Computational models of Language Development
Daniel Freudenthal and Afra Alishahi

Human language development 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 cognitive plausibility criteria (such as memory and processing limitations that humans face), and to explain the developmental stages observed in children, especially in the light of nativist views that claim language is too complex to be learned without innate linguistic knowledge. 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, 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 mechanisms involved in human language development.

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. First, 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. Second, 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 makes it possible to precisely specify those aspects of the model that are deemed innate (the learning mechanism and representations) and those aspects that are learned. Third, 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. Moreover, the performance of two different mechanisms on the same data set can be compared against each other, something that is almost impossible in an experimental study

on children. Finally, 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 that have not been previously studied.

Despite these advantages, computational modeling should not be viewed as a substitute 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 insights about which representations and processes are most plausible in light of the experimental findings on child language development. 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 models 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.

Computational techniques have been applied to different domains of language acquisition, including word segmentation and phonology, morphology, syntax, semantics and discourse. Among these, the acquisition of word meaning, syntax, and the association between syntax and semantics have been more carefully studied using cognitively plausible computational models that simulate the behavioural patterns observed in humans.

Acquisition of Lexicon

Several computational models of word learning have been proposed in the literature. On a high level, we can distinguish two main groups of models: those which study the association of words and meanings in isolation, and those which study word learning in a sentential context. In the first group, a number of connectionist models learn to link labels (or word forms) to referents (or meanings) from input that consists of pairings of a distributed phonological representation of the word form with a distributed representation of the referent of the word. These models show gradual sensitivity to the phonological properties of the

word form and the relevant meaning distinctions (e.g., shape), and can simulate the facilitation of learning second labels for familiar objects. The second group of models use cross-situational evidence to constrain hypotheses about the meaning of each word. These models use full utterances paired with (often noisy) representations of the visual context as input, and simulate various behavioural patterns such as a sudden increase in the rate of vocabulary growth and the acceleration of word learning with exposure to more input. Extensions of some of these models integrate syntactic or social cues into cross-situational word learning.

Computational modeling of the process of word learning in children has been one of the more successful cases of using computational techniques to study an aspect of human language acquisition. Several experimental studies hint at a change of behaviour in most children during the learning process (e.g., vocabulary spurt), and many conflicting hypotheses have been proposed to account for this pattern. However, many computational models have shown that most of these patterns can be a by-product of the statistical properties of the input that children receive. Most importantly, computational studies of word learning suggest that children's behaviour in this task is not necessarily due to a change in the underlying learning mechanism, or to the application of highly task-specific constraints or biases.

Acquisition of Linguistic Structure

Learning the syntactic structure of language has long been considered as the core challenge of learning a language. It has been argued that general learning and problem solving mechanisms are not sufficient to explain humans language acquisition, and some innate knowledge is also needed to account for their exceptional linguistic skills. Most notably, the Universal Grammar theory proposes that each infant is born with a detailed and innately specified representation of a grammar which determines the universal structure of natural languages. This universal grammar describes those aspects of grammar that are invariant across languages, with the differences being captured by a set of parameters which have to be adjusted over time to the language the child is exposed to.

Several models have been developed to simulate this process. Typically, these models are symbolic and assume a relatively abstract representation of the input. Early models attempted to set the parameters by parsing individual utterances as they come in. A problem with this

approach is that many utterances are ambiguous with respect to certain parameters. What's more, language is noisy and contains ungrammatical input, which is problematic for these models. More recent models take a probabilistic approach: parameters are assigned weights that reflect the amount of evidence in the input that support their setting. Such probabilistic models are more successful in the face of noise and ambiguity, but still struggle with the fact that parameters can interact and the number of parameters that needs to be set is relatively large (30-40), leading to a search space well in excess of a billion permutations.

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. Analyses of large collections of data on child-parent interactions have raised questions about the inadequacy of linguistic data. It has been shown that child-directed data provides rich statistical cues about the abstract structures and regularities of language.

Recent psycholinguistic findings hint at a `bottom-up' process of language acquisition, usually referred to as the usage-based or empirical view. Advocates of this view 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, starting with individual items and gradually inducing more general and abstract structures. However, the challenge is to show how children move from an initial phase during which their speech may be 'frozen' and lexically specific to a state where children show adult-like syntactic competence.

One approach aims to show that the grammar that best describes corpora of child speech starts out relatively simple and increases in complexity as children grow older. Within this approach, simple utterances are viewed as templates containing slots that can be progressively filled by more complex materials. Corpora of child speech collected at different ages are then analysed with respect to number and type of fillers or substitutions. Models within this approach differ with respect to the level of abstraction that is needed to describe the filler elements (ranging from lexically specific strings to abstract tree structures) but typically find that the number and complexity of substitutions increases with age, even when controlling for utterance length. While these models technically describe the child's grammar, they have considerable computational overhead, and do not specify how the child's evolving grammar derives from the input to which they are exposed.

A second class of models attempts to show how certain aspects of child language can be understood in terms of the input to which children are exposed. Such models have been employed to meet the Learnability problem head-on by showing that certain aspects of grammar that are thought to be unlearnable can, in fact, be learned from the input. A phenomenon that has received a considerable amount of attention is that of Auxiliary Fronting in question formation.

1.The boy is hungry

2.Is the boy hungry?

3.The boy who is smiling is happy

4.Is the boy who is smiling happy?

5.*Is the boy who smiling is happy?

According to linguistic theory, question formation works through a process of "movement" which results in the auxiliary is from sentence 1 to appear at the front of sentence 2. Sentence 3 however, poses a potential problem as there is a choice of two auxiliaries that might be moved. The 'correct' (structure dependent) rule is to move the auxiliary in the main clause of the declarative to the beginning of the sentence (leading to 4). A child could equally well, however, entertain the (structure independent) notion that the left-most auxiliary needs moving, leading to the incorrect sentence 5. In practice however, children rarely make errors such as 5, despite the fact that correct examples such as 3 rarely occur in the input. One line of modeling has focused on showing that direct evidence such as (3) is not required in order to select the correct alternative from 4 and 5. Instead, it is argued, there is sufficient indirect evidence in the form of word transition statistics. Thus, utterance 4 is more likely to be an English utterance than utterance 5 because the word transitions in 4 are more representative of the word transitions that are found in corpora of English child-directed speech. While models of this sort can provide evidence that certain dependencies are learnable from the input, they tend to focus on one detailed dependency, rather than show how child speech as a whole is shaped by the input.

A third approach aims to directly simulate (corpora of) child speech by learning from the input children hear. This approach is less concerned with learnability and abstraction, but instead focuses on the fact that much of children's early multi-word speech is incomplete and contains many errors. Work within this approach has shown that early child speech can be understood as incomplete utterances or chunks that have been

learned directly from the input through cognitively plausible biases such as primacy and recency effects. Common errors such as errors of inflection (e.g. he go), are not the result of missing abstract linguistic features like Tense or Agreement, but can be produced by omitting the modal can from utterances like he can go or can he go. These error rates decrease, not as a result of a maturing grammar, but because increased lexical learning results in utterances becoming longer and more complete. Models within this approach derive much of their validity from the fact that common processing constraints can simulate data from different languages. While these models can successfully show what areas of language acquisition can be understood without assuming abstract knowledge, their lack of abstraction means they are likely to struggle to explain later stages of development.

Models of syntax acquisition have thus shown that the grammars that describe children's speech can be characterized as developing from very concrete to increasingly abstract as children grow older. They have furthermore confirmed that there is more information in the input than has traditionally been assumed, suggesting that children's ability to avoid certain errors reflects input characteristics rather than innate linguistic knowledge. Additionally, it has been shown that certain frequent errors in child speech can be understood, not in terms of missing abstract linguistic features, but instead in terms of input-driven learning resulting in omission from target utterances. However, while individual models have been successfully applied to specific areas of the learnability problem, many challenges remain for a complete model of language acquisition.

Form-Meaning Associations

The representation and acquisition of syntax have historically been studied independently of the meaning of the words in an utterance. However, words are put together to convey a meaning, and the acceptability of an utterance is determined by the semantic properties of its relational words and the arguments they take. This is particularly noticeable in the domain of verb use, on which much research has focused. The main challenge for models of verb learning is to explain how the semantic and syntactic relations between a verb and its arguments are acquired, and how such verb-specific associations are related to more general and abstract regularities or constructions.

As in other aspects of language acquisition, the association between

syntactic form and semantic content has been attributed to innate representations of linguistic knowledge. According to the semantic bootstrapping hypothesis, innate linking rules map semantic roles (e.g. agent, patient) onto potential syntactic functions (subject, object). Alternatively, a usage-based approach suggests that children move from relatively verb-specific roles such as "hitter" and "hittee" to more general schemas that pair a specific syntactic form (e.g. the directed motion transitive) with a meaning (X causes Y to move).

Computational modeling has been used to investigate each of these proposals. Nativist models that attempt to learn the associations between the syntactic and semantic properties of words tend to rely on extensive prior knowledge, either in the form of innate linguistic categories and combination rules or in the form of a structured hypothesis space and its prior probabilities. In contrast, a number of connectionist models simulate the assignment of thematic role to the arguments of common verbs on the basis of a number of cues such as the identity and semantics of verbs and their arguments.

More recent work has focused on the acquisition of abstract form-meaning associations (or constructions) on the basis of individual verb usages. These models are trained on child-directed utterances, each paired with the semantic properties of the corresponding event and its participants (including their thematic roles). Computational simulations of these models demonstrate an initial stage characterised by conservative employment of the more frequent usages for each individual verb, followed by a phase when more general patterns are grasped and applied overtly. This phase leads to occasional overgeneralization errors, and an eventual recovery from making such errors by receiving more input.

Models of learning form-meaning associations in language suggest that meaningful associations between syntactic forms and semantic features can be learned using appropriate statistical learning techniques. More importantly, probabilistic frameworks for representing and using these associations reflect a natural interaction between item-specific mappings between words and their arguments, and more general associations at the level of constructions. However, a detailed model of the acquisition of form-meaning associations that reflects the semantic complexities of naturalistic input data is still lacking.

Main Challenges

Developing computational algorithms that capture the complex structure

of natural languages is still an open problem. It is often difficult to compare different models and analyze and compare their findings due to incompatible resources and evaluation techniques they employ. Moreover, there are few resources available that provide realistic semantic information which resembles the input data that children receive.

The complex structure of natural languages, has resulted in computational models focusing on restricted and isolated aspects of learning a language. Simplification is usually unavoidable when studying a complex problem, but the interaction between various aspects of linguistic knowledge and the timeline of their acquisition is one of the main open questions that needs to be investigated.

Daniel Freudenthal, University of Liverpool, United Kingdom.

Afra Alishahi, Tilburg University, The Netherlands.

See also: Poverty of the Stimulus argument, Principles and Parameters Framework, Item based learning.

Further Reading:

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

Chater, N. & Manning, C. (2006). Probabilistic models of language processing and acquisition. Trends in Cognitive Sciences, 10, 287-291.

Clark, A. & Lapin, S. (2011). Linguistic Nativism and the poverty of the stimulus. Wiley-Blackwell.

Goldberg, A. (2006). Constructions at work. Oxford University Press Oxford, United Kingdom.

Tomassello, M. (2003). Constructing a language: a usage-based theory of language acquisition. Harvard University Press.

Yang, C. (2002). Knowledge and Learning in natural language. Oxford University Press, USA.