Artificial Intelligence 242 (2017) 23–51

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

A model of language learning with semantics and meaning-preserving corrections

Dana Angluin [a], Leonor Becerra-Bonache [b,*]

[a] Department of Computer Science, Yale University, New Haven, CT, USA
[b] Univ Lyon, UJM-Saint-Etienne, CNRS, Laboratoire Hubert Curien UMR 5516, F-42000, Saint-Etienne, France

ARTICLE INFO

Article history:
Received 25 August 2015
Received in revised form 30 September 2016
Accepted 4 October 2016
Available online 11 October 2016

Keywords:
Semantics
Corrections
Language learning
Grammar learning

ABSTRACT

We present a computational model that takes into account semantics for language learning and allows us to model meaning-preserving corrections. The model is constructed with a learner and a teacher who interact in a sequence of shared situations by producing utterances intended to denote a unique object in each situation.

We test our model with limited sublanguages of 10 natural languages exhibiting a variety of linguistic phenomena. The results show that learning to a high level of performance occurs after a reasonable number of interactions. Comparing the effect of a teacher who does no correction to that of a teacher who corrects whenever possible, we show that under certain conditions corrections can accelerate the rate of learning.

We also define and analyze a simplified model of a probabilistic process of collecting corrections to help understand the possibilities and limitations of corrections in our setting.

1. Introduction

Children acquire their native language easily, quickly and without any specific training. However, *interaction* with other human speakers is crucial for their language development. This is evidenced by the fact that children cannot learn language from just listening to the radio. Studies of hearing children of deaf parents who had access to English only from the radio or television show that children did not learn to speak English from this kind of input [21].

In a child's early language learning, the *correspondence* between utterances and the situations in which they are made also seems to be a very important source of information for both the child trying to figure out how to use the language and the adult trying to make sense of the child's often imperfect utterances. Moreover, adults often reformulate these imperfect utterances to make sure that they have understood them correctly [14]. These *corrections* (in the form of reformulations) *preserve the meaning* that the child intends to express.

We propose a simple computational model of language learning that takes into account all these aspects. Our model has a teacher and a learner who interact in a sequence of shared situations by producing utterances intended to denote a unique object in each situation. There is no explicit semantic annotation of the utterances; the learner uses cross-situational correspondences to learn to comprehend the teacher's utterances and produce its own appropriate utterances.

This setting allows us to study several questions: whether the learner can achieve a high level of performance after a reasonable number of interactions, whether the teacher can offer meaningful corrections to the learner, whether the learner can detect intended corrections by the teacher, and whether the presence of corrections by the teacher has an effect on

* Corresponding author.
E-mail addresses: dana.angluin@yale.edu (D. Angluin), leonor.becerra@univ-st-etienne.fr (L. Becerra-Bonache).

http://dx.doi.org/10.1016/j.artint.2016.10.002
0004-3702/

Utterance	*The green circle*
Situation	$\{bi1(t_1), gr1(t_1), ci1(t_1), le2(t_1, t_2), bi1(t_2), bl1(t_2), sq1(t_2)\}$

Fig. 1. Example of an utterance-situation pair. The situation consists of a big green circle to the left of a big blue square.

language acquisition by the learner. For our model, the answer to each of these questions is yes, and while the model is in many respects artificial and simplified, we believe it sheds new light on these issues.

This paper is an extension of our previous work on semantics and corrections [3–7]. In [3,4], we proposed a non-probabilistic model of meaning and denotation using finite-state transducers. We gave an algorithm to learn a meaning function and proved that it finitely converges to a correct result under a specific set of assumptions about the transducer and examples used. This work was based only on the learner's comprehension, and it was just a first step in the formulation of our current model. In [6], we considered a statistical approach to model comprehension and production (like the one used in this paper), which produced a more powerful version of our initial model and allowed us to model meaning-preserving corrections. This work was based on the effects of meaning-preserving corrections on language learning.

The paper is organized as follows. We review related work in section 2. In section 3 we describe the main components of our model, and in section 4 the algorithms used by the teacher and the learner during the learning process. Section 5 presents the results obtained by testing our model with limited sublanguages of 10 natural languages in a common domain of situations. In section 6, we give a simple model of the effect of corrections based on the coupon-collector problem, in which corrections may speed up learning by 50%. The paper closes in section 7 with a discussion and suggestions for future work.

2. Related work

In this section we review work on computational models of language learning incorporating semantics, and work in linguistics related to the issues of positive and negative data and corrections in natural language acquisition. In order to better understand the relevance of the related work, we first give a brief overview of our approach and then discuss the related work.

2.1. Overview of our approach

The main focus of our model is to investigate the role of semantics and corrections in the process of learning to understand and speak a natural language. We adopt similar assumptions to the ones in [22]. We assume that in the early stages of language acquisition, children learn the meaning of words through exposure to many utterances that refer to things that are perceptible in their environment [49]. Moreover, when children hear an utterance in a concrete (and observable) situation/scene, they can establish links between the utterance and the elements that appear in the scene [22]. Hence, following these assumptions, we use pairs consisting of an utterance and a situation in which this utterance has been produced.

Our representation of situations reflect the assumption that children, at an early age, can recognize objects and different concepts even if they are not able to communicate yet. For example, they can recognize objects such a ball, and that "green" and "blue" have something in common (i.e., they are colors), even if they do not have a word for them. In our model, a situation is a logical description of the elements that a learner can perceive in a given scene. An example of an utterance and its corresponding situation is presented in Fig. 1.

The (utterance, situation)-pairs contain similar properties to those of input received by children, for example: i) alignment ambiguity: the mappings between words in an utterance and meaning elements in the corresponding scene representation are not explicitly indicated; ii) referential uncertainty: the scene representation may contain meaning elements that do not correspond to words in the utterance.

In our model, the learner and the teacher interact as follows. First of all, a situation is presented to both of them. Then, the learner and the teacher try to produce an utterance that refers to one of the objects in the situation. At the beginning, since the learner does not have any knowledge about the language, it will not be able to produce any utterance. As soon as enough data has been provided, the learner will start to construct its own grammar and will attempt to produce appropriate utterances (i.e., it learns in an incremental way the language it is exposed to). The teacher can either produce a correction of the learner's utterance or produce a random denoting utterance for the given situation. Hence, both the teacher and the learner engage in *comprehension* and *production* of utterances which are intended to be appropriate to their shared situation.

The goal of the learner is to be able to produce every appropriate utterance in any given situation. That is, given a situation, the learner should eventually produce only correct denoting utterances for that situation, and should be able to produce all of them. In fact, our model is probabilistic, and what we require is that the probability of learner errors be reduced to very low levels.

2.2. Computational models of language learning

The study and development of computational models of language learning is of great interest not only to better understand the process of language acquisition, but also for practical applications of language learning by machines; the advantages of having a machine able to understand and speak a natural language would be innumerable. However, the

complexity of these tasks has led researchers to focus on individual sub-problems. As a result most work has focused only on language comprehension, and on just a single phenomenon (e.g., single word learning [41] or syntactic category acquisition [39]). The problem of focusing on just one phenomenon is that we can have a model that is good at performing a task but cannot perform other tasks (for example, it can be good at lexical acquisition but not at syntactic acquisition).

Some of the more empirical works that are focused on comprehension and take into account semantics are those of Siskind [45,46], Marcken [19], Regier [40] Bailey [8], Roy and Pentland [42] and Yu and Ballard [54]. The systems developed by Siskind [45] and Marcken [19] learn word semantics in simulation; it is assumed that the learner already has concepts for each of the words to be learned, and word learning becomes a problem of matching words to concepts. The representation used by these systems is based on compositional semantics. Moreover, Siskind investigated the use of cross-situational analysis to model lexical acquisition. The systems introduced by Regier [40] and Bailey [8] use more realistic simulated data; they focus on learning particular words in isolation, rather than in a linguistic context. And finally, systems using embodied sensors, such as that of Roy and Pentland [42] and Yu and Ballard [54], have focused on learning definitions for concrete nouns or physical verbs. The learner does not know anything about the language, except its phonemes. These systems can produce associations between phoneme sequences and visual stimuli, but they cannot produce grammatical utterances from what they have learned.

There is a large body of research dealing with computational modeling of cognitive aspects of language learning [50]. One of their main goals is to study the language learning task under cognitively plausible criteria and to explain the developmental stages observed in children. Many models have been focused on word learning and adopt a probabilistic approach [22,23,53]. One of the closest works to our research is that of Fazly et al. [22]. They presented a computational model of early word learning to shed some light on the mechanisms that underly children's language acquisition. The model learns word meanings by using an incremental and probabilistic approach (word meaning is defined as a probabilistic association between a word and a concept). It is robust to noise (i.e., an utterance may contain words whose appropriate meanings are not included in the scene) and uncertainty in the input (i.e., the scene contains elements that are not relevant to the utterance). The results suggest that much about word meanings can be learned from naturally occurring utterances, paired with meaning representations. Moreover, the model successfully accounts for many of the observed patterns in children's language acquisition. Our model is based on similar assumptions, but it differs in three main aspects: i) utterances given to our learner offer a partial or total description of the scene, but they do not refer to things that are not in the scene; ii) we are not limited to word learning (our learning task is more complex: our system learns to generate and understand relevant utterances in a given scene); iii) we use a richer semantic representation.

Several robotics researchers have also worked on word learning. One of the most closely related research efforts is that of Kevin Gold and his collaborators [25–27]. They introduced a word learning system called TWIG (Transportable Word Intension Generator) that allows a robot to learn compositional meanings for new words. An important feature of TWIG is the use of decision trees to build word definitions. Unlike other approaches, TWIG learns tree structures for storing word semantics in an unsupervised manner. Moreover, it uses words learned in complete grammatical sentences for production, comprehension, or referent inference. Our model also accommodates comprehension and production tasks, but we use a different representation of word meaning, that is not based on Frege's principle of compositionality; this (sometimes) allows us to assign a correct meaning to a sentence even if the sentence is not grammatically correct. For example, a sentence that is ungrammatical because it is missing an article may still be meaningful. We also use decision trees for learner production, not to provide the definition of a word but to help the learner to decide the choice of a phrase in a given position. (Decision trees of this kind were also used in an earlier version of TWIG.) Unlike in TWIG, our learner has no initial knowledge of words or grammar, there is no restriction on the number of new words in a sentence given to the learner, and our learner is allowed to interact with the teacher and possibly receive corrections.

Our work is also related to the work developed by Mooney and colleagues [12,13,33–35,51,52]. They introduced several systems that learn either from unambiguous or ambiguous data. In the first case, pairs consisting of sentences and their meaning representations are given to the learner. In the second case, the learner receives sentences paired with potential meaning representations (only one of the meanings is correct). Some of the systems are focused only on the comprehension task (e.g., [33]), but others on both comprehension and generation tasks (e.g., [12]). Our work mainly differs from theirs in that our learner learns from pairs consisting of a sentence and the situation in which this sentence has been produced. A situation is a description of what the learner can see in the world, and not a meaning or a set of candidates meanings for that sentence. Therefore, our learner constructs the meanings of the sentences from scratch. Moreover, our learner interacts with the teacher and can be corrected.

Jack introduced in [31] a computational model that produces behavior comparable to the stages of children's linguistic development. The model takes syllable-segmented input, learns to associate words with meanings under referential uncertainty, discovers compositional combinations of words, learns to make appropriate use of word ordering, and derives grammar rules and classes describing a language fragment. Our model has clear connections with Jack's model, but our learner is quite different from his learner: the input given to our learner can contain sentences that describe the situation only partially; our learner interacts with the teacher and can be corrected; our learner is not able to produce any sentences after receiving just one sentence from the teacher; and our learner can produce sentences that have a telegraphic character (i.e., sentences that are grammatically incomplete, but perfectly understandable).

A common feature of many of these approaches is the use of a miniature language environment. A task known as *Miniature Language Acquisition* was formalized by Feldman et al. [30]. The task consists of learning a subset of a natural

language from sentence-picture pairs that involve geometric shapes with different properties (color, size and position). Although this task is not as complex as those faced by children, it involves enough complexity to be compared to many real-word tasks. We use a simplified version of Feldman's task.

Other works that have also provided inspiration for our model are by Collins Hill [28] and Schaerlaekens [44]. They model the language of 2-year old children, and point out the relevance of semantics for the construction of the initial grammar of the child. Collins Hill introduced a repetition and response model in which the model produces responses to or repetitions of adult input sentences. This model fits the data collected from a single 2-year old girl, Claire. An interesting feature of this model is the use of a grammar composed of templates, that is, patterns for understanding and producing speech. Based on evidence provided by the study of two sets of Dutch-speaking triplets, Schaerlaekens proposed a semantic relations model focused on the 2-word stage. Schaerlaekens considers that 2-word sentences are semantically interpretable and have their own syntax. She also points out the importance of the context to analyze and interpret 2-word sentences. Using the semantic relations model, Schaerlaekens constructs a grammar that contains a number of possible sentence patterns (each expressing one semantic relation) found in children's 2-word sentences. Using this methodology, she constructs a descriptive grammar for the first 200 2-word sentences of each of the six children in the study.

Most works in the field of Grammatical Inference reduce the language learning problem to syntax learning and tend to omit semantic information. Isabelle Tellier was one of the first researchers in this area who tried to link syntax and semantics. In [47], she gives a very interesting theoretical account of the possible relationship between semantics and syntactic learning, and suggests that "the acquisition of a conceptual representation of the world is necessary *before* the acquisition of the syntax of a natural language can start." Our system reflects a similar assumption that the learner embarking on language learning has a considerable stock of semantic knowledge. Furthermore, Alexander Clark has also deeply studied the link between Grammatical Inference and First Language Acquisition. Most of his work is focused on unsupervised learning of natural languages and its relevance to first language acquisition. In [15,16], Alexander Clark and Shalom Lappin examine several formal models of learning and its connections to linguistic studies of language acquisition. Their analysis leads to favoring distributional information when considering the task of language acquisition, i.e., the learning of probabilistic models. Their arguments run parallel to the analysis we propose in this paper.

2.3. Positive versus negative evidence in language acquisition

Formal models of language acquisition have mainly focused on learning from positive data, that is, utterances that are grammatically correct. But a question that remains open is: Do children receive negative data and can they make use of it?

There have been three main proposals related to this question: i) Children do not receive negative information and rely only on innate information; ii) Children receive negative information in the form of different reply-types given in response to ungrammatical versus grammatical child utterances; iii) Children receive negative information in the form of reformulations (that is, utterances adults use in checking up on what children intended to say), and they not only can detect them but also make use of them. The issue of whether or not children receive negative evidence depends in part on how one defines this concept. Therefore, it seems important to define negative evidence exactly. Next we review these three proposals.

Chomsky's poverty of stimulus argument has been used to support the idea of human innate linguistic capacity. It is claimed that there are principles of grammar that cannot be learned from positive data only, and negative evidence is not available to children. Hence, since children do not have enough evidence to induce the grammar of their native language, the additional knowledge language learners need is provided by some form of innate linguistic capacity.

E.M. Gold's negative results in the framework of formal language learning have also been used to support this position. Gold proved that superfinite languages are not learnable from positive data only, which implies than none of the language classes defined by Chomsky to model natural language is learnable from positive data [24]. However, he suggests several hypotheses about how children overcome this theoretical hurdle. The first hypothesis is that the class of possible natural languages is much smaller than we would expect from our current models of syntax. The second hypothesis is that the child receives negative information by being corrected in a way we do not recognize. And a third possibility is that there is an a priori restriction on the class of texts that can occur, such as a restriction on the order of text presentation. Theoretical works have shown that the first hypothesis can lead to successful learning (e.g., results on learning restricted classes of formal languages using positive data by Angluin [2], Sakakibara [43] and Kanazawa [32]). In linguistics, it is also generally assumed that the first hypothesis holds. We suggest that it is also worth exploring the second hypothesis pointed out by Gold.

Brown and Hanlon [11] studied negative evidence understood as explicit approvals or disapprovals of a child's utterance (e.g.,"That's right" or "That's wrong"). They showed that there is no dependence between these kinds of answers and the grammaticality of children's utterances. These results were taken as showing that children do not receive negative data. But do these results really show this? It seems evident that parents rarely address their children in that way. During the first stages of language acquisition children make a lot of errors, and parents are not constantly telling them that their sentences are wrong; rather the important thing is that they can communicate with each other. However, it is worth studying whether other sources of negative evidence are provided to children. Is this the only form of negative data? Do adults correct children in a different way?

Some researchers have studied other kinds of negative data based on reply-types (e.g., Hirsh-Pasek et al. [29], Demetras et al. [20] and Morgan and Travis [38]). These studies argue that parents provide negative evidence to their children by

using different types of reply to grammatical versus ungrammatical sentences. Marcus analyzed such studies and concluded that there is no evidence that this kind of feedback (he called it "noisy feedback") even exists [37]. He argued for the weakness, inconsistency and inherently artificial nature of this kind of negative feedback. Moreover, he suggested that even if such feedback exists, a child would learn which forms are erroneous only after complex statistical comparisons. Therefore, he concluded that internal mechanisms are necessary to explain how children recover from errors in language acquisition.

Since the publication of the work of Marcus, the consensus seemed to be that children do not have access to negative data. However, a study carried out by Chouinard and Clark shows that this conclusion may be wrong [14]. First, they point out that the reply-type approach does not consider whether the reply itself also contains corrective information, and consequently, replies that are corrective are erroneously grouped with those that are not. Moreover, if we consider only reply-types, they may not help to identify the error made. Hence, Chouinard and Clark propose another view of negative evidence that builds on Clark's principle of contrast [17,18]. Parents often check up on a child's erroneous utterances, to make sure they have understood them. They do this by reformulating what they think the child intended to express. Hence, the child's utterance and the adult's reformulation have the same meaning, but different forms. Because children attend to contrasts in form, any change in form that does not mark a different meaning will signal to children that they may have produced an utterance that is not acceptable in the target language. In this way, reformulations identify the locus of any error, and hence the existence of an error. Chouinard and Clark analyze longitudinal data from five children between two and four years old, and show that adults reformulate erroneous child utterances often enough to help learning. Moreover, these results show that children not only detect differences between their own utterance and the adult reformulation, but that they make use of that information.

Our model is inspired also by Chouinard and Clark's results. Corrections (in form of reformulations) have a semantic component that has not been taken into account in previous studies. Hence, we propose a new computational model of language learning that gives an account of meaning-preserving corrections, and in which we can address questions such as: What are the effects of corrections on learning syntax? Can corrections facilitate the language learning process?

We note that the role of corrections in language learning has been studied in the field of Grammatical Inference, but from a syntactic point of view. In [9], a new type of query called a correction query was proposed for language learning. In a correction query, the learner asks the teacher whether a string is in the target language, and if the answer is negative the teacher returns a correction to the learner. In this work, a correction was chosen as a shortest extension of the queried string in the language, and it was shown that it is possible to learn DFA (Deterministic Finite Automata) from corrections with a considerably reduced number of queries. Later, these results were extended to k-reversible languages and pattern languages [48]. Another kind of correction based on edit distance was introduced in [10], and an efficient algorithm using this new type of correction was given to learn topological balls of strings (i.e., classes of languages defined via edit distance) [10]. These results were subsequently extended to pattern languages and regular expressions (which describe regular languages) [36]. All of these approaches consider syntactic corrections based on proximity between strings, and do not use semantic information. Our work takes into account the semantic component of corrections, and can be viewed as a first attempt to incorporate semantics in the field of Grammatical Inference.

3. The model

We describe the components of our model, and give examples drawn from the primary domain we have used to guide the development of the model.

3.1. Situation, meanings and utterances

A **situation** is composed of some objects and some of their properties and relations. These are not intended to be an exhaustive description of the state of the world, but to pick out some aspects of it that are of joint interest to the teacher and the learner. We assume that these objects, properties and relations are recognizable to both the learner and teacher from the outset. A situation is represented as a set of ground atoms over some constants (denoting objects) and predicates (giving properties of the objects and relations between them). For example, a situation s_1 consisting of a big purple circle to the left of a big red star is represented by the following set of ground atoms:

$$s_1 = \{bi1(t_1), pu1(t_1), ci1(t_1), le2(t_1, t_2), bi1(t_2), re1(t_2), st1(t_2)\}.$$

Here the two objects are represented by the constants t_1 and t_2, and the unary predicates are big (*bi1*), purple (*pu1*), circle (*ci1*), red (*re1*) and star (*st1*), and the sole binary predicate is the relation of one object being to the left of another (*le2*). The teacher and learner each have access to this representation of the situation.

Formally, we have a finite set P of **predicate symbols**, each of a specific arity (number of arguments), where the arity is appended to the name of the predicate. We also have a countable set of **constant symbols** $t_1, t_2, \ldots$, which are used to represent distinct objects. A **ground atom** is an expression formed by applying a predicate symbol to the correct number of constant symbols as arguments, for example, $re1(t_1)$ or $ab2(t_2, t_1)$.

We also have a countable number of **variables** $x_1, x_2, \ldots$. A **variable atom** is an expression formed by applying a predicate symbol to the correct number of variables as arguments, for example, $re1(x_2)$ or $le2(x_4, x_3)$. A **meaning** is a finite

sequence of variable atoms. Note that the atoms do not contain constants, and the order in which they appear is significant. Examples of meanings:

$$m_1 = (st1(x_1))$$
$$m_2 = (st1(x_1), le2(x_2, x_1), pu1(x_2), ci1(x_2))$$
$$m_3 = (bi1(x_1))$$
$$m_4 = (pu1(x_1), ci1(x_2))$$

A meaning is **supported** in a situation if there exists a **support witness**, that is, a mapping of its variables to *distinct* objects in the situation such that the image under the mapping of each atom in the meaning appears in the situation. If a meaning is supported in a situation by a unique support witness then it is **denoting** in the situation.

The meaning m_1 is supported in the situation s_1 via the witness $x_1 \to t_2$, the meaning m_2 is supported in the situation s_1 via the witness $x_1 \to t_2, x_2 \to t_1$, and the meaning m_3 is supported in the situation s_1 by either the witness $x_1 \to t_1$ or the witness $x_1 \to t_2$. The meaning m_4 is not supported in the situation s_1 because any support witness must map the variables x_1 and x_2 to distinct objects. Meanings m_1 and m_2 are denoting in the situation s_1, but meaning m_3 is not; intuitively, the predicate *bi1* by itself could refer to either object. We assume that both the teacher and learner can determine whether a meaning is denoting in a situation.

We also have a finite alphabet W of words. An **utterance** is a finite sequence of words, for example, *the star* or *the star to the right of the purple circle* or *star of circle small the green*. The **target language** is the set containing every utterance such that there exists some situation in which the teacher could produce that utterance. In our example, this includes utterances like *the star* or *the triangle to the right of the purple circle* but not *star of circle small the green*. We assume each utterance in the target language is assigned a unique meaning; in our examples, a finite state transducer is used both to recognize utterances in the target language and to assign them meanings.

An utterance is **denoting** in a situation if the meaning assigned to utterance is denoting in the situation. Intuitively, an utterance is denoting if it uniquely picks out the objects it refers to in a situation. Thus, in this model, an utterance is "appropriate" to a situation if it is denoting in that situation. For example, if the utterance *the star* is assigned the meaning m_1 then it is denoting in the situation s_1 and if the utterance *the star to the right of the purple circle* is assigned the meaning m_2, it is also denoting in the situation s_1. If the situation contained two stars, then the utterance *the star* would not be denoting. Note that a denoting utterance does not have to be "minimal" – it may specify more information than necessary to pick out one object.

3.2. The target language and meaning transducers

We would like a method of specifying the linguistic competence of the teacher, including the target language of utterances and their meanings. This representation should enable production by the teacher, that is, given a situation, the teacher should be able to choose an utterance in the target language that is denoting for the situation. For the model developed in this paper, we represent this competence by a finite state transducer that both recognizes the utterances in the target language and translates each correct utterance to its meaning.

Recall that W is the finite set of words and P is a finite set of predicates. Let A denote the set of all variable atoms over P. We define a **meaning transducer** M with input symbols W and output symbols A as follows. M has a finite set Q of states, an initial state $q_0 \in Q$, a finite set $F \subseteq Q$ of final states, a deterministic transition function δ mapping $Q \times W$ to Q, and an output function γ mapping $Q \times W$ to $A \cup \{\varepsilon\}$, where ε denotes the empty sequence.

The transition function δ is extended to define $\delta(q, u)$ to be the state reached from q following the transitions specified by the utterance u. The **language** of M, denoted $L(M)$ is the set of all utterances $u \in W^*$ such that $\delta(q_0, u) \in F$. A state $q \in Q$ is **live** if there exists an utterance u such that $\delta(q, u) \in F$, and **dead** otherwise. For each utterance u, we define the **output** of M, denoted $M(u)$, to be the finite sequence of non-empty outputs produced by starting at state q_0 and following the transitions specified by u, that is, $M(u)$ is the **meaning** of u.

As an illustration, we describe a limited sublanguage of Spanish involving geometric shapes and their properties and relative locations. W contains the words *el*, *la*, *circulo*, *cuadrado*, *triangulo*, *rojo*, *azul*, *verde*, *a*, *izquierda*, *derecha*, *encima*, and *del*. P contains the predicate symbols *ci1*, *sq1*, *tr1*, *bi1*, *re1*, *bl1*, *gr1*, referring to the properties of being a circle, a square, a triangle, big, red, blue, green, and also *le2* and *ab2*, referring to the relations of one object being to the left of or above another object. Note that in this example there is a predicate for big but no word for big in the vocabulary. Also, there are words in W for the relations left (*izquierda*), right (*derecha*) and above (*encima*), but the word for below (*debajo*) is omitted from W in this example sublanguage.

We define a meaning transducer M_1 for the limited sublanguage as follows. The states are 0 through 10; 0 is the initial state and the final states are $\{2, 3, 8, 9\}$. The transition function is given in Table 1, and the automaton is pictured in Fig. 2. Unspecified transitions go to the non-final dead state, 10.

The language of this transducer is finite and contains 444 utterances each with a distinct meaning. Examples are *el triangulo rojo* and *el circulo a la derecha del triangulo azul*, which have meanings of $(tr1(x_1), re1(x_1))$ and $(ci1(x_1), le2(x_2, x_1), tr1(x_2), bl1(x_2))$, respectively.

Table 1
Transitions and outputs of the transducer M_1.

State	Word	Next state	Output
0	*el*	1	ε
1	*circulo*	2	$ci1(x_1)$
1	*cuadrado*	2	$sq1(x_1)$
1	*triangulo*	2	$tr1(x_1)$
2	*rojo*	3	$re1(x_1)$
2	*azul*	3	$bl1(x_1)$
2	*verde*	3	$gr1(x_1)$
2	*a*	4	ε
3	*a*	4	ε
4	*la*	5	ε
5	*izquierda*	6	$le2(x_1, x_2)$
5	*derecha*	6	$le2(x_2, x_1)$
2	*encima*	6	$ab2(x_1, x_2)$
3	*encima*	6	$ab2(x_1, x_2)$
6	*del*	7	ε
7	*circulo*	8	$ci1(x_2)$
7	*cuadrado*	8	$sq1(x_2)$
7	*triangulo*	8	$tr1(x_2)$
8	*rojo*	9	$re1(x_2)$
8	*azul*	9	$bl1(x_2)$
8	*verde*	9	$gr1(x_2)$

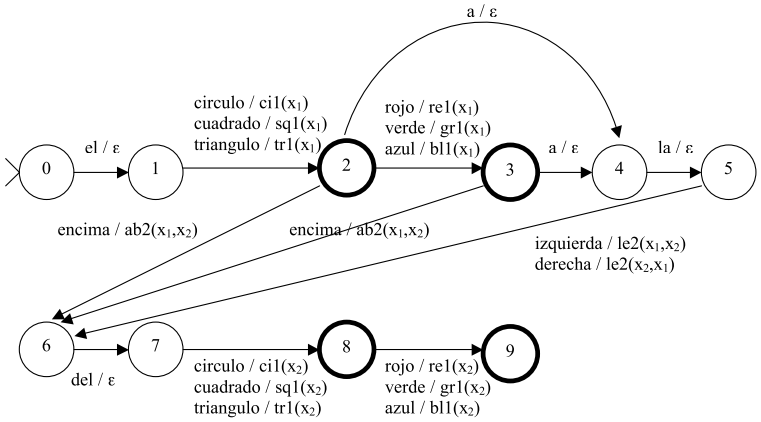

Fig. 2. Meaning transducer M_1.

3.3. The learning task

The learning task is characterized by a meaning transducer, a process that produces situations, and the behavior of the teacher. The goal of the learner is to learn a grammar for the language that will enable it to produce all and only the denoting utterances for any given situation.[1]

The learner gathers information about the language by engaging in a sequence of interactions with the teacher, where each interaction is related to a new situation. In the sequence of interactions between teacher and learner, each interaction is numbered consecutively starting with zero.

Initially the teacher and learner know all of the predicates P that will occur in situations, and are able to determine whether a given meaning is denoting in a given situation. A situation is a collection of objects, properties and relations, given to the learner and teacher as a set of ground atoms. Thus the learner and teacher share the same description of the situation.

The learner and teacher both also know a shared set of **categories** that classify a subset of the predicates into similarity groups. The categories facilitate generalization by the learner, and are used by the teacher in analyzing incorrect utterances of the learner. In our geometric shape examples, shape predicates are one category, color predicates are another, and size predicates are a third category, but no category contains the positional relations.

[1] We refer to the learner's representation as a grammar, although it does not take the form of a classical grammar from formal languages, but, as will be seen, a collection of general forms and decision trees.

Initially the teacher also has the meaning transducer for the target language, which it uses both to produce utterances appropriate to the given situation, and to analyze the utterances of the learner. Initially the learner has no language-specific knowledge, just the predicates and categories, which are common to all the target languages. The interactions of the learner and teacher are described in detail in the next section.

4. The interaction of learner and teacher

Here is a summary of one interaction between learner and teacher. In this section we describe the algorithms used by the learner and teacher to carry out the steps of this process. In addition, we present the pseudocode of these algorithms in Appendix C.

1. A new situation is generated and presented to the learner and teacher.
2. The learner uses its current grammar to attempt to produce an utterance appropriate to the situation.
3. The teacher analyzes the learner's utterance (if any) in the context of the situation. The learner's utterance may be correct, have an error in form or an error in meaning, or be uninterpretable.
4. If the learner's utterance has an error in form or an error in meaning, the teacher decides randomly (using the correction probability) whether to correct the learner.
5. The teacher produces a random utterance appropriate to the situation: either a correction of the learner's utterance or a randomly drawn denoting utterance.
6. The learner analyzes the teacher's utterance and updates its current grammar for the language as appropriate.

4.1. Beginning to learn words: the co-occurrence graph

In each interaction, the learner is presented with a new situation, and attempts to produce an utterance appropriate to that situation. Because the learner initially has no language-specific knowledge, for the first few interactions the learner produces no utterance. When the learner produces no utterance, there is nothing for the teacher to respond to, so the teacher produces a randomly drawn denoting utterance for the situation, and the learner receives that utterance and uses it to learn more about the target language and its semantics.

In processing the teacher's utterance, the learner records the words that occur in it and the predicates that occur in the corresponding situation in a **co-occurrence graph** G. The nodes of the graph correspond to words and predicate symbols, and there is an undirected edge between every pair of nodes, corresponding to two words, two predicate symbols, or a word and a predicate symbol. Each node u has an occurrence count, $c(u)$, recording the number of utterances or situations its corresponding word or predicate symbol has occurred in. Each edge (u, v) also has an occurrence count, $c(u, v)$, recording the number of utterance/situation pairs in which the word or predicate symbol at the two endpoints have occurred together.

From G the learner derives another graph, the **implication graph** H, which is a directed graph with the same set of nodes as G and edges defined as follows. For each ordered pair of nodes (u, v), define $p(u, v) = c(u, v)/c(u)$. Intuitively, $p(u, v)$ may be thought of as a conditional probability: given that u occurs in an utterance/situation pair, $p(u, v)$ is the probability that v also occurs. If $p(u, v)$ is close to 1, then nearly every time u occurs, v also occurs. Because we want to be able to tolerate errors at some rate, there is a noise threshold θ (currently 0.95) and we say that u **approximately implies** v if $p(u, v)$ is not less than θ. The directed edges (u, v) of the implication graph are those pairs (u, v) such that u approximately implies v. The implication graph is recomputed after each situation/utterance pair.

If the teacher produces utterances using the meaning transducer M_1, after the learner receives a number of situation/utterance pairs, we expect that the word *cuadrado* will approximately imply the predicate symbol *sq1* and will not approximately imply any other predicate symbol, thus giving part of the meaning of the word *cuadrado*. Similarly, after a sufficient number of situation/utterance pairs, we expect that the word *rojo* will approximately imply the predicate symbol *re1* and the words *izquierda* and *derecha* will approximately imply the predicate symbol *le2*. Because the word *el* occurs in every situation/utterance pair, we expect that it will not approximately imply any predicate symbol. Thus, to a first approximation, it seems that the predicate symbol(s) (if any) approximately implied by a word should give part of the meaning of that word, after enough situation/utterance pairs.

As an example of some of the complications that arise in using this idea, consider the word *a*, which in this setting only occurs in the phrases *a la izquierda* and *a la derecha*. After sufficiently many situation/utterance pairs, the word *a* will also approximately imply the predicate symbol *le2* and will not imply any other predicate symbol, and similarly for the word *la*. It might be argued that this is an accidental consequence of the fact that we deal with a very limited sample of the language, which is true. However, children must deal robustly with a limited (but expanding) sample of the language they are learning, so it seems important to be able to cope in some way with such artifacts.

If we think about the relationships between the words *a*, *la*, *izquierda* and *derecha* in our setting, *izquierda* and *derecha* imply the presence of *a* and *la*, and the words *a* and *la* always co-occur. In some sense, the "proximate cause" of presence of the predicate symbol *le2* in the situation is the co-occurring pair of words *a* and *la*. This notion of "proximate cause" is captured graph-theoretically by taking a **transitive reduction** H^r of the implication graph H as follows [1]. First, edges from predicates to words are removed. Then, each strongly connected component of the resulting graph is contracted to a vertex,

the transitive reduction of the resulting acyclic graph is computed by removing all edges (u, v) such that there is a directed path of length at least two from u to v, and the strongly connected components are expanded again.

In our example the effect of this operation is to remove the implication edges from *izquierda* and *derecha* to the predicate symbol *le2*, while retaining the implication edges from *a* and *la* to the predicate symbol *le2*. In an expanded language sample in which *a* and *la* were used in situations not involving the predicate symbol *le2*, this artifact would disappear, and there would be implication edges from *izquierda* and *derecha* to the predicate symbol *le2*.

So, to summarize, the learner uses the situation and the teacher's utterance to update the co-occurrence graph, the implication graph, and the transitively reduced implication graph. The information contained in these structures is used by the learner in its attempts to produce its own utterances and to comprehend further utterances of the teacher.

4.2. Comprehension by the learner

Given a situation and an utterance by the teacher, the learner uses the transitively reduced implication graph H^r in an attempt to determine the meaning of the teacher's utterance as follows. For each successive word of the teacher's utterance, the learner finds the list of all predicate symbols such that there is an edge from the word to the predicate symbol in H^r. Some of the predicate symbols may then be removed from each word's list as follows.

A **background predicate** is one whose predicate symbol has occurred in nearly all situations, that is, in a fraction of situations greater than the noise threshold θ. Such a predicate symbol is approximately implied by every word, and is removed from the list of each word in the utterance. Also removed from each list is any predicate symbol whose arity is greater than the minimum arity of any predicate symbol on the list. For example, if there are both binary and unary predicates on the list, the binary predicates are removed. Finally, if two (or more) different words in the utterance are in the same strong component of H^r, all the predicate symbols are removed from the list of each of the equivalent words except for the one that occurs rightmost in the utterance. The choice of one of the set of equivalent words is to avoid repeated predicate symbols; that the choice is rightmost is somewhat arbitrary. The result of this process is a list of lists of predicate symbols, one for each word in the teacher's utterance.

For example, early in one run of learning with M_1, the learner translated the teacher's utterance *el cuadrado rojo* into the following sequence of lists of predicate symbols:

$$((), (gr1, sq1), (re1)).$$

In the reduced implication graph at that point, no predicate symbols were associated with the word *el*, the two predicate symbols *gr1* and *sq1* were associated with the word *cuadrado* and the predicate symbol *re1* was associated with the word *rojo*. Evidently the data at that point were not sufficient to rule out the predicate symbol *gr1* as a possible meaning for *cuadrado*.

The learner forms a set of sequences of predicate symbols by taking (in order) one predicate from each non-empty list of predicate symbols in all possible ways. For the preceding example, that would produce two possible sequences of predicate symbols, namely, $(gr1, re1)$ and $(sq1, re1)$.

The resulting set S of sequences of predicate symbols is then compared with the situation to try to determine the teacher's meaning. In particular, the learner computes a variable-normalized representative of every meaning that is both supported by the situation and is such that its sequence of predicate symbols is in S – these become the learner's guess of the **possible meanings** of the teacher's utterance in the given situation. To continue the preceding example, in the interaction in question, the situation consisted of a big red square above a big blue circle. Thus, the meaning $(sq1(x_1), re1(x_1))$ is both supported by the situation and has one of the two predicate sequences in S, namely, $(sq1, re1)$. This is the only possible meaning that the learner guesses for this utterance because there is no supported meaning with the predicate sequence $(gr1, re1)$, and no other inequivalent supported meaning with the predicate sequence $(sq1, re1)$.

If the learner finds exactly one possible meaning for the teacher's utterance, the learner takes this to be "the" meaning intended by the teacher, an assumption that may or may not be correct. In this case, the learner takes the unique meaning as the basis of a possible general form for meanings in the target language. Specifically, the learner uses its prior knowledge of the categories of the predicate symbols to generalize the unique meaning to a **general form**, by replacing each predicate symbol by its generalization (if any) in the set of categories. In our example, the category of *sq1* is *shape1* and the category of *re1* is *color1*, so the learner generalizes the unique possible meaning $(sq1(x_1), re1(x_1))$ to the general form $(shape1(x_1), color1(x_1))$.

The general form is added to the (initially empty) set of possible general forms of meanings in the target language. The general forms acquired by the learner are the basis of the learner's own attempts to produce utterances relevant to a given situation. In particular, at any point, the current production grammar of the learner is represented by the list of general forms it has acquired, together with rules (not discussed yet) for how they may be instantiated.

To facilitate the "aging out" of incorrect general forms that are acquired early in the learner's experience, the learner records the interaction number of the most recent teacher utterance with a unique possible meaning that matched the general form. These values are used in the learner's production process in such a way that general forms that are not repeatedly matched to teacher utterances become less and less likely to be used by the learner.

If the learner finds no possible meanings, or more than one, for the teacher's utterance, the learner does not attempt to update the information in its set of general forms. In this model we do not attempt to quantify the learner's comprehension,

preferring to measure the learner's progress by its productions. The comprehension process of the learner could be refined in various ways; for example, if there are several possible meanings and only one is compatible with the learner's current grammar, that one could be taken to be "the" teacher's meaning. We have not pursued this direction here.

4.3. Production by the learner

The general forms acquired by the learner are the basis on which it attempts to produce utterances appropriate to situations. Recall that each general form is a sequence of variable atoms, where the predicates may be from P or may be category symbols. Each general form is a template denoting a set of possible meanings. That is, all the meanings obtained by substituting a corresponding predicate from P for each category symbol is the set of meanings generated by that general form. For example, by substituting one of the predicates *sq1*, *ci1* or *tr1* for *shape1* and one of the predicates *bl1*, *gr1*, or *re1* for *color1* in the general form $(shape1(x_1), color1(x_1))$, we get one of nine possible meanings, for example, $(ci1(x_1), bl1(x_1))$. The learner's categories are the basis it uses to generalize from the single meaning $(sq1(x_1), re1(x_2))$ to this set of nine possible meanings.

In attempting to produce an utterance appropriate to the current situation, the learner finds all the meanings generated by its general forms using predicates from the current situation, and tests each meaning to see if it is denoting in the current situation, producing a set of possible denoting meanings for this situation. If the set is empty, the learner produces no utterance. Otherwise, it attempts to translate each denoting meaning into an utterance with that meaning, producing a set of possible denoting utterances; this process is described in the next section.

To produce one utterance for the current situation, the learner selects one of the possible denoting utterances with a probability proportional to the square of the interaction number stored with the general form used to produce the corresponding meaning. Each time a general form is matched, the interaction number will be updated, so that repeatedly matched general forms will continue to be used by the learner, while incorrect general forms will be matched only a few times, and will decline in probability towards zero. This allows the learner gradually to abandon incorrect general forms as its understanding improves.

4.4. The learner: from meaning to utterance

A meaning is a sequence of variable atoms, for example,

$$(ci1(x_1), re1(x_1), le2(x_2, x_1), tr1(x_1)).$$

The learner processes this sequence to produce a sequence of words, that is, an utterance. A very basic approach is to process atoms in turn, replacing each atom with a word that approximately implies the atom's predicate. For this example, the result could be the utterance *circulo rojo derecha triangulo*, a kind of "telegraphic speech" that is understandable but not grammatically correct. A somewhat more complex strategy by the learner generally achieves grammatical correctness in our example domain. However, this strategy is clearly not sufficient in general for natural language.

A meaning is a sequence $(a_1, a_2, \ldots, a_k)$ of atoms. We associate with it two sequences of positions: the **atom positions** $1, 2, \ldots, k$ and the **gap positions** $0, 1, \ldots, k$. The atom positions refer to their corresponding atoms, and the gap position i refers to the position to the right of atom i, except that gap position 0 is the position to the left of atom a_1. The learner generates an utterance by generating a sequence of zero or more words for each position of the meaning in left to right order, that is, first gap position 0, then atom position 1, then gap position 1, then atom position 2, and so on, until gap position k. The sequences of words generated are concatenated to form the final utterance.

In our system, the choice of what sequence of words to produce for each position in a meaning is represented by two sets of decision trees, for the atom positions and the gap positions. For example, once the learner has learned enough about the 68-form Spanish mini-language, the meaning

$$(ci1(x_1), le2(x_2, x_1), el1(x_2), re1(x_2))$$

could be rendered into the words *el circulo a la derecha de la elipse roja*, by using the decision tree for gap 0 to generate *el*, the decision tree for atom 1 ($ci1$) to generate *circulo*, the decision tree for gap 1 to generate no words, the decision tree for atom 2 ($le2(x_2, x_1)$) to generate *a la derecha*, the decision tree for gap 2 to generate *de la*, and the decision tree for atom 3 ($el1$) to generate *elipse*, the decision tree for gap 3 to generate no words, the decision tree for atom 4 ($re1$) to generate *roja*, and the decision tree for gap 4 to generate no words. More details of the atom and gap decision trees follow.

For each variable atom that the learner has encountered in a unique teacher meaning, there is a decision tree that determines what sequence of words to produce for that atom in the context of the whole meaning. As an example, consider the atom $re1(x_2)$. In the sublanguage of Spanish represented by the transducer in Fig. 2, this is unconditionally translated as *rojo*. However, in the 68-form mini-language, in which there are feminine nouns for shapes, the decision tree to decide how to translate $re1(x_2)$ branches on the value of the shape predicate applied to x_2 to select either *rojo* or *roja* as appropriate. These trees are used to determine the sequence of words produced for the atoms occurring in the atom positions of a word.

To handle the gap positions, for each generalization of a variable atom that has been encountered, there is a decision tree. The purpose of this decision tree is to determine the sequence of words produced for the gap position i immediately following atom position i, where the atom in atom position i matches the generalized atom associated with this decision

tree. As in the case of decision trees for atom positions, decision trees for gap positions branch on atoms in the meaning being translated to words. Gap position 0 does not follow any atom position, and has a separate decision tree.

For example, in the 68-form mini-language for Spanish, there are both masculine and feminine nouns for shapes and the learner learns a decision tree for a gap position following the atom $le2(x_2, x_1)$ that branches on the value of the atom in the meaning that matches $shape1(x_2)$, choosing *del* for shapes denoted by masculine nouns and *de la* for shapes denoted by feminine nouns.

If there is no decision tree associated with a given atom or gap position in a meaning, the learner falls back on a "telegraphic speech" strategy as follows. For a gap position with no decision tree, no words are produced. For an atom position whose atom has no associated decision tree, the learner searches the transitive reduction of the approximate implication graph for words that approximately imply the predicate of the atom. If there are several such words, the one that has occurred in the most situations is chosen to replace the atom in the meaning.

4.5. The teacher's response to the learner's utterance

If the learner produces an utterance, the teacher analyzes it and then chooses its own utterance for the situation. The teacher may find the learner's utterance correct, incorrect but correctable, or incorrect and uncorrectable. In the case that the learner's utterance is found incorrect but correctable, the teacher chooses a possible correction for it. Then the teacher randomly chooses whether or not to use the correction as its utterance; a **correction probability** parameter governs this choice. If the learner produced no utterance or an uncorrectable one, or if the teacher did not choose to correct the learner's utterance, then the teacher's utterance is chosen uniformly at random from the denoting utterances for the situation. Thus, when the learner's utterance is correct there is some probability that the teacher will simply repeat it.

The process used by the teacher to analyze the learner's utterance is as follows. If the learner's utterance is equal to one of the correct denoting utterances for the situation, the teacher classifies the learner's utterance as **correct**. If the learner's utterance is not correct, the teacher "translates" the learner's utterance into a sequence of predicates by using the adult meaning transducer for the language. In particular, each word for which there is a non-empty output in the transducer is replaced by the predicate from one such output. For example, the teacher translates the incorrect learner utterance *el elipse pequeno* into the predicate sequence $(el1, sm1)$ despite the errors of agreement in the utterance. In more detail, the teacher's transducer has no output for *el*, has an output of *el1* for *elipse*, and an output of *sm1* for *pequeno*, which yields the indicated teacher translation of the incorrect utterance.

If the resulting sequence of predicates is the same as the sequence of predicates in at least one meaning obtained from a correct denoting utterance, the learner's utterance is classified as having an **error in form**. In this case, the learner is judged to have chosen a correct meaning but an incorrect form to express that meaning. The goal of the teacher is to choose a possible correction with the same meaning as that intended by the learner. A possible correction is chosen by the teacher by considering the set of denoting utterances whose meanings have the same sequence of predicates and choosing one such utterance using a measure of similarity to the learner's utterance. For example, if $(el1, sm1)$ corresponds to a denoting utterance for the situation, the teacher may choose *la elipse pequena* as a possible correction for *el elipse pequeno*.

If the learner's utterance is not correct and its corresponding sequence of predicates is not equal to the sequence of predicates for any denoting utterance for the situation, the teacher uses a measure of similarity between the sequence of predicates for the learner's utterance and the sequences of predicates corresponding to denoting utterances for the situation to determine whether there is a "close enough" match between the predicate sequences, determined by a threshold on a weighted edit distance on the sequences. If so, the teacher classifies the learner's utterance as having an **error in meaning** and chooses as the possible correction a denoting utterance whose predicate sequence is judged "most similar" to the predicate sequence for the learner's utterance. For example, if the learner's utterance is *el pequeno* and the predicate sequence $(el1, sm1)$ corresponds to a denoting utterance for the situation, the teacher may choose *la elipse pequena* as the possible correction.

If the learner produces an utterance but none of these cases (correct, error in form, error in meaning) apply, then the teacher classifies the learner's utterance as **uninterpretable** and does not offer a correction.

The teacher chooses a random denoting utterance for the situation and communicates it to the learner. Then, the learner analyzes the teacher's utterance and updates its grammar of the language as reflected in the co-occurrence graph, the general forms, and the decision trees for word choice. The decision trees are updated using a computed alignment between the teacher's utterance and the learner's understanding of the teacher's meaning, which assigns a (possibly empty) subsequence of words from the utterance to each gap or atom position in the meaning. Each subsequence of words is then added to the data for the corresponding decision tree.

If the learner has produced an utterance and finds that the teacher's utterance has the meaning intended by the learner, but is expressed differently, then the learner classifies the teacher's utterance as a **correction**. The learner could treat a perceived correction differently from other teacher utterances, but our system in fact treats them identically. This capability is included to explore the question of whether the learner can reliably detect corrections intended by the teacher. Of course, the learner may fail to recognize an intended correction of the teacher, and may also be mistaken when it classifies the teacher's utterance as a correction.

This analysis completes one interaction. A new situation is generated, and the cycle of learner production, teacher production, and learner analysis is repeated with the new situation.

To illustrate the process, we provide a commented excerpt from the first fifty interactions of a learner and teacher in Appendix A. These interactions show the learner beginning to comprehend the teacher's utterances and acquiring and using both incorrect and correct general forms, producing both incorrect and correct denoting utterances. It also shows the teacher comprehending enough of the learner's meaning to offer meaning-preserving corrections, some of which the learner recognizes as corrections.

5. Empirical results

We have implemented and tested our learning and teaching procedures in order to explore the following questions:

(1) Can the learner accomplish the learning task to a high level of correctness and coverage from a "reasonable" number of interactions (that is, well short of the number needed to memorize every legal situation/utterance pair)?

(2) What are the effects of correction or non-correction by the teacher on the learner's accomplishment of the learning tasks?

5.1. The learning tasks

The learning tasks we consider use the following set of situations. Each situation has two objects, each with three attributes (shape, color and size), and one binary relation between the two objects (above or to the left of). The attribute of shape has six possible values (circle, square, triangle, star, ellipse, and hexagon), that of color has six possible values (red, orange, yellow, green, blue, and purple), and that of size three possible values (big, medium, and small). Thus there are 108 distinct objects and 23,328 distinct situations. Situations are generated uniformly at random.

We consider limited sublanguages of natural language utterances related to these situations for several natural languages. The utterances in a situation are phrases intended to denote one of the objects, for example, *the circle*, *the orange star*, or *the small purple hexagon below the medium green square*. In our languages, the shape attribute is expressed as a noun and must appear, while the size and color attributes are expressed as adjectives and may be omitted. In the transducer for each language, the order of adjectives for size and color is fixed for that language. If the binary relation between the objects is "above", it may be expressed using either above or below; if it is "to the left of", it may be expressed using either left or right. Thus, there are 168 meanings referring to a single object and 112,896 meanings referring to two objects, for a total of 113,064 possible meanings. The number of meanings that are denoting in a situation varies from 32 (for a situation with two objects with identical attributes) to 40 (for a situation with objects of two different shapes).

In each of these languages, the 113,064 possible meanings are instances of 68 general forms: 4 referring to a single object and 64 referring to two objects. For English, examples of general forms are

$$(shape1(x_1)),$$

$$(color1(x_1), shape1(x_1))$$

and

$$(size1(x_1), shape1(x_1), le2(x_2, x_1), shape1(x_2)).$$

We refer to these languages as the **68-form languages**.

The artificial languages we constructed were each based on the words and grammar of a natural language, and for each one, we consulted at least one speaker of the language to help us construct a meaning transducer to translate appropriate phrases in the language to all 113,064 possible meanings. Each transducer was constructed to have exactly one accepted phrase for each possible meaning. Words were represented by strings of lower-case English letters, which necessitated using straightforward transliterations of languages other than English. Examples of teacher utterances in the 68-form languages with some notes on their construction are in Appendix B. The procedure implementing the teacher is language-independent, and takes as parameters the meaning transducer for the language and the desired correction probability.

To help understand the effect of different aspects of the learning problem, we also considered reduced sublanguages, consisting of the utterances that refer to a single object (168 utterances) and those that refer to two objects, but include all three attributes of both (46,656 utterances). Thus, for English the utterances in the reduced sublanguage include *the star*, *the blue triangle* and *the medium purple ellipse to the left of the medium red square* but not *the circle below the yellow hexagon*. Each meaning in the reduced sublanguage is an instance of one of 8 general forms, but most of the lexical and syntactic complexity of the language is preserved. We refer to these reduced sublanguages as the **8-form languages**.

5.2. How many interactions are needed to learn?

The level of performance of a learner attempting to learn a given language L is measured using two quantities: the correctness and completeness of the learner's utterances in a given situation. The learning procedure is equipped with a test mode, in which no learning takes place. In the test mode, the learner receives a situation, and responds with all the

Table 2
Number of interactions to reach the specified levels of performance for 68-form languages and probability 0.0 of correction. Each number is the median of 10 trials.

Level	0.60	0.70	0.80	0.90	0.95	0.99
English	200	200	300	400	500	700
German	200	300	300	400	550	800
Greek	**400**	**500**	**700**	**1500**	**2200**	**3400**
Hebrew	200	300	400	500	650	900
Hungarian	200	300	350	450	550	750
Mandarin	200	200	300	400	500	700
Russian	**450**	**500**	**850**	**1750**	**2350**	**3700**
Spanish	200	300	350	500	600	1000
Swedish	200	300	300	400	600	1000
Turkish	200	200	300	400	550	800

utterances it could produce in that situation, and, for each one, the learner's probability of producing it. To evaluate the correctness and completeness of the learner's responses, the meaning transducer for L is used to produce all the correct denoting utterances for the given situation. The **correctness** of the learner in the given situation is the sum of the probabilities of the learner's utterances that are in the correct denoting set. The **completeness** of the learner in the given situation is the fraction of the correct denoting utterances that appear in the set of learner utterances. The average of correctness and the average of completeness of the learner in 200 randomly generated situations is used to estimate the overall correctness and the overall completeness of the learner.

We say that a learner reaches a **level p of performance** if both correctness and completeness are at least p. That is, the minimum of the two measurements must be at least p. This is a more stringent measure of performance than the F-score, which would combine them via the geometric mean.

In our first set of trials, we set a target level of performance of $p = 0.99$ and the learner and teacher engage in a sequence of interactions until the learner first reaches this level of performance. Because the testing process is compute-intensive, the learner is tested only at intervals of 100 interactions. In Table 2 we show the number of interactions needed to first reach the given levels of performance for each 68-form language. Each entry is the median of 10 trials. For all these trials the teacher has a correction probability of 0.0.

In these results, there are two clear groups: one group is Greek and Russian, for which the median is at least 3400 interactions to reach level 0.99, and the other group is the rest of the languages, for which the median is at most 1000 interactions to reach level 0.99. The first observation is that the learner is able to achieve correctness and completeness of 0.99 for each of these sublanguages after being exposed to a small fraction of all possible situations and utterances. In the case of Russian, 3700 interactions involve at most 16.5% of all possible situations and at most 3.5% of all possible utterances by the teacher. The bound of 1000 interactions represents fewer than 4.3% of all situations and fewer than 1% of all possible utterances.

To give an idea of how the learner progresses in a single trial, we show the learning curves for a typical run for Hebrew in Fig. 3 and Russian in Fig. 4. The horizontal axis is the number of interactions between learner and teacher. Two of the curves show the measurements of correctness and completeness of the learner's utterances every 100 interactions, tested on 200 random situations as described above. The third curve shows the fraction of the total number of general forms that have been acquired by the learner at intervals of 100 interactions. It is clear from these examples that neither correctness nor completeness is necessarily monotonically increasing.

To achieve completeness of at least 0.99, the learner must acquire all 68 correct general forms. Typically the learner also acquires a small number (3 or 4) of incorrect general forms in early interactions, whose probabilities decrease as they fail to be matched in later interactions. In both of these runs, the learner acquired its last general form between interaction 900 and interaction 1000.

For the Hebrew run, the acquisition of the last general form essentially coincided with the achievement of the 0.99 level, while for the Russian run, levels of correctness and completeness continued to improve for another 2800 interactions after the last general form was acquired.

Russian and Greek require the largest number of interactions, but a natural question arises: is this for computational reasons (the size of the target machines) or linguistic reasons? In order to answer this question, we analyze the size of the alphabets (i.e., the number of words) and the size of the transducers (i.e., the number of states) for the different 68-form languages. Table 3 shows the results.

Concerning the size of the alphabet, we can clearly distinguish two groups: one consisting of Greek and Russian, which each have at least 57 words, and the other consisting of the remaining languages, for which the median alphabet size is 26.5. However, if we focus on the size of the transducers, we do not see such a strong difference between the languages. Greek and Mandarin have the biggest transducers, but close to them are, for example, German and Spanish.

If we compare these results with those in Table 2 (i.e., the number of interactions needed to reach the specified levels of performance for these languages), we see that the languages that required significantly more interactions (Russian and Greek) have the largest alphabet sizes. In particular, the median size of alphabets of languages that require at most 1000 interactions is less than half of that of languages that require at least 3400 interactions. Though Greek has the largest

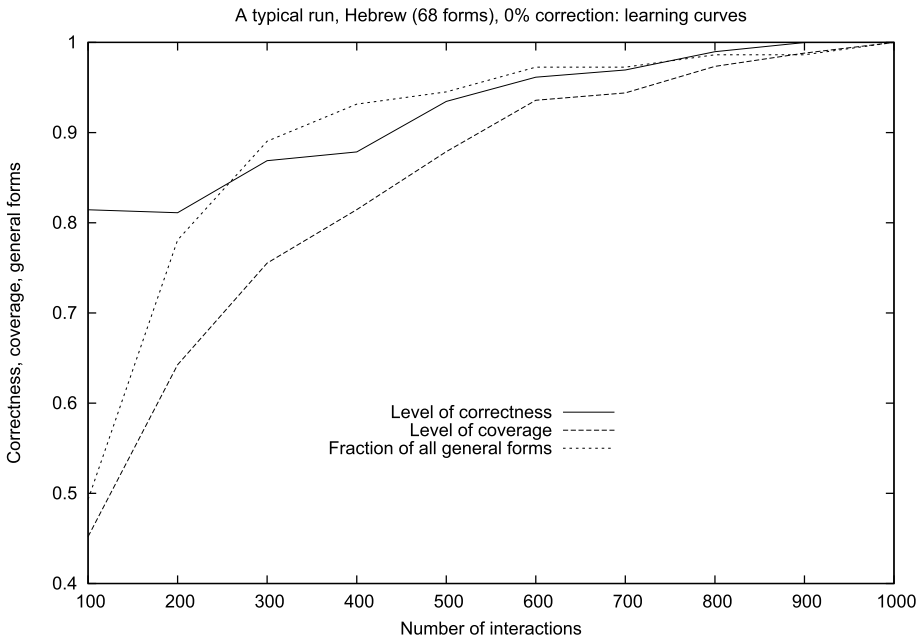

Fig. 3. Learning curves for Hebrew.

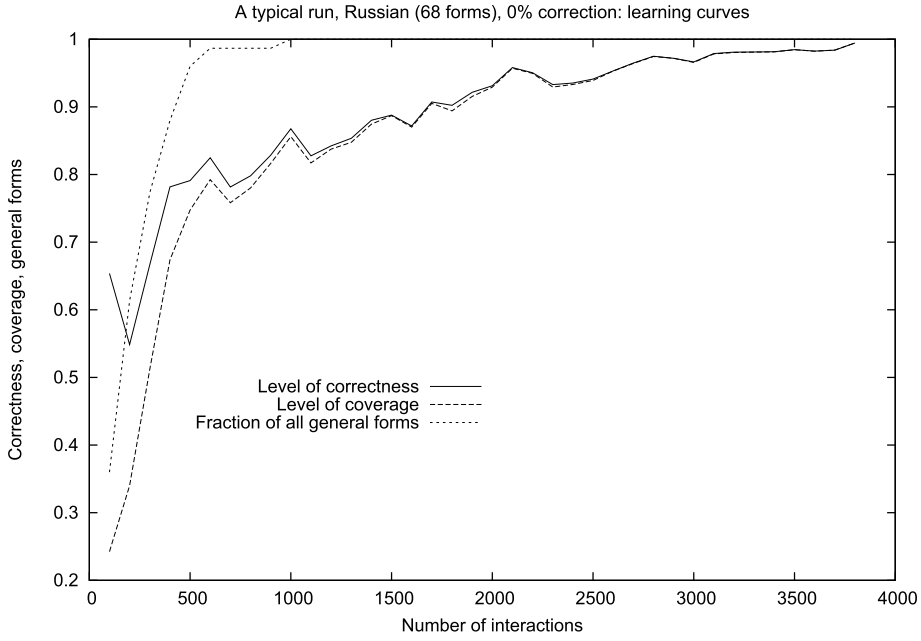

Fig. 4. Learning curves for Russian.

transducer, the size of the Russian transducer is around the overall median. Furthermore, Mandarin, which was one of the languages that required fewer interactions, has an alphabet of half of the size of that for Greek, but the Mandarin transducer has the second largest size.

Hence, the size of the transducers is not directly related to the number of interactions needed to reach a high level of performance, but the size of the alphabet seems to be more related. Therefore, we conclude that Russian and Greek seem to require more interactions due to linguistic reasons rather than computational reasons.

5.2.1. Comparison with an enhanced n-gram model

In this section we explore an enhanced n-gram approach to learning correct and complete models of the ten 68-form mini-languages. The goal is to learn a model that generates random utterances for the target mini-language with correctness

Table 3
Sizes of the alphabet and transducer for each 68-form language. The two largest sizes in each category are in bold type.

Language	Size of the alphabet	Size of the transducer
English	22	13
German	34	25
Greek	**57**	**37**
Hebrew	34	12
Hungarian	26	18
Mandarin	28	**31**
Russian	**60**	20
Spanish	27	24
Swedish	22	12
Turkish	25	8

Table 4
Percentage correctness of n-gram models, varying n.

n	2	3	4	5	6
English	74.5	100.0			
German		80.6	100.0		
Greek		96.3	98.6	100.0	
Hebrew	100.0				
Hungarian			58.9	91.4	100.0
Mandarin	86.0	100.0			
Russian	93.9	98.0	100.0		
Spanish	80.6	97.9	100.0		
Swedish	100.0				
Turkish	100.0				

at least 99% that is also capable of generating at least 99% of all the utterances in the target mini-language. This goal, which entirely ignores the corresponding situations, contrasts with the setting for our system, which must also learn the mapping from situations to denoting utterances.

We consider a model based on n-grams. Because each mini-language is finite (containing exactly 113,064 utterances), 100% correctness is definitely achievable by taking n to be the maximum length of any utterance in the language. However, to achieve at least 99% completeness with training sets of reasonable sizes, a much smaller value of n is desirable.

A basic n-gram model for a fixed value of n is trained as follows. Given a list of utterances, each is prefixed by $(n-1)$ start words and suffixed by one (distinct) end word, where neither the start word nor the end word occur in the list of utterances. For each consecutive tuple of $(n-1)$ words that occur in any utterance in the list, the 1-word continuations that occur in any utterance are recorded, together with their relative frequencies. This model is used to generate an utterance by starting with an $(n-1)$-tuple of the start word and randomly generating a 1-word continuation according to their relative frequencies, then shifting the $(n-1)$-tuple right by one word, and repeating this step until the end symbol is generated. The start and end words are removed from the resulting utterance.

This basic model tends to "loop" in some circumstances. For example, with $n=3$ and a sufficient sample of the English mini-language, the model may generate the utterance "The square to the left of the triangle above the ellipse." This utterance is not in the target language, which permits at most one occurrence of "above", "below", "left" or "right" in any utterance.

To avoid this kind of behavior, the basic n-gram model was enhanced to reject randomly generated utterances that are either (1) of a length not observed in the input list of utterances, or (2) contain words w_1 and w_2 in different positions of the utterance that have occurred in utterances fairly frequently, but never together in the same utterance. (The technical condition for "fairly frequently" was that the product of their frequencies in the sample be at least 0.03, a value that was chosen empirically to achieve good performance for the model.)

For this enhanced n-gram model, we measured correctness and completeness when learning from samples of the ten 68-form mini-languages. The measures of correctness are given in Table 4, which shows the effects of the choice of n for the different mini-languages. The values shown are the medians of five trials, each trained on an independent sample of 4000 utterances (produced by a non-correcting teacher for the domain) and tested on a sample of 1000 utterances generated by the trained enhanced n-gram model, using the teacher's transducer to judge correctness. Note that 100% correctness is attained for each language for a sufficiently large n. (The reason that n must be at least 6 to attain 100% correctness for the Hungarian mini-language is that the positional words "alatt" and "folott" appear at the ends of the utterances rather that between the phrases for the two objects.)

The measures of completeness are shown in Table 5, where we consider just $n=2$ and $n=3$. Each value is the median of five trials of the number of training utterances (tested at intervals of 100) to reach coverage of at least 99%, as measured by the fraction in an independent sample of 1000 utterances (generated by a non-correcting teacher) that could possibly be generated by the trained model.

Table 5
Number of utterances for 99% completeness, varying n.

n	2	3
English	200	1100
German	400	1900
Greek	1200	2700
Hebrew	900	5400
Hungarian	600	2400
Mandarin	200	400
Russian	1200	4200
Spanish	500	1800
Swedish	300	1700
Turkish	500	2700

Table 6
Comparison of n-gram models with our system.

	n	n-gram	Ours
English	3	1100	700
Greek	3	2700	3400
Hebrew	2	900	900
Mandarin	3	400	700
Russian	3	4200	3700
Spanish	3	1800	1000
Swedish	2	300	1000
Turkish	2	500	800

In Table 6 we compare the number of utterances to attain 99% completeness for 68-form mini-languages Hebrew, Swedish, and Turkish with $n = 2$ and for English, Greek, Mandarin, Russian and Spanish with $n = 3$ to the number of utterances required to attain 99% correctness and 99% completeness for these mini-languages by our system, as previously listed in Table 2. (These values of n were chosen as the first for which Table 4 gives at least 95% correctness, a more relaxed criterion. German and Hungarian are omitted because they require at least $n \geq 4$.)

The results in Table 6 do not show a clear advantage of one model over the other in terms of numbers of utterances to attain roughly comparable levels of performance in terms of correctness and completeness. The figures are within a factor of two of each other for all the mini-languages except Swedish, where enhanced bigrams gives 300 utterances to our system's 1000. However, our system also deals with German (800 utterances) and Hungarian (750 utterances), for which the n-gram method would require very large numbers of samples to achieve 99% completeness for $n = 4$ and $n = 6$, respectively. This comparison does suggest that the sample sizes required by our system to learn the 68-form mini-languages are reasonable.

5.3. How do corrections affect learning?

We have seen (in the detailed analysis in Appendix A) that the teacher can detect and classify errors on the part of the learner, and offer corrections related to its interpretation of the learner's meaning. For syntactic errors (errors in form), the teacher's correction has the same meaning as its interpretation of the learner's meaning, and for semantic errors (errors in meaning), the teacher's correction has a meaning close to its interpretation of the learner's meaning. We have also seen that the learner can detect corrections intended by the teacher. The learner may also classify a teacher utterance as a correction when it is not intended as a correction. In this section we attempt to quantify some of the effects of corrections on the learning process.

5.3.1. Corrections and 68-form languages

Table 7 shows the results of a set of trials parallel to those reported in Table 2, in which the teacher's correction probability is set to 1.0. This means that the teacher offers a correction to the learner every time it classifies the learner's utterance as an error in form or an error in meaning.

These results fall into the same two groups: Greek and Russian versus the rest of the languages tested. For Greek, the median number of interactions to reach the 0.99 level of performance with correction probability 0.0 is 3400 and with correction probability 1.0 it is 2600, a decrease of about 24%. For Russian, the median with correction probability 0.0 is 3700 interactions and with correction probability 1.0 is 2900, a decrease of about 21%. Thus for these two languages there is a clear decrease in the number of interactions required to achieve level of performance 0.99. However, the languages in the other group show no clear effect of corrections in these trials.

It is important to note that in these trials the learner processes a teacher utterance in *exactly the same way* regardless of whether the learner classifies it as a correction. That is, the learner does not do anything special for perceived corrections. Thus, the improvements in performance for Greek and Russian depend entirely on the difference in behavior of the teacher. In particular, the non-correcting teacher and the correcting teacher produce different distributions on utterances, and this

Table 7
Numbers of interactions needed to reach the given levels of performance with 68-form languages and probability 1.0 of correction. Each number is the median of 10 trials.

Level	0.60	0.70	0.80	0.90	0.95	0.99
English	200	200	300	400	500	750
German	200	300	400	500	500	750
Greek	400	500	700	1300	1700	2600
Hebrew	300	300	400	500	600	900
Hungarian	300	300	400	450	550	800
Mandarin	200	300	300	400	550	800
Russian	450	600	850	1500	2000	2900
Spanish	300	300	350	500	600	850
Swedish	200	300	300	500	600	900
Turkish	200	250	300	400	550	900

Table 8
Cumulative numbers of interactions, errors and corrections at 0.99 level of performance for 68-form languages, with 1.0 probability of correction. Each number is the median of 10 trials. The last column gives the percentage that the corrections column is of the utterances column.

	Utterances	Errors	Corrections	Percentage: c/u
English	750	25.0	11.5	1.5%
German	750	71.5	52.5	7.0%
Greek	2600	344.0	319.0	12.3%
Hebrew	900	89.5	62.5	6.9%
Hungarian	800	76.5	58.5	7.3%
Mandarin	800	50.0	31.5	3.9%
Russian	2900	380.0	357.0	12.3%
Spanish	850	86.0	68.0	8.0%
Swedish	900	54.0	43.5	4.8%
Turkish	900	59.0	37.0	4.1%

change in the learner's environment changes the performance of the learner. This suggests that studies of correction in natural languages should consider the possibility that corrections may affect the learning process *even when the learner cannot (or chooses not to) detect corrections.*

To illuminate these results further, Table 8 presents the numbers of utterances, errors and corrections to reach level of performance 0.99 for the trials shown in Table 7. This shows that the teacher offers corrections for a substantial fraction of learner errors, and the fraction increases with increasing numbers of errors. The last column shows the percentage of the total number of teacher utterances that were corrections. Greek and Russian show the highest percentages of corrections, while English shows the lowest. The percentages for the other languages are intermediate; for these languages, we might expect that corrections could have a demonstrable (though more subtle) effect on learner performance.

As the level of the learner's performance improves, there are several processes at work. The learner's comprehension of teacher utterances improves as more information about words and predicates accumulates in the co-occurrence graph. The learner acquires new correct general forms, and earlier-acquired incorrect general forms decrease in probability as they are not matched by teacher utterances. The learner also acquires more accurate rules for choosing phrases for meanings that are instances of general forms. The attainment of the 0.99 level of performance may be limited by the need to acquire all the correct general forms (as seems to be the case for the Hebrew trial in Fig. 3), or by the need to improve the correctness of the phrase choices (as seems to characterized the Russian trial in Fig. 4).

The difference in which process (form acquisition versus phrase choice improvement) is the bottleneck in attaining performance level 0.99 distinguishes the group consisting of Greek and Russian from the group consisting of the other languages in our sample. Information about the acquisition of general forms in these runs is shown in Table 9 for correction probability 0.0 and Table 10 for correction probability 1.0. In the case of Greek and Russian, most of the runs had acquired their last general form by the time the 0.90 level of performance was reached. The other languages show a more substantial change in the number of general forms between the 0.95 level and the 0.99 level in both conditions.

This data suggests that the bottleneck for Greek and Russian is the improvement of phrase choice, while for the other languages it is the acquisition of all 68 correct general forms. Because the teacher's corrections generally do not help with the acquisition of new general forms (in fact, very often the general form in a correction is the same one the learner just used), but do tend to improve the correctness of phrase choice, we do not expect correction to accelerate the attainment of the 0.99 level of performance when the bottleneck is the acquisition of general forms. This observation led us to construct reduced sublanguages with just 8 general forms to see if correction would have more effect when the bottleneck of acquiring general forms was removed. The next section describes results for these reduced sublanguages.

Table 9
Cumulative number of general forms acquired at each level of performance for 68-form languages and probability 0.0 of correction. Each number is the mean of 10 trials.

Level	0.60	0.70	0.80	0.90	0.95	0.99
English	51.7	54.9	61.1	66.5	69.5	71.6
German	50.2	60.4	62.4	68.0	69.9	71.4
Greek	58.7	65.6	69.4	72.9	73.0	73.1
Hebrew	55.9	61.7	66.2	68.9	70.6	71.6
Hungarian	54.3	61.4	64.2	67.1	69.4	71.6
Mandarin	49.3	53.9	59.8	66.5	68.6	70.3
Russian	65.9	68.9	72.0	72.5	72.5	72.5
Spanish	53.9	59.6	63.4	67.9	70.1	71.1
Swedish	49.7	57.9	60.3	65.9	68.0	70.1
Turkish	51.1	55.3	60.4	65.8	69.3	71.1

Table 10
Cumulative number of general forms acquired at each level of performance for 68-form languages and probability 1.0 of correction. Each number is the mean of 10 trials.

Level	0.60	0.70	0.80	0.90	0.95	0.99
English	48.5	55.0	59.2	65.7	69.0	70.9
German	47.0	56.8	60.7	66.5	68.7	70.5
Greek	66.3	68.4	70.6	71.5	71.5	71.5
Hebrew	52.7	58.0	62.8	66.6	68.7	71.2
Hungarian	55.3	57.8	62.9	66.4	68.9	71.6
Mandarin	48.0	55.2	59.7	66.0	68.4	70.7
Russian	61.2	67.7	72.6	74.0	74.0	74.0
Spanish	51.3	56.1	59.2	64.3	67.0	69.3
Swedish	47.1	57.5	60.3	65.4	68.0	70.3
Turkish	47.1	53.7	58.3	65.2	67.9	70.3

Table 11
Numbers of interactions needed to reach the given levels of performance with 8-form Spanish. Row labels are probabilities of correction. Levels of performance were tested every 50 interactions. Each number is the median value for the 100 trials.

Level	0.60	0.70	0.80	0.90	0.95	0.99
0.00	86.5	128.0	200.5	359.5	541.0	908.0
0.25	91.0	123.0	186.0	316.5	472.5	841.5
0.50	90.0	125.5	189.0	297.5	438.5	730.5
0.75	92.5	121.0	169.5	283.0	421.0	674.0
1.00	93.5	122.0	170.5	280.0	386.5	630.5

5.3.2. Correction and 8-form languages

The reduced sublanguages have just 8 general forms, which are acquired relatively early. In Table 11 we show the results of 100 trials each of the 8-form Spanish sublanguage, with correction probabilities of 0.0, 0.25, 0.50, 0.75 and 1.0. Levels of performance were tested every 50 interactions for these trials.

These results show an improvement of over 30% in the number of interactions to reach performance level 0.99 in going from correction probability 0.0 to correction probability 1.0. The intermediate correction probabilities give intermediate values. The corrected learners seem to be slightly slower to reach the 0.60 level of performance. The data on numbers of general forms for these runs show that learners had overwhelmingly acquired their last general form by the time they reached the 0.90 level of performance, where the superior performance of the corrected learners becomes very evident.

To give a more detailed sense of the comparison between the uncorrected and corrected learners, we include histograms of the numbers of interactions to reach the 0.99 level of performance in these trials for probability of correction 0.0 (in Fig. 5) and 1.0 (in Fig. 6). The histograms are displayed on the same scale and convincingly show the advantage of the corrected learners in this task.

Even though in the case of the 8-form languages there are only 8 correct general forms to acquire, the distribution on utterances with one object versus utterances with two objects is quite different from the case of the 68-form languages. For a situation with two objects of different shapes, there are 40 denoting utterances in the case of 68-form languages, of which 8 refer to one object and 32 refer to two objects. In the case of the 8-form languages, there are 10 denoting utterances, of which 8 refer to one object and 2 refer to two objects. Thus, for situations with two objects of different shapes (which are 5/6 of the total), utterances referring to two objects are 4 times more likely in the case of 68-form languages than in the case of 8-form languages. This means that if the learner needs to see utterances involving two objects in order to master

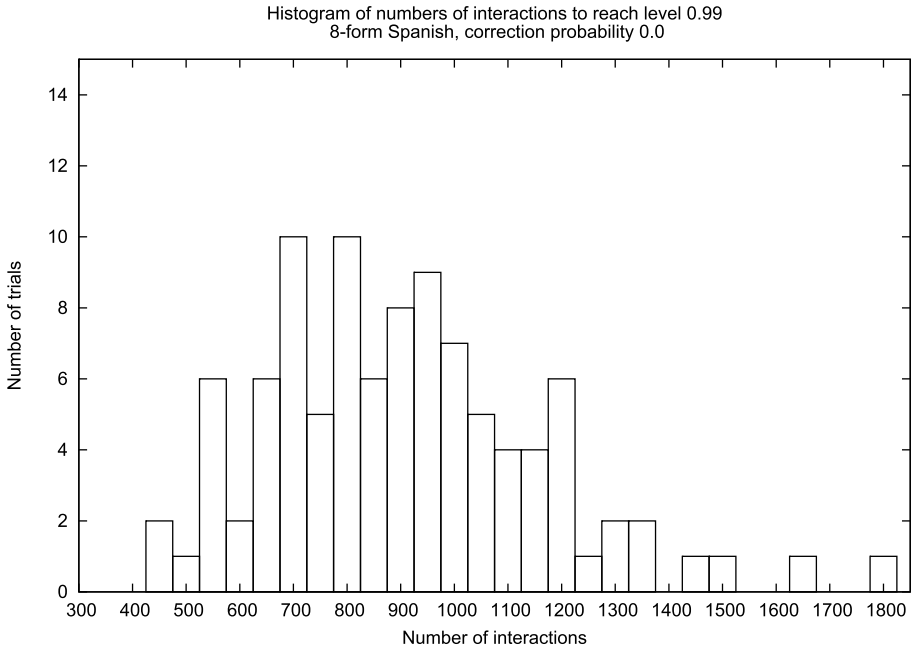

Fig. 5. Histogram of number of interactions to achieve level 0.99 in 100 trials of learning 8-form Spanish with correction probability 0.0. Levels of performance were tested every 50 interactions.

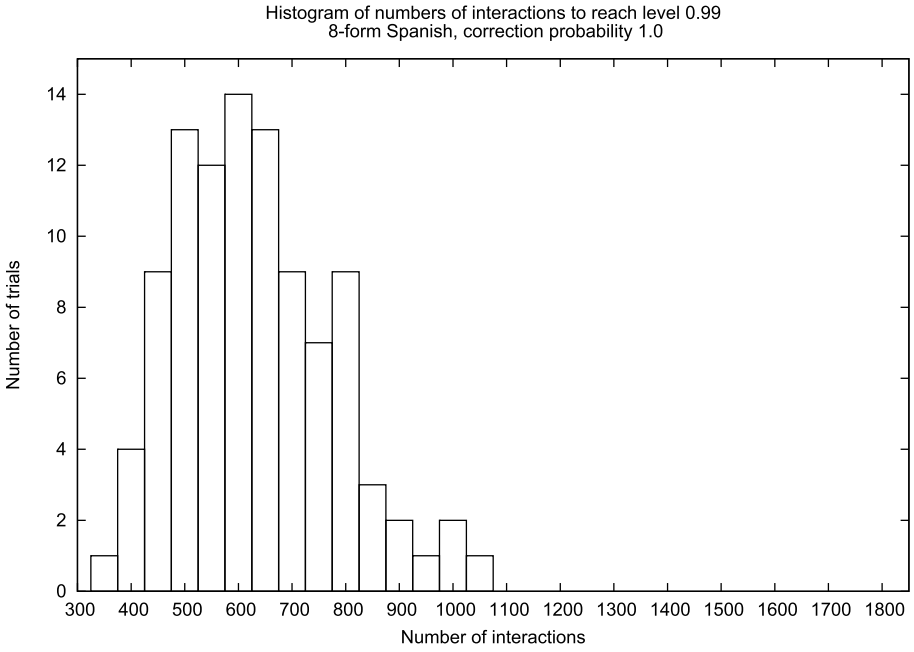

Fig. 6. Histogram of number of interactions to achieve level 0.99 in 100 trials of learning 8-form Spanish with correction probability 1.0. Levels of performance were tested every 50 interactions.

certain aspects of syntax (for example, cases of articles, adjectives and nouns), the waiting time is noticeably longer in the case of 8-form languages.

This longer waiting time emphasizes the effects of corrections, because the initial phase of learning is a smaller fraction of the whole. In Table 12 we show the percentage reduction in the number of interactions to reach the 0.99 level of performance from correction probability 0.0 to correction probability 1.0 for the 8-form languages. For each language, corrections produce a reduction, ranging from a low of 11.7% for Swedish to a high of 38.1% for Greek. This confirms our hypothesis that corrections demonstrably help the learner when the problem of acquiring all the general forms is not the primary bottleneck.

Table 12
Percentage reduction in the number of interactions to reach level 0.99 (except: 0.95 for Turkish) from probability of correction 0.0 to probability of correction 1.0 for 8-form sublanguages. Levels of performance were tested every 50 interactions. The numbers in the first two columns are means of 100 trials (except: 20 trials each for Greek and Russian). The third column shows the percentage reduction from the first column to the second.

	0.0	1.0	% reduction
English	247.0	202.0	18.2%
German	920.0	683.5	25.7%
Greek	6630.0	4102.5	38.1%
Hebrew	1052.0	771.5	26.7%
Hungarian	1632.5	1060.5	35.0%
Mandarin	340.5	297.5	12.6%
Russian	6962.5	4640.0	33.4%
Spanish	908.0	630.5	30.6%
Swedish	214.0	189.0	11.7%
Turkish	1112.0	772.0	30.6%

6. Collecting corrections

In this section we define and analyze a simplified model of the interactions between learner and teacher to get a better understanding of the possibilities and limitations of corrections. We model the process of learning with corrections as a modified version of the coupon collector process.

In the basic coupon collector process, there is a set of n tokens, each bearing a distinct positive integer from 1 to n. An agent draws tokens uniformly at random from this set with replacement, until it has drawn each number from 1 to n at least once. The expected number of draws until the agent has drawn each distinct token at least once is $nH(n)$, where $H(n)$ is the nth harmonic number, that is

$$H(n) = \sum_{k=1}^{n} 1/k.$$

Asymptotically this quantity is $n \ln n + O(n)$.

To model learning with corrections in our setting, we define a **correction collection** process, as follows. The learner has a set of n tokens, each bearing a distinct positive integer from 1 to n and also a **correctness bit** indicating whether the token is correct or not. The teacher has a similar set of tokens, except that each correctness bit is set to 1. For each token, the learner can only perceive the number on the token, not its correctness bit, while the teacher can perceive both the number on the token and its correctness bit. The tokens represent possible utterances of the learner and the teacher. We assume that the learner cannot determine the correctness of utterances for itself but that the teacher can. Moreover we assume that the teacher's utterances are all correct.

Now the learner and teacher engage in a sequence of interactions as follows. The learner draws one of its tokens uniformly at random with replacement and shows it to the teacher. The teacher chooses with replacement one of its tokens and shows it to the learner. If the teacher's token is numbered i, then the correctness bit on the learner's token numbered i is set to 1, indicating that it is now correct.

This is a very simplified model of the interactions in our system: different tokens represent different possible (atomic) utterances, the teacher can always understand what the learner intended to say, and the learner completely corrects its utterance with one exposure to the correct version. The interaction cycle is as follows: the learner produces an utterance (token), then the teacher produces an utterance (token), and the learner uses the teacher's utterance (token) as a model for correct utterances (tokens) of that kind in the future. Because utterances are treated atomically, the task of the learner is to collect correct versions of all n possible utterances.

We consider two possible strategies for the teacher's choice of a token to show to the learner. The non-correcting teacher ignores the token chosen by the learner and chooses one of its tokens uniformly at random to show to the learner. This is analogous to the teacher in our system using correction probability 0.0.

The correcting teacher looks at the token chosen by the learner and chooses its action based on the correctness bit. If the learner's token has correctness bit 0 and number i, then the teacher chooses its own token with number i, thereby ensuring that the learner's token i will henceforth be correct. If the learner's token has correctness bit 1, then the teacher chooses one of its own tokens uniformly at random. This is analogous to the teacher in our system using correction probability 1.0.

We assume that all of the learner's tokens initially have correctness bit 0 and analyze the expected number of interactions until all of the learner's tokens have correctness bit 1. In the case of the non-correcting teacher, the process reduces immediately to the coupon collection process: the choices of the learner are irrelevant, and each learner's token has its correctness bit changed from 0 to 1 the first time the teacher randomly draws its own token with the same number. Thus in this case, the expected number of interactions until the learner's tokens are all corrected is $n \ln n + O(n)$.

In the case of the correcting teacher, for each value of c from 0 to n, we define the random variable Y_c to be the number of the interaction in which the number of correct learner tokens first reaches c. Clearly, $Y_0 = 0$ and $Y_1 = 1$. We are interested in $E(Y_n)$. We define $X_c = Y_{c+1} - Y_c$ for $c = 0, 1, \ldots, n-1$.

If the learner currently has c correct tokens, then in one interaction it can come to have $c+1$ correct tokens in two ways: by choosing one of its $(n-c)$ incorrect tokens and being corrected by the teacher, or by choosing one of its c correct tokens and then having the teacher randomly choose a token corresponding to one of the learner's $(n-c)$ incorrect tokens, which is then corrected. Thus, the probability p_c of moving from c correct tokens to $c+1$ tokens in one interaction is

$$p_c = \frac{(n-c)}{n} + \frac{c}{n}\frac{(n-c)}{n} = \frac{n^2 - c^2}{n^2}.$$

Then we may bound $E(Y_n)$ as follows.

$$\begin{aligned}
E(Y_n) &= \sum_{c=0}^{n-1} E(X_c) \\
&= \sum_{c=0}^{n-1} \frac{1}{p_c} \\
&= n^2 \sum_{c=0}^{n-1} \frac{1}{n^2 - c^2} \\
&\leq n^2 \left(\int_0^{n-1} \frac{dx}{n^2 - x^2} \right) + \frac{n^2}{2n-1} \\
&< \frac{1}{2} n \ln n + O(n).
\end{aligned}$$

Asymptotically, the expected number of interactions until no uncorrected tokens remain for the learner interacting with the correcting teacher is $1/2$ of the value for the learner interacting with the non-correcting teacher. Despite the simplicity of this model, these results are roughly consistent with the rates of improvement seen in our empirical data for learners interacting with teachers correcting with probability 1.0, none of which exceeded about 40%.

This simplified model could be generalized to allow intermediate probabilities of correction by the teacher, and perhaps more complicated effects such as differential attention paid by the learner to perceived corrections. This model highlights the fact that neither teacher nor learner takes into account the history of their interactions, relying on the fact that random choices cover the whole domain with reasonable efficiency.

7. Discussion and future work

We have presented a computational model that takes into account semantics for language learning. We have shown that a system can learn to comprehend and generate utterances from pairs consisting on a situation and an utterance denoting one of the objects in that situation. Unlike previous approaches, there is no explicit semantic annotation of the utterances, the learner constructs the meanings of the utterances from scratch. The learner collects information about the co-occurrence of words and predicates, and uses the semantic situation to help it translate the teacher's utterances to meanings. The teacher's meanings are generalized via the semantic categories to general forms that are the basis of the learner's own utterances. The learner refines its ability to use the general forms to produce grammatical utterances by collecting data from alignments of the teacher's utterances with the general forms. This data is used to construct decision trees to choose sequences of words for parts of an intended meaning. Thus, the learner's grammar is represented by its initial semantic categories and the information it has collected about co-occurrence of words and predicates, about what general forms represent meanings, and about how to choose sequences of words for parts of an intended meaning.

Decomposing the learner's task into learning a grammar of meanings (in our system, the general forms) and a set of rules for expressing meanings in words (in our system, the decision trees) seems to simplify the learning task. For the tasks we consider, simply accumulating a list of general forms (with a mechanism to "age out" incorrect ones) and translating the atoms and gaps of the meaning into consecutive sequences of words provides a feasible solution. However, it is likely that more complex language learning tasks require a more complex model of the relationship between meanings and utterances.

Our model of language is very simplified, and there are many issues it does not deal with properly, including multi-word phrases bearing meaning, morphological relations between words, phonological rules for word choice, words with more than one meaning and meanings that can be expressed in more than one way, languages with free word-orders and meaning components expressed by non-contiguous sequences of words. Questions of generality, scalability and noise tolerance should also be addressed. Other desirable directions to explore include more sophisticated use of co-occurrence information, more

powerful methods of learning the grammars of meanings, feedback to allow the learning of productions to improve comprehension, better methods of alignment between utterances and meanings, and methods to allow the learner's semantic categories to evolve in response to language learning.

Our model and results have also allowed us to demonstrate that a relatively simple model of a teacher can offer meaning-preserving corrections to the learner, and that for certain learning tasks such corrections can significantly reduce the number of interactions for the learner to reach a high level of performance. Moreover, this improvement does not depend on the learner's ability to detect corrections: the effect depends on the change in the distribution of teacher utterances in the correcting versus non-correcting conditions. This result suggests re-visiting discussions in linguistics that assume that the learner must detect a teacher correction as a correction in order for it to have an influence on the learning process.

In fact, in our model the learner can detect meaning-preserving corrections of the teacher. A future direction of research is to see if a learner can use the ability to detect corrections to accelerate the learning process further. Interestingly, preliminary experiments in assigning greater weight to detected corrections in the decision tree construction process did not seem to have a significant positive effect on the attainment of high levels of performance.

Our model of a teacher is very simple and purely "reactive." That is, the teacher does not keep any history for the learner that it is interacting with (other than accumulating summary statistics), and reacts to learner utterances using the same procedures each time. The only dimension along which teachers differ is their propensity to correct the learner (that is, the probability of correction). An interesting future direction would be to explore the effects of making the teacher "more responsive" to the learner, possibly taking into account the history of interactions with this learner. One possibility would be to try a simple model of "motherese" in which the distribution on the teacher's utterances might be weighted toward utterances slightly more complex than the current average for learner utterances. Another possibility would be to have the teacher keep track of learner errors, and weight its distribution of utterances to provide more models of correct utterances in regions of learner error.

It would be very interesting to extend our results to more practical domains such as data mining and information retrieval. In the framework of data mining, semantic information might allow us to tackle the problem of scalability by reducing the amount of data the algorithms must consider. In information retrieval, specifically in the context of web search, a search guided by semantic information might allow users to get results more relevant to their queries when the desired information is hard to find purely syntactically. Semantics might also facilitate communication between the user and the search engine; the user could ask for information in a more natural way (not just using keywords), and the search engine could understand the query, even if it is not completely grammatically correct. We believe that incorporating semantics is crucial for future search engines.

Acknowledgements

The work of the first author was supported by the National Science Foundation under Grant CCF-0916389. The work of the second author was supported by a Marie Curie International Fellowship within the 6th European Community Framework Programme and was performed in part during an appointment as a Postdoctoral Scholar in the Computer Science Department of Yale University. The authors would like to thank Ahmet Aktay, Suna Bensch, Oya Bermek, Xuejin Chen, Ronny Dakdouk, Robert Frank, Kevin Gold, Johanna Högberg, Melis Inan, Gaja Jarosz, Edo Liberty, José Oncina, Lev Reyzin, Brian Scassellati, Nikhil Srivastava, Antonis Stampoulis, György Vaszil, and Yinghua Wu for their help with aspects of this work. We also thank the anonymous reviewers of AI journal for their useful comments and suggestions.

Appendix A. Excerpted interactions

Below we comment on a few of the first fifty interactions drawn from one run of a learner learning the 68-form Spanish sublanguage from a teacher with correction probability 1.0.

```
> (test-x-sp 1 50 1.0)
(
((interaction-number 0)
(situation ((sm1 t1) (pu1 t1) (el1 t1) (le2 t1 t2)
            (bi1 t2) (pu1 t2) (tr1 t2)))
(learner-intended-meaning ())
(learner-utterance ())
(teacher-classification no-utterance)
(intended-correction? #f)
(teacher-utterance (la elipse purpura a la izquierda
                    del triangulo))
(learner-detected-correction? #f)
(new-general-form ())
(learner-understanding ())
)
```

In this first interaction, the situation is a small purple ellipse to the left of a big purple triangle. Because the learner has no knowledge of the language, it does not attempt an utterance. The teacher's classification is "no utterance" and it does not attempt a correction, but rather produces a randomly-drawn denoting utterance for the situation, which in English would be *the purple ellipse to the left of the triangle*. The learner is not able to understand the utterance, does not acquire a new general form, and does not perceive any correction by the teacher.

However, information about co-occurrence of words and predicates is incorporated into the learner's co-occurrence graph. Things continue in this way until the fifth interaction, when the learner's understanding of the teacher's utterance (which in English would be *the green triangle above the blue hexagon*) is the meaning `((tr1 x1)(he1 x2))`. This meaning is partially correct, but omits a variable atom giving the relation between the triangle and the hexagon. This meaning is generalized to the general form `((shape1 x1)(shape2 x2))`, which is also incorrect. However the learner subsequently uses this general form to produce denoting utterances of its own.

```
((interaction-number 4)
(situation ((sm1 t1) (gr1 t1) (tr1 t1) (ab2 t1 t2)
            (sm1 t2) (bl1 t2) (he1 t2)))
(learner-intended-meaning ())
(learner-utterance ())
(teacher-classification no-utterance)
(intended-correction? #f)
(teacher-utterance (el triangulo verde encima del hexagono azul))
(learner-detected-correction? #f)
(new-general-form ((shape1 x1) (shape1 x2)))
(learner-understanding ((tr1 x1) (he1 x2)))
)
```

The learner uses its newly-acquired (but incorrect) general form in the following interaction. The learner intends the meaning `((el1 x1) (ci1 x2))`, that is, *ellipse circle*, but the word-choice part of its grammar yields a kind of word salad (though containing the word for ellipse). The teacher finds the word salad "uninterpretable" and produces a randomly-chosen denoting utterance for the situation.

```
((interaction-number 6)
(situation ((bi1 t1) (gr1 t1) (ci1 t1) (ab2 t1 t2)
            (sm1 t2) (gr1 t2) (el1 t2)))
(learner-intended-meaning ((el1 x1) (ci1 x2)))
(learner-utterance (el elipse verde encima del mediana azul))
(teacher-classification uninterpretable)
(intended-correction? #f)
(teacher-utterance (la elipse verde debajo del circulo grande))
(learner-detected-correction? #f)
(new-general-form ())
(learner-understanding ())
)
```

The next interaction is similar, though in this case the learner produces correct words for both the star and the triangle.

```
((interaction-number 7)
(situation ((me1 t1) (re1 t1) (tr1 t1) (le2 t1 t2)
            (sm1 t2) (re1 t2) (st1 t2)))
(learner-intended-meaning ((st1 x1) (tr1 x2)))
(learner-utterance (el estrella verde encima del triangulo azul))
(teacher-classification uninterpretable)
(intended-correction? #f)
(teacher-utterance (la estrella pequena y roja a la derecha
                    del triangulo mediano))
(learner-detected-correction? #f)
(new-general-form ())
(learner-understanding ())
)
```

Interactions continue, and the learner acquires another incorrect general form and attempts to use both of its incorrect general forms. As its word-choice rules are refined, the learner produces utterances closer to its meaning, and in the following interaction, the teacher is first able to understand the learner's meaning enough to offer a correction, which is not perceived by the learner.

```
((interaction-number 15)
(situation ((bi1 t1) (or1 t1) (el1 t1) (ab2 t1 t2)
            (bi1 t2) (ye1 t2) (he1 t2)))
(learner-intended-meaning ((el1 x1) (he1 x2)))
(learner-utterance (el elipse hexagono))
(teacher-classification error-in-meaning)
(intended-correction? #t)
(teacher-utterance (la elipse encima del hexagono))
(learner-detected-correction? #f)
(new-general-form ())
(learner-understanding ())
)
```

In the following interaction, the learner again uses its first incorrect general form, and the teacher attempts to correct the utterance. Despite not perceiving the correction, the learner acquires its first correct general form, `((shape1 x1)(le2 x2 x1)(shape1 x2))`, from the teacher's utterance. The learner continues to try to use all three of its acquired general forms, but the two incorrect ones will gradually lose probability because they will not continue to be matched by the teacher's utterances once the learner's understanding of the teacher's utterances improves sufficiently.

```
((interaction-number 18)
(situation ((sm1 t1) (bl1 t1) (tr1 t1) (le2 t1 t2)
            (bi1 t2) (bl1 t2) (st1 t2)))
(learner-intended-meaning ((st1 x1) (tr1 x2)))
(learner-utterance (el estrella triangulo))
(teacher-classification error-in-meaning)
(intended-correction? #t)
(teacher-utterance (la estrella a la derecha del triangulo))
(learner-detected-correction? #f)
(new-general-form ((shape1 x1) (le2 x2 x1) (shape1 x2)))
(learner-understanding ((st1 x1) (le2 x2 x1) (tr1 x2)))
)
```

Here is the first interaction in which the learner perceives a correction by the teacher, because the meaning intended by the learner and the learner's understanding of the teacher's utterance are the same, but the actual utterances are different (the learner's utterance has an incorrect choice for the article for the word triangle). Note that although the learner perceives a correction, the teacher's utterance is processed the same way as it would be if the learner did not perceive a correction.

```
((interaction-number 24)
(situation ((me1 t1) (gr1 t1) (tr1 t1) (le2 t1 t2)
            (me1 t2) (gr1 t2) (tr1 t2)))
(learner-intended-meaning ((tr1 x1) (le2 x2 x1) (tr1 x2)))
(learner-utterance (la triangulo a la derecha del triangulo))
(teacher-classification error-in-form)
(intended-correction? #t)
(teacher-utterance (el triangulo a la derecha del triangulo))
(learner-detected-correction? #t)
(new-general-form ())
(learner-understanding ((tr1 x1) (le2 x2 x1) (tr1 x2)))
)
```

In the following interaction, the learner for the first time produces an utterance classified as correct by the teacher. The learner continues to produce unintelligible utterances, utterances with errors in form and meaning, and correct utterances, while refining its co-occurrence graph (allowing it to understand the teacher's utterances more accurately), acquiring new general forms, and improving its rules for word choice.

```
((interaction-number 25)
(situation ((bi1 t1) (gr1 t1) (tr1 t1) (ab2 t1 t2)
            (bi1 t2) (pu1 t2) (el1 t2)))
(learner-intended-meaning ((el1 x1) (ab2 x2 x1) (tr1 x2)))
(learner-utterance (la elipse debajo del triangulo))
(teacher-classification correct)
(intended-correction? #f)
(teacher-utterance (el triangulo grande encima de la elipse purpura))
(learner-detected-correction? #f)
(new-general-form ())
(learner-understanding ())
)
```

The final interaction in the sequence of 50 is followed by the teacher's summary of its classifications and corrections. In this particular run, the teacher's classifications were: 12 correct utterances, 10 utterances with an error in meaning, 9 utterances with an error in form, 10 uninterpretable utterances, and 9 interactions in which the learner produced no utterance. Because the teacher's correction probability was 1.0, it offered a correction for every error in meaning or form, for a total of 19 corrections.

```
((interaction-number 49)
(situation ((bi1 t1) (or1 t1) (tr1 t1) (ab2 t1 t2)
            (bi1 t2) (or1 t2) (he1 t2)))
(learner-intended-meaning ((tr1 x1) (ab2 x1 x2) (he1 x2) (or1 x2)))
(learner-utterance (el triangulo encima de la hexagono naranja))
(teacher-classification error-in-form)
(intended-correction? #t)
(teacher-utterance (el triangulo encima del hexagono naranja))
(learner-detected-correction? #t)
(new-general-form ())
(learner-understanding ((tr1 x1) (ab2 x1 x2) (he1 x2) (or1 x2)))
)
```

Appendix B. Notes on the sublanguages

The 68-form languages include utterances referring to one object or to two objects with a relation between them. The objects are one of six shapes (square, circle, triangle, star, hexagon, or ellipse), have one of three sizes (big, medium, or small) and have one of six colors (red, orange, yellow, green, blue, or purple). The relation between them expresses that one object is to the left of, to the right of, above, or below the other object. For each language, a grammar (in the form of a meaning transducer) was constructed so that exactly one utterance produces each of the 113,064 possible expressible meanings, each of which is an instance of one of 68 general forms.

Examples of teacher utterances for each language follow. Individual words are represented by lower-case English letters, with the exceptions noted.

English:

the purple square above the big yellow hexagon
the big star to the right of the big triangle
the purple triangle to the right of the square
the medium blue hexagon
the red circle

German:

der orangne kreis
der mittlere stern uber dem grunen kreis
das kleine grune quadrat links vom dem orangnen stern
das kleine blaue hexagon unter dem quadrat
das grosse orangne dreieck links von der ellipse

Greek:

to exagono
to mikro kokkino tetragono sta aristera tou mesaiou portokali kyklou
i mikri ellipsi
i ellipsi sta deksia tou kyklou
to portokali trigono pano apo to mikro astro

Phonological rules for articles were not included.

Hebrew:

hakochav hagadol hakachol
haelipsa mismol lameshushe
hameshulash hakatan me!al laelipsa haktana hatsehuva
haigul mitachat laelipsa ha!aduma
hameshulash miyamin lameshushe hagadol hasagol

Hungarian:

a/az piros negyzet
a/az kicsi zold negyzet a/az kozepes csillag folott
a/az piros kor balra a/az negyzettol
a/az csillag a/az kicsi sarga haromszog alatt
a/az kicsi zold negyzet jobbra a/az lila kortol

The phonological rule to select the article *a* or *az* is omitted, so the article is *a/az* throughout. The grammar we used has the words for left and right separating the two objects, but the words for above and below following both objects.

Mandarin:

da de lu se xing1
liu bian xing zuo bian de huang se zheng fang xing
huang se zheng fang xing
xiao de zheng fang xing xia mian de zhong jian da-xiao de lu se san jiao xing
zhong jian da-xiao de lu se yuan zuo bian de xiao de hong se tuo-yuan

Because of limitations of our system in dealing with words with more than one meaning and multi-word phrases, we distinguished the senses of *xing* and *xing1* and hyphenated *da-xiao* (medium) and *tuo-yuan* (ellipse).

Russian:

bolshoy fioletoviy shestiugolnik
triugolnik pod sinim ellipsom
shestiugolnik c leva ot srednevo ellipsa
kvadrat c prava ot bolshovo shestiugolnika
bolshoy krasniy shestiugolnik c leva ot bolshoy siney zvezdi

There are three cases of each noun and adjective in the grammar we used.

Spanish:

el triangulo
el hexagono azul debajo del circulo
la estrella encima de la elipse roja
el cuadrado pequeno y amarillo a la izquierda del cuadrado mediano y naranja
el hexagono grande y azul a la derecha de la elipse grande

In the grammar we use, the order of adjectives is fixed as size followed by color.

Swedish:

cirkeln
den lilla bla fyrkanten
den normala triangeln under den normala oranga stjarnan
den lila ellipsen till hoger om den lilla fyrkanten
den normala roda ellipsen till vanster om hexagonen

Turkish:

kare
mavi karenin ustundeki sari elips
orta yesil cemberin solundaki yildiz
buyuc yildizin altindaki altigen
orta sari cemberin altindaki mavi kare

Appendix C. Pseudocode

Algorithm 1 Computing the graphs G, H, and H^r.

Require: threshold θ, sequence of n pairs (s_i, d_i) of a situation s_i and a denoting utterance d_i
Ensure: graphs G, H, and H^r

{*Construct the* ***co-occurrence graph*** G}
Let V be the set of predicates and words occurring in any (s_i, d_i)
for each vertex $u \in V$ **do**
  Let $c(u) =$ count of (s_i, d_i) pairs in which u occurred
end for
for each pair of distinct vertices (u, v) **do**
  Let $c(u, v) =$ count of (s_i, d_i) pairs in which u and v both occurred
end for

{*Construct the* ***implication graph*** H *from* G}
for each vertex $u \in V$ **do**
  Let $p(u) = c(u)/n$
end for
for each pair of distinct vertices (u, v) **do**
  Let $p(u, v) = c(u, v)/c(u)$
  if $p(u, v) \geq \theta$ **then**
    Include a directed edge between (u, v)
  end if
end for

{*Construct the* ***transitive reduction graph*** H^r *from* H}
Remove edges (u, v) where u is a predicate and v is a word
Contract each strongly connected component to a vertex
Transitively reduce the contracted graph
Expand the strongly connected components to get H^r

Algorithm 2 Learner's production.

Require: situation s, general forms F, decision trees T
Ensure: an utterance for s

Let M be all meanings obtained from general forms in F instantiated with predicates from s
Let D be $\{m \in M \mid m$ is denoting in $s\}$
if $D = \emptyset$ **then**
  Produce the empty utterance
else
  Let U be the translations of all $d \in D$ into utterances (using decision trees T)
  Produce a random utterance u from U (with probability proportional to the square of the interaction number stored with the general form of u)
end if

Algorithm 3 Teacher's comprehension and classification.

Require: transducer A, situation s and learner utterance u
Ensure: classification of utterance u and (possibly) a correction of it
  Let D be the denoting utterances for s (using A)
  Let M be the meanings of utterances in D
  Let M' be the sequences of predicates of elements of M
  if u is in D **then**
    Classify u as "correct", with no correction
  else if u is empty **then**
    Classify u as "no utterance", with no correction
  else
    Let m' be the sequence of predicates obtained from u (using A)
    if $m' \in M'$ **then**
      Classify u as "error in form" with correction equal to some $u \in D$ with sequence of predicates m'.
    else if m' is "close enough" to some $m'' \in M'$ **then**
      Classify u as "error in meaning" with correction equal to some $u \in D$ with sequence of predicates m''.
    end if
  else
    Classify u as "uninterpretable" with no correction
  end if

Algorithm 4 Teacher's (possibly correcting) production.

Require: transducer A, situation s, learner utterance u, correction probability p
Ensure: a denoting utterance for s

Let c be the result of flipping a coin with probability p
Let D be the denoting utterances for s (using A)
if c is "tails" **then**
  Return a randomly chosen element of D
else if Teacher's classification of u is "error in form" or "error in meaning" **then**
  Return Teacher's correction of u
else
  Return a randomly chosen element of D
end if

Algorithm 5 Learner's comprehension, detection of correction, and updates.

Require: threshold θ, situation s, denoting utterance d for s, learner's intended meaning m_L for s, general forms F and decision trees T
Ensure: learner understanding m of d, relevant updates

for each word $w_i \in d$ **do**
  Let P_i be all predicates p_j with an edge (w_i, p_j) in H^r
  for each predicate $p_j \in P_i$ **do**
    if $p(p_j) \geq \theta$ **or** p_j is not of minimum arity in P_i **then**
      Remove p_j from P_i
    end if
  end for
end for
for maximal sets W of words of d in same strong component of H^r **do**
  Set P_i to $\emptyset$ for all w_i in W except the word rightmost in d
end for

Let Q be all sequences of predicates obtained by taking in order one predicate from each nonempty P_i
Let R be all denoting meanings for s derived from Q
Update graphs G, H, H^r with the new pair (s, d)

{*If exactly one denoting meaning is found, use it*}
if $R = \{m\}$ **then**
  Generalize meaning m to general form f
  Include f in F, recording the current interaction number
  Find the alignment of d and m
  Use the alignment to update decision trees T
  Learner detects a correction if m equals m_L
  Learner returns learner understanding m
else
  Learner returns empty learner understanding
end if

References

[1] A.V. Aho, M.R. Garey, J.D. Ullman, The transitive reduction of a directed graph, SIAM J. Comput. 1 (2) (1972) 131–137.
[2] D. Angluin, Inference of reversible languages, J. Assoc. Comput. Mach. 29 (3) (1982) 741–765.
[3] D. Angluin, L. Becerra-Bonache, Learning Meaning Before Syntax, Tech. rep. YALE/DCS/TR1407, Computer Science Department, Yale University, 2008.
[4] D. Angluin, L. Becerra-Bonache, Learning meaning before syntax, in: ICGI 2008 – 9th International Colloquium on Grammatical Inference, 2008, pp. 1–14.
[5] D. Angluin, L. Becerra-Bonache, A Model of Semantics and Corrections in Language Learning, Tech. rep. YALE/DCS/TR1425, Computer Science Department, Yale University, 2010.
[6] D. Angluin, L. Becerra-Bonache, Effects of meaning-preserving corrections on language learning, in: CoNLL 2011 – 15th Conference on Computational Natural Language Learning, 2011, pp. 97–105.
[7] D. Angluin, L. Becerra-Bonache, An overview of how semantics and corrections can help language learning, in: WI-IAT 2011 – Web Intelligence and Intelligent Agent Technology – Workshops, IEEE Computer Society, 2011, pp. 147–150.
[8] D. Bailey, A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs, Ph.D. thesis, U.C. Berkeley, 1997.
[9] L. Becerra-Bonache, On the Learnability of Mildly Context-Sensitive Languages Using Positive Data and Correction Queries, Ph.D. thesis, Rovira i Virgili University, 2006.
[10] L. Becerra-Bonache, C. de la Higuera, J. Janodet, F. Tantini, Learning balls of strings from edit corrections, J. Mach. Learn. Res. 9 (2008) 1841–1870.
[11] R. Brown, C. Hanlon, Derivational complexity and the order of acquisition in child speech, in: J. Hayes (Ed.), Cognition and the Development of Language, Wiley, New York, 1970, pp. 11–54.
[12] D.L. Chen, J. Kim, R.J. Mooney, Training a multilingual sportscaster: using perceptual context to learn language, J. Artif. Intell. Res. 37 (2010) 397–435.
[13] D.L. Chen, R.J. Mooney, Learning to sportscast: a test of grounded language acquisition, in: ICML 2008 – 25th International Conference on Machine Learning, 2008, pp. 128–135.
[14] M. Chouinard, E. Clark, Adult reformulations of child errors as negative evidence, J. Child Lang. 30 (2003) 637–669.
[15] A. Clark, S. Lappin, Linguistic Nativism and the Poverty of the Stimulus, Wiley–Blackwell, Malden, MA, 2011.
[16] A. Clark, S. Lappin, Computational learning theory and language acquisition, in: R. Kempson, N. Asher, T. Fernando (Eds.), Philosophy of Linguistics, Elsevier, 2012, pp. 445–475.
[17] E. Clark, The principle of contrast: a constraint on language acquisition, in: B. MacWhinney (Ed.), Mechanisms of Language Acquisition, Erlbaum, Hillsdale, NJ, 1987, pp. 1–33.
[18] E. Clark, The Lexicon in Acquisition, Cambridge University Press, Cambridge, 1993.

[19] C. de Marcken, Unsupervised Language Acquisition, Ph.D. thesis, MIT, 1996.
[20] M.J. Demetras, K.N. Post, C.E. Snow, Feedback to first language learners: the role of repetitions and clarification questions, J. Child Lang. 13 (1986) 275–292.
[21] S. Ervin-Tripp, Some strategies for the first two years, in: Cognitive Development and the Acquisition of Language, Academic Press, 1973, pp. 261–286.
[22] A. Fazly, A. Alishahi, S. Stevenson, A probabilistic computational model of cross-situational word learning, Cogn. Sci. 34 (6) (2010) 1017–1063.
[23] M.C. Frank, N.D. Goodman, J.B. Tenenbaum, A Bayesian framework for cross-situational word-learning, in: NIPS 2007 – 21st Annual Conference on Neural Information Processing Systems, 2007, pp. 457–464.
[24] E. Gold, Language identification in the limit, Inf. Control 10 (1967) 447–474.
[25] K. Gold, Using Sentence Context and Implicit Contrast to Learn Sensor-Grounded Meanings for Relational and Deictic Words: The TWIG System, Ph.D. thesis, Yale University, 2008.
[26] K. Gold, M. Doniec, C. Crick, B. Scassellati, Robotic vocabulary building using extension inference and implicit contrast, Artif. Intell. 173 (1) (2004) 145–166.
[27] K. Gold, B. Scassellati, A robot that uses existing vocabulary to infer non-visual word meanings from observation, in: AAAI 2007 – 22nd Conference on Artificial Intelligence, 2007, pp. 883–888.
[28] J.A.C. Hill, A Computational Model of Language Acquisition in the Two-Year-Old, Indiana University Linguistics Club, Indiana, 1983.
[29] K. Hirsh-Pasek, R.A. Treiman, M. Schneiderman, Brown and Hanlon revisited: mothers' sensitivity to ungrammatical forms, J. Child Lang. 2 (1984) 81–88.
[30] J.A. Feldman, G. Lakoff, A.S.S. Weber, Miniature language acquisition: a touchstone for cognitive science, in: Proceedings of the 12th Annual Conference of the Cognitive Science Society, MIT, Cambridge, MA, 1994, pp. 686–693.
[31] K. Jack, A Computational Model of Staged Language Acquisition, Ph.D. thesis, University of Dundee, 2006.
[32] M. Kanazawa, Learnable Classes of Categorial Grammars, Cambridge University Press, New York, NY, 1998.
[33] R.J. Kate, R.J. Mooney, Learning language semantics from ambiguous supervision, in: AAAI 2007 – 23rd Conference on Artificial Intelligence, 2007, pp. 895–900.
[34] J. Kim, R.J. Mooney, Generative alignment and semantic parsing for learning from ambiguous supervision, in: COLING 2010 – 23rd International Conference on Computational Linguistics, 2010, pp. 543–551.
[35] J. Kim, R.J. Mooney, Unsupervised PCFG induction for grounded language learning with highly ambiguous supervision, in: EMNLP-CoNLL 2012 – Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 433–444.
[36] E.B. Kinber, On learning regular expressions and patterns via membership and correction queries, in: ICGI 2008 – 9th International Colloquium on Grammatical Inference, 2008, pp. 125–138.
[37] G. Marcus, Negative evidence in language acquisition, Cognition 46 (1993) 53–95.
[38] J. Morgan, L. Travis, Limits on negative information in language input, J. Child Lang. 16 (1989) 531–552.
[39] M. Redington, N. Chater, S. Finch, Distributional information: a powerful cue for acquiring syntactic categories, Cogn. Sci. 22 (1998) 425–469.
[40] T. Regier, The Human Semantic Potential: Spatial Language and Constrained Connectionism, MIT Press, Cambridge, MA, 1996.
[41] T. Regier, The emergence of words: attentional learning in form and meaning, Cogn. Sci. 29 (2005) 819–865.
[42] D. Roy, A. Pentland, Learning words from sights and sounds: a computational model, Cogn. Sci. 26 (2002) 113–146.
[43] Y. Sakakibara, Efficient learning of context-free grammars from positive structural examples, Inf. Process. Lett. 97 (1992) 23–60.
[44] A. Schaerlaekens, The Two-Word Sentence in Child Language Development: A Study Based on Evidence Provided by Dutch-Speaking Triplets, The Hague, Mouton, 1973.
[45] J. Siskind, Lexical acquisition in the presence of noise and homonymy, in: AAAI 1994 – 12th National Conference on Artificial Intelligence, MIT Press, 1994, pp. 760–766.
[46] J. Siskind, A computational study of cross-situational techniques for learning word-to-meaning mappings, Cognition 61 (1996) 39–61.
[47] I. Tellier, Meaning helps learning syntax, in: ICGI 1998 – 4th International Colloquium on Grammatical Inference, 1998, pp. 25–36.
[48] C. Tirnăucă, T. Knuutila, Polynomial time algorithms for learning k-reversible languages and pattern languages with correction queries, in: ALT 2007 – 18th International Conference on Algorithmic Learning Theory, 2007, pp. 272–284.
[49] E. Veneziano, Displacement and informativeness in child-directed talk, First Lang. 21 (63) (2001) 323.
[50] A. Villavicencio, T. Poibeau, A. Korkhonen, A. Alishahi, Cognitive Aspects of Computational Language Acquisition, Springer, 2013.
[51] Y.W. Wong, R.J. Mooney, Generation by inverting a semantic parser that uses statistical machine translation, in: HLT-NAACL 2007 – Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, 2007, pp. 172–179.
[52] Y.W. Wong, R.J. Mooney, Learning synchronous grammars for semantic parsing with lambda calculus, in: ACL 2007 – 45th Annual Meeting of the Association for Computational Linguistics, 2007, pp. 960–967.
[53] C. Yu, The emergence of links between lexical acquisition and object categorization: a computational study, Connect. Sci. 17 (3–4) (2005) 381–397.
[54] C. Yu, D. Ballard, A multimodal learning interface for grounding spoken language in sensory perceptions, ACM Trans. Appl. Percept. 1 (1) (2004) 57–80.