these details are of secondary importance to the process that the model aims to study. It is of utter importance for the developers of a computational model to clearly specify which theoretical assumptions about the implemented model or the characteristics of the input data are fundamental, and which implementation decisions are arbitrary. Moreover, they must show that the overall performance of the model does not crucially depend on these trivial decisions.

Finally, the level of processing targeted by a model must also be taken into account. One of the first (and most influential) categorizations of cognitive models was proposed by (47), who identifies three levels of describing a cognitive process. First is the *Computational* level, which identifies what knowledge is computed during the process. This is the highest level a model can aim for: the focus is on what is needed or produced during the cognitive process under study, abstracting from any learning or processing algorithm that is used for computing or applying this knowledge. Next comes the *Algorithmic* level, which specifies how computation takes place. At this level, the focus is on the mechanisms involved in the computational process. Finally there is the *Implementation* level, which simulates how the algorithms are actually realized in brain. At this level, every implementational detail is a vital component of the model. It is important on the modelers' side to specify, and on the evaluators' side to take into account, the intended level of the model to be assessed. If the simulation of a model aimed at a computational level of describing a process results in a behavioural pattern that is inconsistent with that of children, it might be due to an inappropriate choice of algorithm or other implementational details, and not because the specification of the proposed computation itself is flawed.

One important constraint when developing a cognitive model is *cognitive plausibility*. In the field of natural language processing, many automatic techniques have been developed over the years for extracting various types of linguistic knowledge from large collections of text and speech, and for applying such knowledge to different tasks. In this line of research, the main goal is to perform the task at hand as efficiently and accurately as possible. Therefore, any implementation decision that results in better performance is desired. For instance, to induce wide coverage grammars from corpora, supervised learning methods based on annotated data such as the Penn Treebank have been usually employed, with grammars that tend to be less than or equal to context free grammars in expressive power and which may not be linguistically adequate to capture human grammar (60). However, cognitive models of language learning and processing are not motivated by improving performance on a certain task. Instead, they are aiming at simulating and explaining how humans perform that task. Such models have to conform to the limitations that humans are subject to.

A model which attempts to simulate a cognitive process has to make realistic assumptions about the properties of the input data that are available to children during that process. For example, a model of syntax acquisition cannot assume that children are being corrected when producing an ungrammatical sentence, since various analyses of child-directed data have shown that such information is not consistently provided to them (44). Also, when modeling any aspect of child language acquisition, it cannot be assumed that children receive *clean* input data, since the data almost always contain a high level of noise and ambiguity. Sometimes it is inevitable to make simplifying assumptions about the structure of data in order to keep calculations feasible or to focus on one specific aspect of learning. However, if a model makes obviously false assumptions about the input, any finding by such a model might not be generalizable to realistic situations.

Also, a cognitive model must draw on language-independent strategies. Children around the world learn a variety of languages with drastically different characteristics, such as their sound system or structure. It is highly implausible to assume that children use different learning mechanisms for learning different languages. Thus a model of language learning must avoid any language-specific assumptions or learning strategies. For example, a model of learning syntax which assumes a rigid word order cannot be extended to families of languages with a more relaxed word order.

Finally, cognitive models must conform to the memory and processing limitations of humans. The architecture of the human brain and its processing capacities and memory resources are very different from those of the existing computational systems. Thus many of the machine learning techniques that are developed for applying on large-scale data sets are not suitable for modeling human language processing. For example, it is unlikely that children can remember every instance of usage of a particular word or every sentence that they have heard during their lifetime in order to learn something about the properties of language. This limits the scope of the techniques and algorithms that can be used in cognitive modeling. One of the by-products of human memory and processing limitations is that language must be learned in an incremental fashion. Every piece of input data is processed

when received, and the knowledge of language is built and updated gradually. This is in contrast with many machine learning techniques which process large bodies of input at once (usually through iterative processing of data) and induce an optimum solution (e.g., a grammar) which formalizes the whole data set precisely and parsimoniously.

Although a cognitive model of language is often expected to provide a cognitively plausible explanation for a process, it is the intended description level of the model which determines the importance of various plausibility criteria. For example for a model at the computational level, making realistic assumptions about the characteristics of the input data is crucial. However, conforming to processing limitations (such as incrementality) in the implementation of the model is of secondary importance, since the model is not making any claims about the actual algorithm used for the proposed computation.

2.2 Modeling frameworks

The first generation of models of language were influenced by early artificial intelligence techniques, including the logic-based inference techniques which were widespread in 1960s. Symbolic modeling often refers to an explicit formalization of the representation and processing of language through a symbol processing system. In this approach, linguistic knowledge is represented as a set of symbols and their propositional relations. Processing and updating this knowledge takes place through general rules or schemas, restricted by a set of constraints. Each rule might be augmented by a list of exceptions, i.e., tokens or instances for which the rule is not applicable. The syntactic structure of a language is typically modeled as a rule-based grammar, whereas the knowledge of semantics is modeled through schemas and scripts for representing simple facts and processes. These representations are often augmented by a set of logical rules for combining them and constructing larger units which represent more complex events and scenarios.

Following the Chomskian linguistics tradition, symbolic models of language assume that a language is represented as an abstract rule-based grammar which specifies all (and only) valid sentences, based on judgements of linguistic acceptability (12). In this view, language processing is governed by internally specified principles and rules, and ambiguities are resolved using structural features of parse trees (e.g., the principle of minimal attachment; 24). The influence of lexical information on parsing and disambiguation is often overlooked by these theories. Language acquisition, on the other hand, has been mainly modeled through *trigger-based* models, where the parameters associated with a pre-specified grammar are set to account for the input linguistic data (e.g., 26).

Symbolic models of language are often transparent with respect to their linguistic basis, and are computationally well-understood. However, typical symbolic models do not account for the role of *experience* (or the statistical properties of the input data) on behaviour and are not robust against noise and ambiguity.

Connectionist models of cognitive processes (48) emerged during 1980s as an alternative to symbolic models. The architectural similarities between the connectionist models and the human brain on a superficial level, and their capacity for distributional representation and parallel processing of knowledge made them an appealing choice for modeling human language acquisition and processing. The idea of connectionist models is based on simple neural processing in brain. Each connectionist model (or *artificial neural network*) consists of many simple processing units (or *neurons*), usually organized in layers, which are heavily interconnected through weighted links. Each neuron can receive many input signals, process them and pass the resulting information to other neurons. Linguistic knowledge is represented as distributed activation patterns over many neurons and the strength of the connections between them. Learning takes place when connection weights between neurons change over time to improve the performance of the model in a certain task, and to reduce the overall error rate. A cognitive process is modeled by a large number of neurons performing these basic computations in parallel.

Various versions of artificial neural networks have been proposed which vary in the neuron activation function, the architecture of the network, and the training regime. For modeling language learning, multi-layered, feed-forward networks have been most commonly used. These networks consist of several neurons, arranged in layers. The input and output of the cognitive process under study are represented as numerical vectors, whose dimensions correspond to input units. Such models are normally trained in a supervised fashion: the model produces an output for a given input pattern, and the connection weights are adjusted based on the difference between the produced and the expected output.

Connectionist models have received enormous attention from the cognitive science community due to the learning flexibility they offer compared to symbolic models (e.g., 40; 20), and because they suggest that general knowledge of language can be learned from instances of usage. However, these models often cannot easily scale up to naturalistic data. Moreover, the knowledge they acquire is not transparent, and is therefore hard to interpret and evaluate.

The relatively recent development of machine learning techniques for processing language motivated many researchers to use these methods as an alternative modeling paradigm. Probabilistic modeling allows for combining the descriptive power and transparency of symbolic models with the flexibility and experience-based properties of the connectionist models. Probabilities are an essential tool for reasoning under uncertainty. In the context of studying language acquisition, probabilistic modeling has been widely used as an appropriate framework for developing experience-based models which draw on previous exposure to language, and at the same time provide a transparent and easy to interpret linguistic basis. Probabilistic modeling views human language processing as a rational process, where various pieces of evidence are weighted and combined through a principled algorithm to form hypotheses that explain data in an optimal way. This view assumes that a natural language can be represented as a probabilistic model which underlies sentence production and comprehension. Language acquisition thus involves constructing this probabilistic model from input data.

Many probabilistic models of language are essentially an augmented version of their symbolic counterparts, where each rule or schema is associated with a weight (or probability). For example, Probabilistic Context Free Grammars (PCFG) use a symbolic representation of the syntactic knowledge (CFG), but they also calculate a probability for each grammar rule depending on the number of times that rule has appeared in the input (32). However, an alternative (and more radical) probabilistic view proposes language represented as a bottom-up, graded mapping between the surface form and the underlying structure, which is gradually learned from exposure to input data (e.g., 16; 64).

The acquisition of linguistic knowledge can be formulated as an induction process, where the most likely underlying structures are selected based on the observed linguistic evidence. The basic idea behind this process is to break down complex probabilities into those that are easier to compute, often using Bayes' rule. A family of probabilistic models, generally referred to as Bayesian models, have gained popularity over the past decade. In the context of grammar learning, Bayesian methods specify a framework for integrating the prior information about the grammatical structures and the likelihood of the observed word strings associated with each structure, to infer the most probable grammatical structure from a sentence. The prior probabilities are often used for embedding underlying assumptions about the hypothesis space and for seemlessly integrating biases and constraints into the system. It has been argued that prior information (specifically the prior structure over Bayesian networks) is crucial to support learning (62). These positioning of these models in relation to nativism due to the nature of the prior information adopted remains to be determined. As (51) discuss, there seems to be an agreement along the empiricist-nativist continuum that there must be some innate constraints. Different proposals vary as to which constraints are adopted and how strong and domain-specific they are, given the empiricist ideal of a bottom-up, data-driven learning.

In addition to the probabilistic frameworks that are specifically developed for representing and processing linguistic knowledge, many recent computational models heavily rely on general-purpose statistical machine learning tools and techniques. A variety of such methods have been successfully exploited in more practical natural language processing applications. The efficiency of these methods has motivated their use in modeling human language acquisition and processing, in particular for the purpose of extracting abstract and high-level knowledge from large collections of data. In this context, one approach from Information Theory that has been adopted in various computational language acquisition tasks is the Minimum Description Length (MDL) Principle. MDL is an algorithmic paradigm for evaluating the hypothesis space, based on Occam's Razor, in which the best hypothesis for a given set of data is the one that leads to the best compression of the data (55). The idea is that MDL can be used to order the hypothesis space according to how compact the hypotheses are and how well they generate the data (37). MDL has proved to be a powerful tool in many language related tasks, such as word segmentation (e.g., 17; 4), grammar learning (e.g., 30; 33; 19; 66) and learnability assessment (e.g., 31).

Probabilistic models in general are robust against noise, and are a powerful tool for handling ambiguities. A range of statistical and probabilistic techniques have

been efficiently employed over the last couple of decades to modeling various aspects of language acquisition and use, some examples of which can be seen in the papers in this collection. However, some suggest that probabilistic methods must be viewed as a framework for building and evaluating theories of language acquisition, rather than as embodying any particular theoretical viewpoint (8).

2.3 Research methods

As a response to the nativist claims that some aspects of language (mainly syntax) are not learnable solely from input data, a group of computational models have been proposed to challenge this view and investigate to what extent extracting a grammatical representation of language from a large corpus is in fact possible. These models are not considered as typical cognitive models, since most of them are not concerned with how humans learn language. Instead, their goal is to show that the Primary Linguistic Data is rich enough for an (often statistical) machine learning technique to extract a grammar from it with high precision and without embedding any innate knowledge of grammar into the system. On the other hand, a typical cognitive model cannot be solely evaluated based on its accuracy in performing a task. The behaviour of the model must be compared against observed human behaviour, the errors made by humans must be replicated and explained, and the result must be linguistically and psychologically motivated. Therefore, evaluation of cognitive models depends highly on the experimental studies of language.

We need to compare the knowledge of a cognitive model to that of humans in a particular domain. But there is no direct way to figure out what humans *know* about language. Instead, their knowledge of language can only be estimated or evaluated through how they *use* it in language processing and understanding, as well as in language production. Analysis of child production data provides valuable cues about the trajectory of their learning the language. Many developmental patterns are revealed through studying the complexity of the utterances that children produce, the errors that they make and the timeline of their recovery from these errors. On the other hand, comprehension experiments reveal information about knowledge sources that children exploit, their biases towards linguistic and non-linguistic cues, and their awareness of the association between certain utterances and events.

Earlier studies of child language acquisition were based on sporadic records of interaction with children, or isolated utterances produced by children which researchers individually recorded. But recent decades have seen a significant growth in the variety and quantity of resources for studying language, and a collective attempt from the computational linguistics and cognitive science communities to use standard formats for the expansion of these resources.

The most well-known and widely used database of transcriptions of dialogues between children and their caregivers is CHILDES (20), a collection of corpora containing recorded interactions of adults with children of different age and language groups and from different social classes. Transcriptions are morphologically

annotated and mostly follow a (semi-)standard format, and occasionally, some semantic information about the concurrent events is added to the conversation (e.g., what objects are in the scene or what the mother points to). The English portion of CHILDES has been annotated with dependency-based syntactic relations (59). Many of the databases in CHILDES also contain audio or video recordings of the interaction sessions, but these recordings are mostly unannotated.

Some of the audio and video recordings in CHILDES have been annotated by individual research groups for specific purposes. For example, (68) and (23) use video clips of mother-infant interactions from CHILDES, and manually label the visible objects when each utterance is uttered, as well as the objects of joint attention in each scene. Other social cues such as gaze and gesture are also marked. A more systematic approach is taken by the TalkBank project, which is accumulating the speech corpus of children with multimodal annotation (43). Other researchers have collected smaller collections of annotated videos from children. One such example is the recording of adults reading story books to 18 month old infants, annotated to identify the physical objects and the spoken words in each frame in the video (69). Another example is a set of videos of a human operator enacting visual scenes with toy blocks, while verbally describing them (18). These resources are sparse, and the annotation scheme or the focus of annotation is rather arbitrarily chosen by the researchers who developed them. Another massive collection of data has been recently gathered by (56). Roy has recorded his son's development at home by gathering approximately 10 hours of high fidelity audio and video on a daily basis from birth to age three. However, the resulting corpus is not structured. These collections are hard to use without some sort of preprocessing or manual annotation. Nevertheless, they are complementary to the textual data from corpora which lack any semantic information.

Ever since the availability of resources like CHILDES (20), both child-directed (utterances by parents and other adults aimed at children) and child-produced data have been extensively examined. Analyses of child-directed data in particular have mainly focused on the grammaticality of the data, its statistical properties (e.g., 31), and the availability of various cues and constructions (e.g., 29)). Such analyses have provided valuable information about what children have access to. For example, child-directed data has been shown to be highly grammatical (e.g., 5), and sufficiently rich with statistical information necessary for various tasks (e.g., the induction of lexical categories (49)).

Utterances produced by children, on the other hand, have been analyzed with a different goal in mind: to identify the developmental stages that children go through in the course of learning a language, and to detect common behavioural patterns among children from different backgrounds. The parameters usually examined in child-produced data are: (a) the size of the vocabulary that they use; (b) the length of the sentences that they produce; (c) the complexity of these sentences (which syntactic constructions they use); (d) the wide-spread errors that they make and the type of these errors; and (e) how each of these factors changes as the child ages. Also, differences between each of these factors have been studied in children of different genders, nationalities and social classes. Such studies have yielded substantial

evidence about children's learning curves in different tasks (e.g., word learning or argument structure acquisition).

Besides the more direct use of adult-child interaction data, properties of these data are also used in evaluating computational models of language. Statistical properties of child-directed data (average sentence length, distributional properties of words, etc.) are normally used as standard when selecting or artificially creating input for many computational models (e.g, 66; 7; 31; 51). Additionally, several models have attempted to simulate or explain the patterns observed in child-produced data.

In addition to these child-focused collections, there are several large corpora of adult-generated text and speech. These corpora, such as the Brown corpus (22), the Switchboard corpus (27), and the British National Corpus (BNC; 34; 6) contain large amounts of data, and are representative of language used by a large number of speakers of a language (mostly English) in different domains and genre. Some of these corpora are entirely or partially annotated with part of speech tags or parsed (e.g., 46). Although these corpora are normally used as input data for models of grammar induction, they have also been used as basis for comparative and even complementary analyses to those reported using child-related data (e.g., 31).

3 Impact of computational modeling on the study of language

Advances in machine learning and knowledge representation techniques have led to the development of powerful computational systems for the acquisition and processing of language. Concurrently, various experimental methodologies have been used to examine children's knowledge of different aspects of language. Empirical studies of child language have revealed important cues about what children know about language, and how they use this knowledge for understanding and generating natural language sentences. In addition, large collections of child-directed and child-produced data have been gathered by researchers. These findings and resources have facilitated the development of computational models of language. Less frequently, experiments have been designed to assess the predictions of some computational models on a particular learning process. Computational cognitive modeling is a new and rapidly developing field. During its short life span, it has been extensively beneficial to cognitive science in general, and the study of natural language acquisition and use in particular.

One of the main impacts of computational models of language acquisition has been to emphasize the importance of probabilistic knowledge and information theoretic methods in learning and processing language. The role of statistical methods in language acquisition for long in the sidelines has been gaining prominence in recent years. The undeniable success of statistical techniques in processing linguistic data for more applied NLP tasks has provided strong evidence for their impact in human language acquisition (8). On the other hand, shallow probabilistic techniques which are not linguistically motivated can only go so far. For example, pure distributional models have been generally unsuccessful in accounting for learning a

natural language in realistic scenarios. Fifty years after the development of the first computational models of language, hybrid modeling approaches that integrate deep structures with probabilistic inference seem to be the most promising direction for future research.

Developing computational algorithms that capture the complex structure of natural languages in a linguistically and psychologically motivated way is still an open problem. Computational studies of language combine research in linguistics, psychology and computer science. Because it is a young field of a highly interdisciplinary nature, the research methods employed by scholars are inevitably varied and non-standard. This is an unfortunate situation: it is often difficult to compare different models and analyze and compare their findings due to incompatible resources and evaluation techniques they employ. It is vital for the community to share resources and data collections, to develop unified schemes for annotation and information exchange, and to converge on standards for model comparison and evaluation.

When it comes to comparing and evaluating computational models, there is even less agreement among researchers in this field. The majority of algorithms used for simulating language acquisition are unsupervised, mainly because it is highly unrealistic to assume that children receive input data which is marked with the kind of linguistic knowledge they are supposed to learn. As a consequence, there is no gold standard for evaluating the outcome of these unsupervised models and the success of their results and contributions may be difficult to assess. Furthermore, the underlying representation of the linguistic knowledge in human brain is unknown; therefore, the knowledge of language that a model acquires cannot be evaluated on its own. Many models apply their acquired knowledge on different tasks, but such tasks are often chosen arbitrarily. With computational modeling becoming more widespread, it is extremely important for the community to converge on standard evaluation tasks and techniques in each domain that can be used for rigorous evaluation of the methodology and replicability of the results, as in the more traditional disciplines that influence the field.

4 This collection

4.1 Methods and tools for investigating phonetics and phonology

Child language corpora are essential for research on language acquisition, yet prohibitively expensive to build. The study of child language acquisition has made great progress in recent years thanks to the availability of shared corpora and tools among researchers. The CHILDES database includes a large number of corpora for growing number of languages like Danish, Portuguese, German, Russian and Cantonese among others. Moreover, the tools associated with CHILDES (e.g. CLAN) enable easy access to the information provided in corpora allowing complex searches that

combine different levels of information. However, the majority of the tools associated with CHILDES focus on morphology, syntax and semantics (50; 58), with a lack of tools for phonetics and phonology. With the development of Phon this is no longer the case. The chapter *Phon 1.4: A Computational Basis for Phonological Database Elaboration and Model Testing* of this collection by Yvan Rose, Gregory J. Hedlund, Rod Byrne, Todd Wareham, and Brian MacWhinney introduces version 1.4 of Phon – an open-source software program designed for the transcription, coding, and analysis of phonological corpora. Phon 1.4. is a versatile program capable of supporting multimedia data linkage, utterance segmentation, multiple-blind transcription, transcription validation, syllabification, and the alignment of target and actual forms. The system is available with a graphical interface and is used by PhonBank, a database project that seeks to broaden the scope of CHILDES into phonological development and disorders.

A framework for phonetic investigations is also the topic of the chapter *Self-Organization of the Consonant Inventories in the Framework of Complex Networks*, by Animesh Mukherjee, Monojit Choudhury, Niloy Ganguly and Anupam Basu. It introduces a computational framework for representing, analysing and synthesising consonant inventories of the world's languages. The framework is capable of integrating the full variety in consonants as well as languages. It is essentially a complex network with two sets (or partitions) of nodes: one for consonants and the other one for languages. The primary objective is to provide the means to systematically analyse and synthesise the structure of the Phoneme-Language Network (PlaNet) and thereby, explain the distribution with which consonants occur across languages.

4.2 Classifying words and mapping them to meanings

The task for learning the meanings of words can be viewed as children learning the association between a word form and a concept after hearing repeated instances of the word form used in reference to the concept. Despite its misleading apparent simplicity, there are many challenges to this task. First, few words are used in isolation. Children usually hear words in a sentential context. Secondly, a sentence can potentially refer to many different aspects of a scene, as a typical learning situation usually involves a large number of objects. For a learner who does not know the meanings of words yet, it can be a real challenge to figure out the exact aspect (or the relational meaning) that the sentence conveys. Thirdly, child-directed speech has been shown to contain a substantial level of noise and ambiguity. Therefore learning the correct mapping between each word and its meaning is a complex process that needs to be accounted for by a detailed model.

In addition to the acquisition of word meanings, psycholinguistic studies suggest that early on children acquire robust knowledge of some of the abstract lexical categories such as nouns and determiners. For example, (25) show that two-year-olds treat novel words which do not follow a determiner (e.g., *Look! This is Zag!*) as a proper name which refers to an individual. In contrast, they tend to interpret novel

words which do follow a determiner (e.g., *Look! This is a zag!*) as a mass noun. However, learning lexical categories takes place gradually, and not all categories are learned at the same time. For example, (65) show that two-year-olds are more productive with nouns than with verbs, in that they use novel nouns more frequently and in more diverse contexts. How children gain knowledge of syntactic categories is an open question. Children's grouping of words into categories might be based on various cues, including the phonological and morphological properties of a word, distributional information about its surrounding context, and its semantic features.

The chapter of Fatemeh Torabi Asr, Afsaneh Fazly, and Zohreh Azimifar's *From cues to categories:A computational study of children's early word categorization* focuses on the acquisition of word categories. The authors investigate the types of information children require in order to learn these categories. The paper proposes a computational model which is capable of acquiring categories from distributional, morphological, phonological, and semantic properties of words. The results show that syntax plays an important role in learning word meaning (and vice versa). Additionally, the proposed model can predict the semantic class of a word (e.g., action or object) by drawing on the learned knowledge of the word's syntactic category.

The chapter *Lexical Category based Computational Model of Construction Acquisition*, by Bichuan Zhang and Xiaojie Wang presents a computational model of language acquisition which derives lexical information from a corpus of child speech. The model associates each word with a part-of-speech tag and a specific construction with an accuracy above 70%. This is the first time lexical acquisition techniques of this type have been applied to a Chinese child language corpus. In addition, the authors present a small comparison of child-produced speech, child-directed speech and adult language. They report interesting observations which are in line with observations reported in linguistic literature on language acquisition.

In the chapter *In learning nouns and adjectives remembering matters: a cortical model*, Alessio Plebe, Vivian De la Cruz and Marco Mazzone investigate different artificial models which can be used to mimic the acquisition of mapping of words to meanings. The authors use a hierarchy of artificial cortical maps to develop models of artificial learners that are subsequently trained to recognize objects, their referents, and the adjectives pertaining to their color. In doing so, they address several fundamental issues such as noun acquisition as well as adjective acquisition (which is known to be a much more difficult task). The relation between nouns and adjectives also plays a role in their meaning, so the model accounts for an embryonic syntax. Results reported in the chapter explain various developmental patterns followed by children in acquiring nouns and adjectives, by perceptually driven associational learning processes at the synaptic level.

4.3 Learning morphology and syntax

Learning the categories or meanings of words is not enough for successful communication: a language learner has to also master the regularities that govern word

forms, and the acceptable combinations of words into sentences. Natural languages are highly regular in their morphological and syntactic structure. Regularities in syntax, such as the position of the subject and object in relation to the verb, can provide important information for the learner about the structure of the language (e.g. the SVO order in English, and SOV in Japanese). Nevertheless, in each language there are words which do not conform to such general patterns, and one well known case is that of exceptions to the English past tense verb formation (57; 45; 67) with *ed* suffix (e.g. *receive/received* vs *ring/rang/*ringed*). The challenge in learning morphology and syntax is how to grasp the abstract regularities that govern form, as well as the idiosyncratic properties of individual words and constructions based on potentially poor stimulus and/or no consistent negative evidence.

One well known example in the discussion on the poverty of the stimulus centers around the knowledge of structure dependency in question inversion (10), and whether the relevant data provides sufficient information to guarantee successful acquisition. For instance, for native speakers in general the following two sentences are closely related:

1. *The company has bought his shares.*
2. *Has the company bought his shares?*

but for learners they provide a though challenge, as a learner has to identify the relation between the declarative sentence and the corresponding derived interrogative form without overgenerating to possible but ungrammatical, unattested or unnatural forms (e.g. **Bought the company has his shares?* and *?Has bought the company his shares?*).

Related to this discussion is the chapter by Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick in the chapter *Treebank Parsing and Knowledge of Language* who examine some complex linguistic constructions with nonlocal dependencies that are also challenging for computational models, such as tense marking, wh-questions and passives in English, assessing gaps in the *knowledge of language* acquired from large corpora. They investigate if grammars acquired by statistical parsers can successfully account for a full knowledge of language, verifying in which cases poor performance may be due to data sparsity, and in which it might arise from the underlying grammatical frameworks. They propose a general approach which incorporates linguistic knowledge by means of *regularizations* that canonicalize predicate-argument structure, which results in statistically significant improvements in parser performance. The results obtained indicate the contributions of distributional and linguistic properties of data needed for a successful account of language, and where insights from language acquisition can positively inform statistical parsing development.

In other cases, the data may provide enough information for a learning mechanism to obtain results compatible with syntactic evolution in language acquisition. Christophe Parisses's chapter, *Rethinking the syntactic burst in young children*, focuses on the speed and correctness of child language acquisition at the point of "syntactic burst" which usually occurs around age two to three. The author shows that recent theories based on general cognitive capabilities such as perception, memory or

analogy processing do not sufficiently explain the syntactic burst. He then proposes a testing procedure to demonstrate that the acquisition of usage-based and fixed-form patterns can account for the syntactic evolution in language acquisition. The patterns are applied to the Manchester corpus taken from the CHILDES database. The author shows that simple mechanisms explain language development until age three and that complex linguistic mastery does not need to be available early in the course of language development.

4.4 Linking syntax to semantics

Experimental child studies have shown that children are sensitive to associations between syntactic forms and semantic interpretations from an early age, and that they use these associations to produce novel utterances (3; 52; 20). Children's learning of form-meaning associations is not well understood. Specifically, it is not clear how children learn the item-specific and general associations between meaning and syntactic constructions.

One aspect of language that provides a rich testbed for studying form-meaning associations is the argument structure of verbs. The argument structure of a verb determines the semantic relation a verb has with its arguments and how the arguments are mapped onto valid syntactic expressions. This complex structure exhibits both general patterns across semantically similar verbs, as well as more idiosyncratic mappings of verbal arguments to syntactic forms. This is particularly acute in the case of multiword expressions, which vary along a continuum from literal to more idiomatic cases, like the phrasal verbs *carry up* and *come up [with an idea]* (as *propose an idea*), respectively, and from more to less productive expressions (e.g. *carry up/down* vs *come up/?down*).

In addition to regularities at the level of argument structure, research on child language has revealed strong associations between general semantic roles such as Agent and Destination and syntactic positions such as Subject and Prepositional Object (e.g., 21, and related work). Despite the extensive use of semantic roles in various linguistic theories, there is little consensus on the nature of these roles. Moreover, researchers do not agree on how children learn general roles and their association with grammatical functions.

The first chapter on this topic, *Learning to interpret novel noun-noun compounds: Evidence from category learning experiments* by Barry J. Devereux and Fintan J. Costello, focuses on the analysis of one type of multiword expressions: nominal compounds. The interpretation of noun-noun compounds is known to be challenging and difficult to predict since these constructions are highly ambiguous. Two approaches have been proposed for the interpretation of noun-noun compounds: one which assumes that people make use of distributional information about the linguistic behaviour of words and how they tend to combine as noun-noun phrases; another which assumes that people activate and integrate information about the two constituent concepts' features to produce interpretations. Devereux

and Costello propose a model that combines these two approaches. They present an exemplar-based model of the semantics of relations which captures these aspects of relation meaning, and show how it can predict experimental participants' interpretations of novel noun-noun compounds.

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson, in the chapter *Child Acquisition of Multiword Verbs: A Computational Investigation*, address the question of the acquisition of multiword expression by children. They show that multiword expressions have received far less attention than simple words in child language studies. However, in natural language processing, there is a long research tradition on models for the recognition and analysis of idiosyncratic expressions. The authors explore whether this computational work on multiword lexemes could be extended in a natural way to the domain of child language acquisition where an informative cognitive model must take into account two issues: what kind of data the child is exposed to, and what kind of processing of that data is cognitively plausible for a child.

In *Starting from Scratch in Semantic Role Labeling: Early Indirect Supervision* Michael Connor, Cynthia Fisher and Dan Roth investigate the problem of assigning semantic roles to sentence constituents, where a learner needs to parse a sentence, find possible arguments for predicates, and assign them adequate semantic roles. They look at possible starting points for a learner using a computational model, Latent BabySRL, which learns semantic role classification from child-directed speech. They found that even before acquiring any specific verb knowledge this model is able to begin comprehending simple semantics in a plausible setup when initialized with a small amount of knowledge about nouns and some biases.

In *Gradual Acquisition of Verb Selectional Preferences in a Bayesian Model*, Afra Alishahi and Suzanne Stevenson present a cognitive model for inducing verb selectional preferences from individual verb usages. The selectional preferences for each verbal argument are represented as a probability distribution over the set of semantic properties that the argument can possess, *i.e.* a semantic profile. The semantic profiles yield verb-specific conceptualizations of the arguments associated with a syntactic position. The proposed model can learn appropriate verb profiles from a small noisy training data, and can use them to simulate human plausibility judgments and to analyse implicit object alternation.

5 Concluding remarks

These chapters present a cross-section of the research on computational language acquisition, and investigate linguistic and distributional characteristics of the learning environment for different linguistic aspects, adopting a variety of learning frameworks. Computational investigations like these can contribute to research on human language acquisition, challenging the extent to which innate assumptions need to be specified in these models, and how successful they are in each of the specific tasks, providing valuable insights into learnability aspects of the data, the learning

environment and the specific frameworks adopted. This is a new and fast growing multidisciplinary field that has yet much to achieve, evolving along with its foundational areas.

References

[1] Akhtar, N. (1999). Acquiring basic word order: evidence for data-driven learning of syntactic structure. *Journal of Child Language*, 26:339–356.

[2] Alishahi, A. (2010). *Computational Modeling of Human Language Acquisition*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[3] Bowerman, M. (1982). Evaluating competing linguistic models with language acquisition data: implications of developmental errors with causative verbs. *Quaderni di semantica*, 3:5–66.

[4] Brent, M. R. and Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61(1-2):93 – 125.

[5] Broen, P. A. (1972). *The verbal environment of the language-learning child*. American Speech and Hearing Association.

[6] Burnard, L. (2000). Users reference guide for the British National Corpus. Technical report, Oxford University Computing Services.

[7] Buttery, P. and Korhonen, A. (2007). I will shoot your shopping down and you can shoot all my tins: automatic lexical acquisition from the childes database. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 33–40. Association for Computational Linguistics.

[8] Chater, N. and Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Science*, 10(7):335–344.

[9] Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

[10] Chomsky, N. (1975). *The logical structure of linguistic theory*. Plenum press.

[11] Chomsky, N. (1980). *Rules and representations*. Oxford: Basil Blackwell.

[12] Chomsky, N. (1981). *Lectures on government and binding*. Mouton de Gruyter.

[13] Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger Publishers.

[14] Clark, A. and Lappin, S. (2010). *Linguistic nativism and the poverty of stimulus*. Wiley Blackwell, Oxford and Malden, MA.

[15] Clark, E. V. (2009). *First language acquisition*. Cambridge University Press, second edition.

[16] Cullicover, P. W. (1999). *Syntactic nuts*. Oxford University Press.

[17] De Marcken, C. G. (1996). *Unsupervised language acquisition*. PhD thesis.

[18] Dominey, P. and Boucher, J. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence*, 167(1-2):31–61.

[19] Dowman, M. (2000). Addressing the learnability of verb subcategorizations with bayesian inference. In Gleitman, L. R. and Joshi, A. K., editors, *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*.

[20] Elman, J. (2001). *Essential readings in language acquisition*, chapter Connectionism and language acquisition. Oxford: Blackwell.

[21] Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, 31(1):41–81.

[22] Francis, W., Kučera, H., and Mackie, A. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin Harcourt (HMH).

[23] Frank, M., Goodman, N., and Tenenbaum, J. (2007). A Bayesian framework for cross-situational word learning. *Advances in Neural Information Processing Systems*, 20.

[24] Frazier, L. and Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 13:187–222.

[25] Gelman, S. and Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, pages 1535–1540.

[26] Gibson, E. and Wexler, K. (1994). Triggers. *Linguistic Inquiry*, 25:407–454.

[27] Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92.*, volume 1.

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

[29] Goldberg, A. E. (1999). *The Emergence of Language*, chapter Emergence of the semantics of argument structure constructions, pages 197–212. Carnegie Mellon Symposia on Cognition Series.

[30] Grünwald, P. (1996). A minimum description length approach to grammar inference. In Wermter, S., Riloff, E., and Scheler, G., editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, volume 1040 of *Lecture Notes in Computer Science*, pages 203–216. Springer.

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

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

[33] Keller, B. and Lutz, R. (1997). Evolving stochastic context-free grammars from examples using a minimum description length principle. In *Workshop on Automata Induction Grammatical Inference and Language Acquisition, ICML-97*.

[34] Leech, G. (1992). 100 million words of English: the British National Corpus (BNC). *Language Research*, 28(1):1–13.

[35] Legate, J. and Yang, C. (2002). Empirical re-assessment of stimulus poverty arguments. *Linguistic Review*, 19(1/2):151–162.

[36] Leonard, L. (2000). *Children with specific language impairment*. the MIT Press.

[37] Li, M. and Vitányi, P. M. B. (1995). *Computer Science Today*, volume 1000 of *Lecture Notes in Computer Science*, chapter Computational Machine Learning in Theory and Praxis. Springer Verlag, Heidelberg.

[38] MacWhinney, B. (1982). Basic syntactic processes. In Kuczaj, S., editor, *Language Development: Syntax and Semantics*, volume 1, pages 73–136. Hillsdale, N.J., Lawrence Erlbaum.

[39] MacWhinney, B. (1987). The competition model. In MacWhinney, B., editor, *Mechanisms of language acquisition*. Hillsdale, NJ: Erllbaum.

[40] MacWhinney, B. (1993). Connections and symbols: closing the gap. *Cognition*, pages 291–296.

[20] MacWhinney, B. (1995). *The CHILDES project: tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates, second edition.

[42] MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language*, 31:883–914.

[43] MacWhinney, B., Bird, S., Cieri, C., and Martell, C. (2004). TalkBank: Building an open unified multimodal database of communicative interaction. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon*, pages 525–528.

[44] Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46:53–85.

[45] Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., and Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4, Serial No. 228).

[46] Marcus, M., Santorini, B., and Marcinkiewicz, M. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[47] Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

[48] McClelland, J. L., Rumelhart, D. E., and Group, T. P. R. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, volume 2. Cambridge, MA: Bradford Books/MIT Press.

[49] Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

[50] Parisse, C. and Le Normand, M. T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers*, 32:468–481.

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

[52] Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

[26] Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua*, 92:377–410.

[54] Pullum, G. and Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review*, 19(1/2):9–50.

[55] Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

[56] Roy, D. (2009). New horizons in the study of child language acquisition. In *Proceedings of Interspeech 2009, Brighton, England*.

[57] Rumelhart, D. and McClelland, J. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing. *Mechanisms of language acquisition*, pages 195–248.

[58] Sagae, K., Davis, E., Lavie, A., MacWhinney, B., and Wintner, S. (2007). High-accuracy annotation and parsing of childes transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

[59] Sagae, K., Davis, E., Lavie, A., MacWhinney, B., and Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(03):705–729.

[60] Steedman, M., Baldridge, J., Bozsahin, C., Clark, S., Curran, J., and Hockenmaier, J. (2005). Grammar acquisition by child and machine: the combinatory manifesto. Invited Talk at the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). http://www.clips.ua.ac.be/conll2005/pdf/steedman.pdf.

[61] Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632.

[62] Tenenbaum, J., Griffiths, T., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

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

[64] Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Harvard University Press.

[65] Tomasello, M., Akhtar, N., Dodson, K., and Rekau, L. (1997). Differential productivity in young children's use of nouns and verbs. *Journal of Child Language*, 24(02):373–387.

[66] Villavicencio, A. (2002). *The Acquisition of a Unification-Based Generalised Categorial Grammar*. PhD thesis, Computer Laboratory, University of Cambridge.

[67] Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford, New York: Oxford University Press.

[68] Yu, C. and Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165.

[69] Yu, C. and Smith, L. (2006). Statistical cross-situational learning to build word-to-world mappings. In *Proceedings of the 28th annual meeting of the cognitive science society*. Citeseer.