Triangle: Language, Literature, Computation, n. 1, 2010
Publicacions Universitat Rovira i Virgili · ISSN: 2013-939X
https://revistes.urv.cat/index.php/triangle

An Introduction to Grammatical Inference for Linguists

Leonor Becerra-Bonache*

Research Group on Mathematical Linguistics
Rovira i Virgili University, Spain.

1 Motivation

This paper is meant to be an introductory guide to Grammatical Inference (GI), i.e., the study of machine learning of formal languages. It is designed for non-specialists in Computer Science, but with a special interest in language learning. It covers basic concepts and models developed in the framework of GI, and tries to point out the relevance of these studies for natural language acquisition.

How do children acquire their native language? This question has attracted the attention of researchers from different areas, including linguistics, cognitive science and computer science. Traditionally, this question has been addressed by linguists and psychologists. Their approach has specially been focused on making experiments with children that are acquiring their native language, with the ultimate goal of describing the process of natural language acquisition. There are basically two different kinds of studies: longitudinal studies (which focus on one child and collect data regularly to create extensive databases that can be found at CHILDES:

* Supported by a Marie Curie International Fellowship within the 6^{th} European Community Framework Programme.

http://childes.psy.cmu.edu/); transversal studies (experiments are made with a group of children of different ages. Researchers try to test a hypothesis and design specific tasks that have to be performed by the children). Although important results have been obtained from all these studies, there are still many open questions about how children acquire their native language. This is why researchers have tried to approach the problem from a more interdisciplinary point of view, including such different scientific disciplines as Computer Science.

Within the field of Computer Science, *Artificial Intelligence* aims to study and design intelligent machines. This field was founded in the middle of the 50s. It has two different purposes:

> *One is to use the power of computers to augment human thinking, just as we use motors to augment human or horse power. Robotics and expert systems are major branches of that. The other is to use a computer's artificial intelligence to understand how humans think. In a humanoid way. If you test your programs not merely by what they can accomplish, but how they accomplish it, then you're really doing cognitive science; you're using Artificial Intelligence to understand the human mind.* [34].

The founders of Artificial Intelligence were very optimistic about the future of this new field. For example, in 1965, H. Simon predicted that *"... machines will be capable, within twenty years, of doing any work a man can do"* [33]. Although important advances have been made in the last 45 years, this prediction has not come true yet. We have machines that can do "some of the things" that a man can do; for example, play soccer (http://www.robocup.org/), play some instruments (like Toyota's violin-playing robot), express feelings by moving their faces (like the MIT's robots: MDS and Kismet). Nevertheless, so far machines have been unable to *learn to speak*. The advantages of having a machine that can learn and speak a natural language would be innumerable. From a theoretical point of view, for example, we could better understand the process of natural language acquisition. From a practical point of view, to have a machine that is able to speak would definitely facilitate communication between humans and machines.

Within the field of Artificial Intelligence, *Machine Learning* aims to develop techniques that allow computers to learn. Machine Learning is concerned with the design and development of algorithms that allow computers to use data to change their behavior (an algorithm is a finite sequence

of instructions specifying how to solve a particular problem). Some of the Machine Learning applications are: natural language processing, search engines, medical diagnosis, detection of credit card fraud, classification of DNA sequences, speech and handwriting recognition, etc.

Grammatical Inference is a specialized subfield of Machine Learning that deals with the learning of formal languages from a set of examples. The basic framework can be regarded as a game played between two players: a teacher and a learner. The teacher provides data to the learner, and the learner (or learning algorithm), from these data, must identify the underlying language. For example, imagine that the target language (i.e., the language to be learnt) is ab^+ (i.e., a language that contains strings starting with one a, followed by at least one b). The teacher could provide the learner with strings that belong to the language (i.e., positive data), such as ab, abb, $abbb$... The learner uses this information to infer that the target language is ab^+.

As we can see, this process has some similarities with the process of natural language acquisition; instead of a teacher we could have an adult, and instead of a learner, a child. Therefore, GI provides a good theoretical framework to study the problem of natural language acquisition. In fact, the initial theoretical foundations of GI were given by E.M. Gold, who was primarily motivated by the problem of children's language acquisition.

It is worth noting that the theory of formal languages was born in the 50's as a tool to describe natural language syntax. Hence, formal languages are an important tool to study natural languages. Moreover, formal results are also of great interest, because as A. Clark [16] pointed out:

> *Positive results can help us to understand how humans might learn languages by outlining the class of algorithms that might be used by humans, considered as computational systems at a suitable abstract level. Conversely, negative results might be helpful if they could demonstrate that no algorithms of a certain class could perform the task Ð in this case we could know that the human child learns his language in some other way* [16, p. 26].

Therefore, by applying Grammatical Inference to the study of natural language acquisition, we could provide a formal model that explains how children acquire their native language. The study and development of a formal model of language learning is of great relevance, not only to better understand the process of natural language acquisition, but also for the

practical applications that such a model could have (for example, communication between humans and machines could be improved).

The remainder of the paper is organized as follows. We give some basic definitions in Section 2. In Section 3, we review some of the most important formal models investigated in GI, and we analyze them from a linguistic point of view. In section 4, we try to answer the following two questions: what classes of formal languages are interesting from a linguistic point of view? and what source of data should we provide our learning algorithm? In Section 5 we present some new lines of research in GI, motivated by studies of children's language acquisition. Concluding remarks are presented in Section 6.

2 Basic definitions

Formal languages are defined with respect to a given *alphabet*. The alphabet is a finite set of symbols, denoted Σ (e.g., $\Sigma = \{a, b\}$). A finite sequence of symbols chosen from some alphabet is called a *string* (e.g., $a, b, aa, ab, ba, bb, aaa$...). A *language* is a set of strings; among all the possible strings, some of them belong to the language and others do not (e.g., $ab, abb, abbb$ belong to the language ab^+, but $a, ba, abba$ do not). A *grammar* is a finite mechanism that generates the elements of the language.

The Chomsky grammars are particular cases of *rewriting systems*, where the operation used to process the strings is rewriting (the replacement of a "short" substring of the processed string by another short substring). According to the form of their rules, the Chomsky grammars are classified as follows (from less to more expressive power): regular (REG), context-free (CF), context-sensitive (CS), recursively enumerable (RE). We call this the *Chomsky hierarchy* (see Figure 1). It is worth noting that Chomsky defined these formal grammars/languages with the ultimate goal of modeling the syntax of natural language.

For example, the language ab^+ is a regular language generated by the following regular grammar: $S \rightarrow aB, B \rightarrow b, B \rightarrow bB$.

Automata are recognizer devices that are able to decide whether or not an input string belongs to a specified language. The five basic families of languages in the Chomsky Hierarchy are also characterized by recognizing automata. These automata are: the finite automaton, the one-turn pushdown automaton, the pushdown automaton, the linearly bounded automaton, and the Turing machine, respectively.

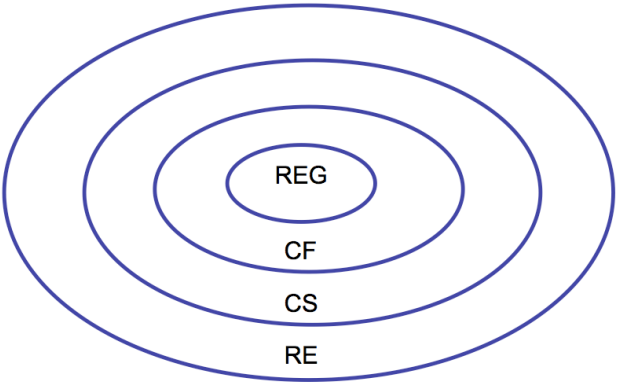

Fig. 1. The Chomsky Hierarchy

A *finite automaton* consists of a finite set of states, a finite alphabet of input symbols, and a set of transition rules. If the next state is always uniquely determined by the current state and the current input symbol, we say that the automaton is *deterministic*. Formally, a deterministic finite automata (DFA) is defined as a 5-tuple $(\Sigma, Q, \delta, q_0, F)$ where: Σ is the alphabet, Q is a finite set of states, T is the transition function ($T : Q \text{ x } \Sigma \rightarrow Q$, that is, from one state and reading a given symbol from the alphabet, we go to another state), q_0 is the initial state, and F the set of final states ($F \subseteq q$). A DFA takes a string as an input, and for each input symbol go to a state by following the transition function. When the last symbol is processed, depending on whether the DFA is in an accepting state or not, the string is accepted or rejected. A DFA characterizes the family of languages REG. See Figure 2 for an example of a DFA; initial state is marked with the symbol $>$ and the final (or accepted) state is marked with a double circle.

3 Formal models in Grammatical Inference

In this section we present two of the most important formal models developed within the field of GI. We also discuss some linguistics aspects of these models.

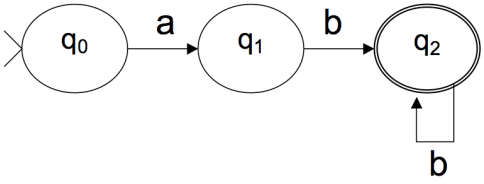

Fig. 2. Example of a Deterministic Finite Automata (DFA). This DFA recognizes the language ab^+.

3.1 Gold: Identification in the limit

In 1967, Gold [21] introduced the model of *identification in the limit*. His final goal was to explain the acquisition of natural languages.

> *The study of language identification described here derives its motivation from artificial intelligence. The results and the methods used also have implications in computational linguistics, in particular the construction of discovery procedures, and in psycholinguistics, in particular the study of child learning (...).*
> *I wish to construct a precise model for the intuitive notion "able to speak a language" in order to be able to investigate theoretically how it can be achieved artificially. Since we cannot explicitly write down the rules of English which we require one to know before we say he can "speak English", an artificial intelligence which is designed to speak English will have to learn its rules from implicit information. That is, its information will consist of examples of the use of English and/or of an informant who can state whether a given usage satisfies certain rules of English, but cannot state these rules explicitly.* [21, pp. 447–448].

Identification in the limit views learning as an infinite process. In this model, the learner passively receives more and more examples, and has to produce a hypothesis of the target language. If the learner receives new examples that are not consistent with his hypothesis, he has to change it. His hypothesis has to converge to a correct hypothesis. We say that the learner identifies the target language in the limit if, after a finite number of examples, he makes a correct guess and does not alter his guess thereafter.

It is worth noting that under this criterion, the learner cannot be certain of having correctly guessed the target language, since he may receive new

examples that are not consistent with his hypothesis. Gold justifies the study of identifiability in the limit in the following way:

> *My justification for studying identifiability in the limit is this: A person does not know when he is speaking a language correctly; there is always the possibility that he will find that his grammar contains an error. But we can guarantee that a child will eventually learn a natural language, even if it will not know when it is correct.* [21, p. 450].

Gold studied two different learning settings: i) Learning from *text*: the learner only receives positive data (strings that belong to the language); ii) Learning from *informant*: the learner receives positive and negative data (i.e., strings that belong to the language and strings that do not).

Gold proved that *superfinite* classes of languages (a class is superfinite if it contains all finite languages and at least one infinite language) cannot be identified in the limit from only positive data. This implies that none of the classes of languages defined by Chomsky to model natural language syntax is identifiable in the limit from only positive data. Therefore, the following question arises: How do children overcome this theoretical hurdle? Gold suggested the following hypothesis:

> *If one accepts identification in the limit as a model of learnability, then this conflict must lead to at least one of the following conclusions:*
>
> 1. *The class of possible natural languages is much smaller than one would expect from our present models of syntax. That is, even if English is context-sensitive, it is not true that any context-sensitive language can occur naturally. Equivalently, we may say that the child starts out with more information than that the language it will be presented is context-sensitive. In particular, the results on learnability from text imply the following: The class of possible natural languages, if it contains languages of infinite cardinality, cannot contain all languages of finite cardinality.*
> 2. *The child receives negative instances by being corrected in a way we do not recognize. If we can assume that the child receives both positive and negative instances, then it is being presented information by an "informant". The class of primitive recursive languages, which includes the class of context-sensitive languages, is identifiable in the limit from an informant. The child may receive the equivalent of negative instances for the purpose of grammar acquisition when it does not get the desired response to an utterance. It is difficult to interpret the actual training*

program of a child in terms of the naive model of a language assumed here.

3. *There is an a priori restriction on the class of texts which can occur, such as a restriction on the order of text presentation. The child may learn that a certain string is not acceptable by the fact that it never occurs in a certain context. This would constitute a negative instance.* [21, p. 453–454].

Studies along these lines have shown that the first path (the class of potential natural language is more restrictive than those defined by Chomsky) can be successful (see, [1,25,31]). In linguistics, it is also generally assumed that the first conclusion holds.

> *Now it seems evident to many linguists (notably, Chomsky [40,43]) that children are not genetically prepared to acquire any arbitrary language on the basis of the kind of casual linguistic exposure typically afforded the young. Instead, a relatively small class $\mathcal{H}$ of languages may be singled out as "humanly possible" on the basis of their amenability to acquisition by children, and it falls to the science of linguistics to propose a nontrivial description of $\mathcal{H}$* [23, p.29].

3.2 Angluin: Query Learning

D. Angluin introduced the query learning model in [2]. In this model, the learner is allowed to make queries to the teacher. The teacher (or oracle) knows the target language and answers the queries made by the learner correctly (he is perfect).

The learner (or learning algorithm) can only make queries from a given set. After asking a finite number of questions, the learner must return a hypothesis. The learner's hypothesis has to be the correct one (that is why this kind of learning is also known as *exact learning*).

There are different kinds of queries available to the learner, but just two of them have established themselves as the standard combination to be used:

- *Membership queries* (MQs): the learner asks if a string w is in the language, and the teacher answers "yes" if w belongs to the target language, and "no" otherwise.
- *Equivalence queries* (EQs): the learner asks if his hypothesis H is correct, and the teacher answers "yes" if H is equivalent to the target language

L and "no" otherwise. If the answer is "no", a counterexample x is returned (i.e., a string in the symmetric difference of H and L).

A teacher that can answer MQs and EQs is called a MAT teacher (minimally adequate teacher). In [2], Angluin gave an algorithm known as L^*, which learns DFA from MAT. She proved that is possible to learn DFA from MQs and EQs in polynomial time, and it was conjectured that richer classes than DFA cannot be inferred through a polynomial use of MAT. Since then, the L^* algorithm has become the main reference and one of the most relevant results in the framework of learning from queries. Below we briefly review the learning algorithm L^*. Details can be found in [2].

The L^* algorithm

The general idea of the algorithm is to repeat the following loop until the answer to an EQ is "yes":

- Find a closed and consistent observation table (representing a DFA) by means of MQs
- Ask an EQ
- If the answer is "no" (it is not the correct acceptor), then use the counterexample to update the table

What is an observation table? The information during the learning process is organized in a table called *observation table*. An observation table is a two-dimensional table, with both rows and columns indexed by strings (for example, see Figure 3).

We can differentiate three main parts in an observation table:

- S: a prefix-closed set of strings. Rows labeled by elements of S are the candidates for states of the automaton being constructed.
- T: in this part of the table we find rows labeled by elements of $S \cdot \Sigma$ (i.e., elements of S concatenated with all the symbols of the alphabet). These rows are used to construct the transition function.
- E: a suffix-closed set of strings. Columns labeled by elements of E correspond to distinguishing experiments for these states.

The observation table will be denoted (S, E, T). By concatenating the string of a row r with the string of a column c we get a string rc. If the string rc is in the language, the corresponding cell contains a 1, and 0 otherwise.

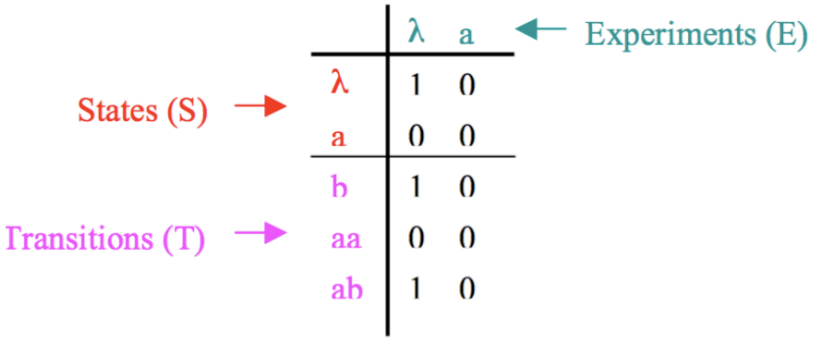

Fig. 3. Observation table. $\Sigma = \{a, b\}$

An observation table is called *closed* if any row of $S \cdot \Sigma$ corresponds with some row in S. An observation table is called *consistent* if every equivalent pair of rows in S remains equivalent after appending any symbol. When we have a closed and consistent table we can build the corresponding DFA and make an EQ.

How do we build a DFA? The L^* algorithm uses the observation table to build one. We define a corresponding automaton $A(S, E, T)$ over the alphabet Σ, with state set Q, initial state q_0, accepting states F, and transition function δ as follows:

- $Q = \{row(s) | s \in S\}$
- $q_0 = row(\lambda)$
- $F = \{row(s) | s \in S \text{ and } T(s) = 1\}$
- $\delta(row(s), a) = row(s \cdot a)$

For example, as the reader can easily verify, the observation table depicted in Figure 3 is closed and consistent. So, we can construct a DFA from this table. There are only two candidates for states: the row labeled λ and the row labeled a. The first contains 10 and the second 00; these values can be considered as a codification of the state. Therefore, we can call q_0 all the rows that have the value 10, and q_1 all the rows with the value 00. Now, by using the other rows, we know that: from q_0, by reading the symbol a, we go to state q_1 (value of the row labelled a), and by reading the symbol b, we go to state q_0 (value of the row labelled b); from q_1, by reading the symbol a we go to state q_1 (value of the row labelled aa), and by reading b we go to

state q_0 (value of the row labelled ab). Moreover, q_0 is both initial and final state. In this way, we can construct the corresponding automaton, which is depicted in Figure 4.

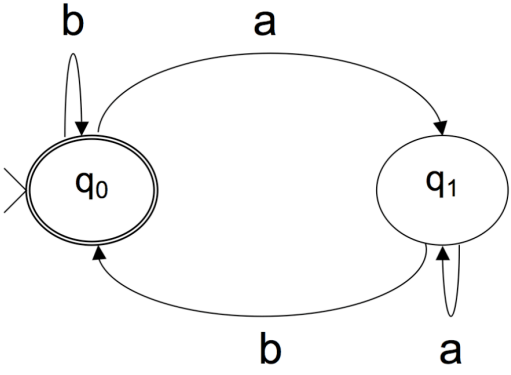

Fig. 4. Automaton corresponding to the observation table depicted in Figure 3

After making the EQ, if the conjectured DFA is the correct one, we will get a positive answer from the teacher. Then, the algorithm halts. If the conjectures DFA is not the correct one, we will get a counterexample. In such a case, we have to: i) Add the counterexample and all its prefixes to S; ii) Update the table using MQs for missing elements. We shall explain all these steps in greater detail using an example.

Running example

Let the alphabet $\Sigma = \{0, 1\}$, and a language $L = (0 + 110)^+$. The minimal automaton associated with the mentioned language is shown in Figure 5.

Initially the learner starts with the following observation table described as Table 1.

This table is not closed because $row(0)$ does not belong to $rows(S)$. L^* chooses to add the string 0 to S, 00 and 01 to $S\Sigma - S$, and then queries 00 and 01 to construct the observation table T_2 shown in Table 2.

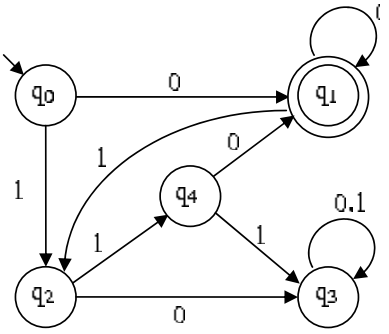

Fig. 5. Minimal automaton associated to the language $L_1 = (0 + 110)^+$

Table 1. $S = \{\lambda\}$, $E = \{\lambda\}$

T_1	λ
λ	0
0	1
1	0

Table 2. $S = \{\lambda, 0\}$, $E = \{\lambda\}$

T_2	λ
λ	0
0	1
1	0
00	1
01	0

This observation table is closed and consistent, so L^* makes a conjecture of the automaton A_1, shown in Figure 6.

A_1 is not a correct automaton for L, so the teacher selects a counterexample. In this case we assume that the counterexample 10 is returned (it is not in L but accepted by A_1).

To process the counterexample 10, L^* adds the strings 1 and 10 to S (the string λ is already in S), and queries the strings 11, 100 and 101 to construct the observation table T_3 shown in Table 3.

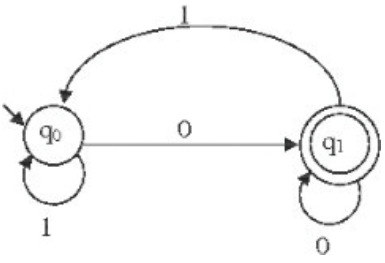

Fig. 6. Associated automaton: A_1

Table 3. $S = \{\lambda, 0, 1, 10\}$, $E = \{\lambda\}$

T_3	λ
λ	0
0	1
1	0
10	0
00	1
01	0
11	0
100	0
101	0

This observation table is closed, but not consistent since $row(\lambda) = row(1)$ but $row(0) \neq row(10)$. Thus L^* adds the string 0 to E, and queries the strings required to construct the observation table T_4 shown in Table 4.

Table 4. $S = \{\lambda, 0, 1, 10\}$, $E = \{\lambda, 0\}$

T_4	λ	0
λ	0	1
0	1	1
1	0	0
10	0	0
00	1	1
01	0	0
11	0	1
100	0	0
101	0	0

This observation table is closed, but not consistent since $row(1) = row(10)$ but $row(11) \neq row(101)$. Thus L^* adds the string 10 to E, and queries the strings required to construct the observation table T_5 shown in Table 5.

Table 5. $S = \{\lambda, 0, 1, 10\}$, $E = \{\lambda, 0, 10\}$

T_5	λ	0	10
λ	0	1	0
0	1	1	0
1	0	0	1
10	0	0	0
00	1	1	0
01	0	0	1
11	0	1	0
100	0	0	0
101	0	0	0

This observation table is closed and consistent, so L^* conjectures the automaton A_2 shown in Figure 7.

A_2 is not a correct acceptor for L, so the teacher answers the conjecture with a counterexample. We assume that the counterexample supplied is 11110, which is not in L but is accepted by A_2.

L^* adds the counterexample and all its prefixes to S and constructs the observation table T_6 shown in Table 6.

This table is found to be closed but not consistent, since $row(\lambda) = row(11)$ but $row(1) \neq row(111)$.

Thus L^* adds the string 110 to E and queries the necessary strings to construct the observation table T_7 shown in Table 7.

This table is closed and consistent. The automaton conjectured by L^* now corresponds to the correct acceptor for the language L, so the Teacher replies to this conjecture with *yes* and L^* terminates with this automaton as its output.

The total number of queries during this run of L^* is 3 EQs (the last one was successful) and 44 MQs.

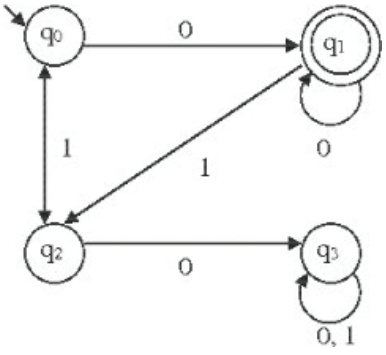

Fig. 7. Associated automaton: A_2

Table 6.
$S = \{\lambda, 0, 1, 10, 11, 111, 1111, 11110\}$,
$E = \{\lambda, 0, 10\}$

T_6	λ	0	10
λ	0	1	0
0	1	1	0
1	0	0	1
10	0	0	0
11	0	1	0
111	0	0	0
1111	0	0	0
11110	0	0	0
00	1	1	0
01	0	0	1
100	0	0	0
101	0	0	0
110	1	1	0
1110	0	0	0
11111	0	0	0
111100	0	0	0
111101	0	0	0

Table 7.
$S = \{\lambda, 0, 1, 10, 11, 111, 1111, 11110\}$,
$E = \{\lambda, 0, 10, 110\}$

T_7	λ	0	10	110
λ	0	1	0	1
0	1	1	0	1
1	0	0	1	0
10	0	0	0	0
11	0	1	0	0
111	0	0	0	0
1111	0	0	0	0
11110	0	0	0	0
00	1	1	0	1
01	0	0	1	0
100	0	0	0	0
101	0	0	0	0
110	1	1	0	1
1110	0	0	0	0
11111	0	0	0	0
111100	0	0	0	0
111101	0	0	0	0

3.3 Linguistic discussion of these models

The formal models presented in the previous section are based on different learning settings (i.e., the type of data used in the learning process and

the way in which these data are provided to the learner is different in both cases) and different criteria for a successful inference (i.e., the conditions under which we say that a learner has been successful in the language learning task are different). But, which one is better for modeling natural language acquisition? Below we review some of the accepted and controversial aspects of these models.

We can find some similarities between learning in Gold's model and first language acquisition. In both cases there is a process of *improvement*: in identification in the limit model, the new conjecture is better than the previous guess; in the case of first language acquisition there is a progressive improvement of the language acquired by the child. However, there are some aspects of Gold's model that are controversial from a linguistic point of view. For example:

- *In the limit* denotes the criterion of success, which assumes that there is no limit on how long it can take the learner to guess the correct language. Hence, considerations of efficiency form a somewhat separate line of analysis from Gold's work, which was concerned with limiting behavior rather than speed of learning. However, from the natural language acquisition point of view efficiency is also important. Although learning a natural language is an infinite process, we are able to learn the language in an efficient way.
- The learner passively receives strings of the language. However, we know that natural language learning is more than that: children also interact with their environment.
- The current hypothesis has to be consistent with all the examples seen so far. Moreover, the learner hypothesizes complete grammars instantaneously. From a linguistic point of view these assumptions are unrealistic (e.g., children are unlikely to remember the entire record of sentences ever addressed to them).

Therefore, the definition of identification in the limit postulates greatly idealized conditions, as compared to the conditions under which children learn language.

Angluin's model addresses an important tool available to a child, i.e., queries to a teacher (usually, a family adult member). Therefore, the query learning model might be useful when representing several aspects of the process of children's language acquisition. However, this model also has some controversial aspects from a linguistic point of view:

- The type of queries introduced with this model are quite un-natural for real learning environments. For example, an equivalency query will never be produced in a real situation; a child will never ask an adult if his grammar is correct.
- The learner does not really interact with the teacher; he can ask MQs or EQs, but he does not really communicate with the teacher by producing sentences, etc. In the communication between children and adults we can see that the role of the children is more active, and not limited to asking this kind of queries.
- Angluin's model is known as exact learning. However, from a linguistic point of view, everybody has (small) imperfections in their linguistic competence.
- The teacher in this model is assumed to know everything and always gives the correct answers. Therefore, he is an ideal teacher, which does not correspond with a real situation.

The third model studied in GI is called the PAC learning model (probably approximately correct), which was introduced by Valiant in [35]. It is a probabilistic model of learning from random examples; the distribution over the examples is unknown, and the examples are sampled under this distribution. The learner is required to be able to learn from this sample and under any probability distribution, but exactitude is not required (a small error is permitted since one may be unlucky during the sampling processes). Taking into account that exact learning is too hard in a real context, approximate learning could be a good way of dealing with children's language acquisition. However, the requirement that the examples have the same distribution throughout the process is too strong for practical situations.

As we have seen, all these models have aspects that make them suitable for studying natural language acquisition to a certain extent, but other aspects of the models make them unsuitable for this task. Therefore, we can conclude that none of these models perfectly accounts for natural language acquisition.

4 Towards a new formal model of language learning

As we have pointed out, the study and development of formal models of language learning is of great interest if we are to better understand the pro-

cess of natural language acquisition. In the section above we have seen that the models that have been proposed so far in GI have many controversial aspects from a linguistic point of view. Part of the reason is because GI studies have been specially focused on obtaining formal results, and they have been more interested in the mathematical aspects of the models than in their linguistic relevance.

Therefore, it would be interesting to develop new formal models of language learning that take greater account of studies of natural language acquisition (in this way, we could avoid some of the controversial aspects of the models proposed so far). In order to do this, it is important to address two questions: what classes of formal languages are interesting from a linguistic point of view?; what source of data should we provide our learning algorithm with? We try to answer these questions below.

4.1 What class of formal languages?

The theory of formal languages arose born in the second half of the 20th century as a tool to describe natural language syntax. As we have pointed out, the goal of GI studies is to learn formal languages from data. Most research into GI has focused on learning two classes of formal languages: regular and context-free languages (two of the classes with least generative power in the Chomsky hierarchy). However, what class of formal languages is more interesting from a linguistic point of view?

In order to answer this question, first, we need to answer the following question: Where are natural languages located in the Chomsky hierarchy? This question has been a subject of debate for a long time. This debate was focused on trying to determine whether natural languages are CF or not. In the late 80s, examples of structures that are not CF were discovered in several natural languages. Here are some examples of such constructions:

- **Dutch**: Bresnan et al. studied cross-serial dependencies in Dutch, arguing against the context-freeness of natural language.

  > *While Dutch may or may not be CF in the weak sense, it is not strongly CF: there is no CFG that can assign the correct structural descriptions to Dutch cross-serial dependency constructions.* [13, p. 314]

  The following example shows a duplication-like structure $\{w\bar{w} \mid w \in \{a,b\}^*\}$, where $\bar{w}$ is the word obtained from w by replacing each letter with its barred copy.

> *...dat Jan Piet Marie de Kinderen zag helpen laten zwemmen*
> (That Jan saw Piet help Marie make the children swim)

This is only *weakly* non-context-free, i.e., only in the deep structure.

- **Bambara**: Bambara, an African language of the Mande family, was studied by Culy in [19]. He provided another argument against context-freeness based on the morphology of words in that language.

> *In this paper I look at the possibility of considering the vocabulary of a natural language as a sort of language itself. In particular, I study the weak generative capacity of the vocabulary of Bambara, and show that the vocabulary is not context-free. This result has important ramifications for the theory of syntax of natural language.* [19, p. 349].

A duplication structure is found in the vocabulary of Bambara, demonstrating a strong non-context-freeness, i.e., on the surface and in the deep structure:

> *malonyininafilèla o malonyininafilèla o*
> (one who searches for rice watchers + one who searches for rice watchers = whoever searches for rice watchers)

This has the structure $\{wcw \mid w \in \{a,b\}^*\}$. But also the *crossed agreement* structure $\{a^n b^m c^n d^m \mid \text{m, n}{>}0\}$ can be inferred.

- **Swiss German**: The paper by Shieber [32], offers evidence for the non-context-freeness of natural language. He collected data from native Swiss German speakers, and provided a formal proof of the non-context-freeness of Swiss German.

> *Using a particular construction of Swiss German, the cross-serial subordinate clause, we have presented an argument providing evidence that natural languages can indeed cross the context-free barrier. The linguistic assumptions on which our proof rests are small in number and quite weak; most of the proof is purely formal. In fact, the argument would still hold even if Swiss German were significantly different from the way it actually is, i.e., allowing many more constituent orders, cases and constructions, and even if the meanings of the sentences were completely different.* [32, p. 330].

The following example is a strong non-context-free structure, again showing crossed agreement:

> *Jan säit das mer (d'chind)^m (em Hans)^n es huus haend wele (laa)^m (hälfe)^n aastriiche*
>
> (Jan said that we wanted to let the children help Hans paint the house)

This has the structure $xwa^m b^n yc^m d^n z$, where *a, b* stand for accusative, dative noun phrases, respectively, and *c, d* for the corresponding accusative, dative verb phrases, respectively.

So, all these studies provide a negative answer to the question of whether natural languages are CF or not. Moreover, they suggest that natural languages can only be described by a generative capacity that is greater than context-free grammar. But, how much power is needed to describe these non-CF constructions?

In 1985, Joshi [24] introduced the notion of the *Mildly Context-Sensitive* family of languages. The general idea was to provide a device that was able to generate CF and non-CF structures, but keep the generative power under control. There are very well known mechanisms for fabricating MCS families: for example, tree adjoining grammars, head grammars, combinatory categorial grammars. In the Chomsky hierarchy thay are somewhere between CF and CS. However, is it necessary for such formalisms to generate all CF languages? We can find natural language constructions that are neither REG nor CF, and also some REG or CF constructions that do not appear naturally in sentences. Therefore, as some authors point out [7, 26, 27], natural languages could occupy an orthogonal position in the Chomsky hierarchy.

So, it would be desirable to find new formalisms that have the following two properties: i) They are able to generate Mildly Context-sensitive languages (i.e., they generate multiple agreement, crossed agreement and duplication structures, and they are computational feasible); ii) They occupy an orthogonal position in the Chomsky hierarchy (i.e., thay contain some REG, some CF, and so on).

4.2 What source of data?

The learning paradigms that have most been studied in GI are: learning from positive data (most of them), and learning from queries. However, if

we want to correctly simulate natural language learning, we should provide our learning algorithm with the same kind of examples that are available to a child. But one of the questions that is still a subject of debate in Linguistics is precisely this: what source of data is available to children during the learning process?

It is widely accepted that children receive positive data; that is, sentences that are grammatically correct. However, the availability of another kind of data (called negative data) is still a matter of substantial controversy. Do children receive negative data and use them during the learning process?

There have been three main responses to this question. The first proposal is that children do not receive negative data and they must rely on innate information to acquire their native language. This proposal is based on the *poverty of stimulus* argument: there are principles of grammar that cannot be learnt from only positive data, and since children do not receive negative data (i.e., evidence about what is not grammatical), one can conclude that the innate linguistic capacity is what provides the additional knowledge that is necessary for language learning. Further justification for innateness was drawn from Gold's negative result on learning from positive data. Moreover, Brown and Hanlon [14] analyzed adult approval and disapproval of child utterances (for example, adult's answers such as"That's right", "Correct", "That's wrong", "No"). They found no relation between this type of answer and the grammaticality of the sentences produced by the children, and this was also taken to show that children do not receive negative data. However, it is worth noting that parents do not usually address their children in this way. Should only explicit disapproval count as negative evidence? Do adults correct children in a different way?

The second proposal is that children receive negative data in the form of *different reply-types* given in response to grammatical versus ungrammatical child utterances. Hirsh-Pasek et al. [22], Demetras et al. [20], and Morgan and Travis [29] proposed that parents respond to ungrammatical child utterances by using different types of answers from those they use when responding to grammatical utterances. Under this view, the reply type would indicate to the child whether an utterance was grammatically correct or not. For example, if parents tend to respond with an expansion when the child's utterance is incorrect, but repeat the sentences that are grammatically correct, then adult use of an expansion would signal that the child's utterance was incorrect. However, Marcus [28] analyzed all these studies and concluded that there is no evidence that this kind of data is necessary to learn a

language or even that they exist. Even if they exist, a child would learn what utterances are correct only after complex statistical comparisons. Therefore, these results were also used to show that internal mechanisms are necessary to explain how children get rid of errors to acquire their native language.

The third proposal is that children receive negative evidence in the form of *reformulations*, and not only do they detect them, they also make use of the information. Chouinard and Clark [15] proposed this new view of negative evidence. They consider that the reply-types proposal does not take into account if the adult's answer contains corrective information (then, answers that are corrective are grouped with those that are not). Therefore, if only the reply-type is taken into account, it could be difficult to identify the error made. On the basis of Clark's theory of contrast [17, 18], Chouinard and Clark proposed adult reformulations as negative evidence. They consider that it is in the to-and-fro of conversation that children receive information about whether their utterances are appropriate for their intended meanings. For example (extracted from CHILDES database, Kuczaj):

Abe: milk milk
Father: you want milk?
Abe: uh-huh
Father: Ok. Just a second and I'll get you some.

In this conversation, Abe is about two years and a half. She produces an incorrect sentence and, immediately after, the father reformulates her sentence by checking on what the child had intended to say. After that, the child acknowledges the reformulation. Therefore, as we can see: i) Adult correction preserves the same meaning of the child; ii) Adult uses the correction to keep the conversation on track (adult reformulates the sentence just to make sure that he has understood the child's intentions); iii) Child utterance and adult correction have the same meaning, but different form. Chouinard and Clark analyzed longitudinal data from five children between two and four years old, and they showed that adults reformulate erroneous child utterances often enough to help learning. Moreover, they showed that children not only detect differences between their own utterance and the adult reformulation, they also make use of the information.

Do corrections give positive or negative information? As we can see, these types of corrections contain positive and negative information at the same time. On the one hand, corrections are positive data, since a correction

is a sentence that is grammatically correct. On the other hand, they also give us negative information; as Chouinard and Clark pointed out:

> *Since, like adults, children attend to contrast in form, any change in form that does not mark a distinct, different, meaning will signal to children that they may have produced something that is not acceptable in the target language. And this fits the classic definition of negative evidence* [15, p. 666]

It is worth noting that during the first stages of children's language acquisition, children receive corrections that preserve the meaning of what they intend to convey. However, this kind of information has not been taken into account in formal models of language learning. Why should it not be taken into account? What is the effect of corrections on the process of language learning? A model that takes corrections into account could allow us to answer this question.

5 New proposals

As we have seen, REG and CF languages have a limited expressive power to describe some aspects of the syntax of natural languages. Moreover, corrections could play an important role during the process of language acquisition. Taking into account all these ideas, we shall briefly review two new lines of research that have been proposed in the last four years.

5.1 Learning Simple External Contextual Languages

We have pointed out in the section above that it would be desirable to have a mechanism that can generate MCS languages and occupy an orthogonal position in the Chomsky hierarchy. Becerra-Bonache [7] proposed and studied a non-classical mechanism that has these interesting properties: *Simple External Contextual* grammars (SEC).

A SEC produces a language starting from a string called *base*, and iteratively adding contexts (i.e., pair of strings) at the ends of the current string. Formally, a SEC grammar is defined as $G = (\Sigma, B, C)$, where:

- Σ: alphabet.
- B: one p-word (i.e., a p-dimensional vector whose components are words/ strings) over Σ, called the base of the grammar.

- C: a finite set of p-contexts (i.e., a p-dimensional vector whose components are contexts) over Σ, called the set of contexts of G.

Here is an example. Let us assume we have a SEC grammar with 2 dimensions, where: $\Sigma = \{a, b, c\}$, $B = \{(\lambda, \lambda)\}$, and $C = \{\ c_1 = [(a, b), (c, \lambda)]\}$. Starting from the base (λ, λ), if we apply the context once we obtain the 2-word $(a\lambda b, c\lambda\lambda) = (ab, c) = abc$. Now, starting from (ab, c), if we again apply the context we obtain $(aa\lambda bb, cc\lambda\lambda\lambda) = (aabb, cc) = aabbcc$. Note that by using this grammar, we can generate the following non-CF language: $L = \{a^n b^n c^n \mid n \geq 0\}$. The generation process is depicted in Figure 8.

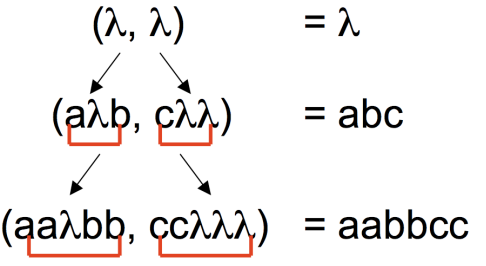

Fig. 8. Derivation process of the SEC grammar $G = (\Sigma = \{a, b, c\}$, $B = \{(\lambda, \lambda)\}$, and $C = \{\ c_1 = [(a, b), (c, \lambda)]\})$

Becerra-Bonache [7] proved that SEC can generate MCS languages and occupies an orthogonal position in the Chomsky hierarchy (see Figure 9). Moreover, the learnability of SEC from positive data has been studied in [8, 12, 30].

5.2 Learning from Positive Data and Corrections

As we have seen in the section above, studies on children's language acquisition show that corrections are available to children. Although the main source of information received during the process of natural language acquisition is positive data, corrections could play a complementary role in the process. Therefore, it is of great interest to study the effects of corrections on language learning.

Taking all this into account, Becerra-Bonache [7] tried to apply the idea of corrections to GI studies, and more concretely to the query learning model