Miniature Language Acquisition: A touchstone for cognitive science

Jerome A. Feldman, George Lakoff, Andreas Stolcke and Susan Hollbach Weber
International Computer Science Institute, Berkeley CA

Abstract

Cognitive Science, whose genesis was interdisciplinary, shows signs of reverting to a disjoint collection of fields. This paper presents a compact, theory-free task that inherently requires an integrated solution. The basic problem is learning a subset of an arbitrary natural language from picture-sentence pairs. We describe a very specific instance of this task and show how it presents fundamental (but not impossible) challenges to several areas of cognitive science including vision, language, inference and learning.

1 Introduction

> touchstone (tuch' ston'). n. 1. a black siliceous stone used to test the purity of gold and silver by the color of the streak produced on it by rubbing it with either metal. 2. a test or criterion for the qualities of a thing.
>
> —Syn. 2. standard, measure, model, pattern.

Among the things that cognitive science has studied most are visual perception, language, inference, and learning [Posner, 1989]. However, these are often studied as if they were isolated from one another. Studies in visual perception rarely address the questions of how we perceive higher-order spatial relations and what systems of spatial concepts there are in the languages of the world. Computer vision and natural language processing are seen as different and unrelated disciplines. In psychology as well, language and vision are seen as distinct subspecialties, where specialists in one have little or nothing to do with the other.

The study of learning is, for the most part, just as isolated. Learning research often proceeds as if the content of what is learned did not matter. This is especially true of connectionist learning, which studies the learning of correlations among microfeatures, independent of content. One partial exception to this is language acquisition in the generative tradition, where a good deal of innateness is assumed [Pinker, 1989]. But there language acquisition is defined in a very limited way: "language acquisition" usually means just syntax acquisition and research has not attempted to characterize how people learn to describe what they see.

One way to start to unify several branches of cognitive science is to ask the question:

How could we learn to describe what we see?

We believe that addressing this question seriously could change the course of research in several subfields in a healthy way. We believe that these fields need to co-evolve, taking into account one another's constraints.

We realize of course that each of these fields is enormous and complex and largely unknown and that anything like a total integration at present is impossible. However, we believe that it is possible to undertake a small but nonetheless significant portion of that task now, and that doing so will have a sobering and an enriching effect on much of cognitive science.

We are proposing a new touchstone problem for cognitive science, a mini-task that is well-defined, very small relative to the overall job to be done, and yet significant enough so that one can learn a great deal. The Miniature Language Acquisition (MLA) task in its most general formulation is to construct a computer system such that:

The system is given examples of pictures paired with true statements about those pictures in an arbitrary natural language.

The system is to learn the relevant portion of the language well enough so that given a new sentence of that language, it can tell whether or not the sentence is true of the accompanying picture.

There are a number of attractive features in this general task. It is strictly behavioral and theory-free: nothing has been said about the theory or methodology that should be employed in the task. The problem is closed in the sense that one cannot appeal to some forthcoming result in a related domain that will complete the story — the system has to do the whole job. There is no stipulation of how much should be built into the system and how much learned. And the requirement that the same system should work for equivalent fragments of any natural language rules out *ad hoc* solutions. The issue now becomes one of feasibility: whether there is an instance of this task that is currently approachable but still rich enough to meet our programmatic goals.

Of course, the MLA task is not a model of human language acquisition or adult language learning. The semantic and pragmatic context of real human communication is far too complex to use as a basis for a Miniature Language Acquisition task. Even the domain of idealized two-dimensional scenes, which is the simplest we could find, involves considerable complexity as some coming examples will illustrate. In fact, one of the greatest appeals of the MLA task is that deep questions in several areas of cognitive science appear in sharp form, even in the very limited domain envisaged here.

2 L_0: A specific formulation

We have been investigating the particular task (which we call L_0) of language acquisition in the domain of simple two-dimensional scenes. In order to define the scope of the task precisely, both in the linguistic and in the conceptual domain, we give a set of syntactic rules for L_0 (see Figure 1). The MLA task requires that the system learn equivalent fragments of any natural language. We (tentatively) characterize an "equivalent fragment" as one that can describe the same range of visual inputs while using the simplest (most unmarked, pragmatically most neutral) grammatical realization the language provides.

We would like to make it clear that using a simple phrase structure grammar for a fragment of English (which, for practical reasons, is our 'base' language) as a specification is a mere matter of convenience. We use it to implicitly constrain the conceptual domain, i.e. the set of objects, their attributes and relations, allowed by L_0. Neither do we want to imply that the learning should derive exactly these syntactic rules, nor that a grammar of similar structure even exist for other languages. Instead we take the conceptual domain as the cross-linguistic common denominator in our learning task.[1] For every language besides English a suitable language fragment has to be

[1] Of course there is the possibility that no such common conceptual denominator exists. This is unlikely given that all languages and corresponding conceptual systems are constrained by a common perceptual apparatus. However, if it turned out that the L_0 definition given above is meaningless due to such fundamental conceptual differences, we would consider this finding in itself a valuable outcome of the research task proposed here.

S = NP | NP VP
NP = DET NP1 | DET NP1 and DET NP1
VP = VI PP | VT NP
NP1 = OBJ | SHADE OBJ | SIZE OBJ | SIZE SHADE OBJ
PP = REL NP
VI = is | are
VT = touches | touch
DET = a
OBJ = circle | square | triangle
SHADE = light | dark
SIZE = small | medium | large
REL = REL1 | far REL1
REL1 = above | below | to the left of | to the right of

Figure 1: A syntactic specification of L_0 for English.

---

determined independently. For many Indo-European languages, the syntactic specification will turn out to be close to the English version, in other cases radically different grammatical structures will have to be used.

L_0 scenes consist of up to four objects drawn from a population of three shapes (circle, square and triangle), and two distinct shades (light and dark) (see Figure 2). Objects can be of arbitrary size and position within the limits imposed by visual discernability, the image boundaries, and the additional constraint that objects may not occlude or overlap one another.

A candidate system is presented with a picture and with one or more sentences that are grammatical in the test language and are true of that picture. The system designer can specify that the training examples be presented according to some specific rule such as lexicographic order or random selection according to some distribution. We explicitly allow an isolated noun phrase (NP) as a sentence fragment as this should simplify the initial stages of learning. After training on no more than half the examples (and hopefully many fewer) the system is tested by being presented with pairs consisting of a picture and a grammatical sentence which may be true or false about the companion picture. Obviously, the system succeeds to the extent that it produces the right answers. It is important to realize that the grammar for the test language is hidden from the system. It is known only to the generator of the training inputs, and the restrictions it embodies (though not these particular rules) must somehow be discovered by the learning system.

From the point of view of linguistics, the task has important attractions. First, there are already enough rich descriptions of spatial systems for various Indo-European and non-Indo-European languages for a start to be made [Rudzhka-Ostyn, 1988; Langacker, 1987; Casad, 1982; Casad and Langacker, 1985; Hershkovits, 1986; Janda, 1986; Talmy, 1983; Talmy, 1972; Talmy, 1985]. Second, the task, even in its simplest form, is demanding enough so that much deeper research on those spatial systems will be required. Third, the task focuses the attention of cognitive science on languages other than English, with special attention to the non-Indo-European languages, where the spatial systems are often very different from what English speakers are used to. Fourth, the task is semantically driven, which will require serious attention to the relation between syntax and semantics.

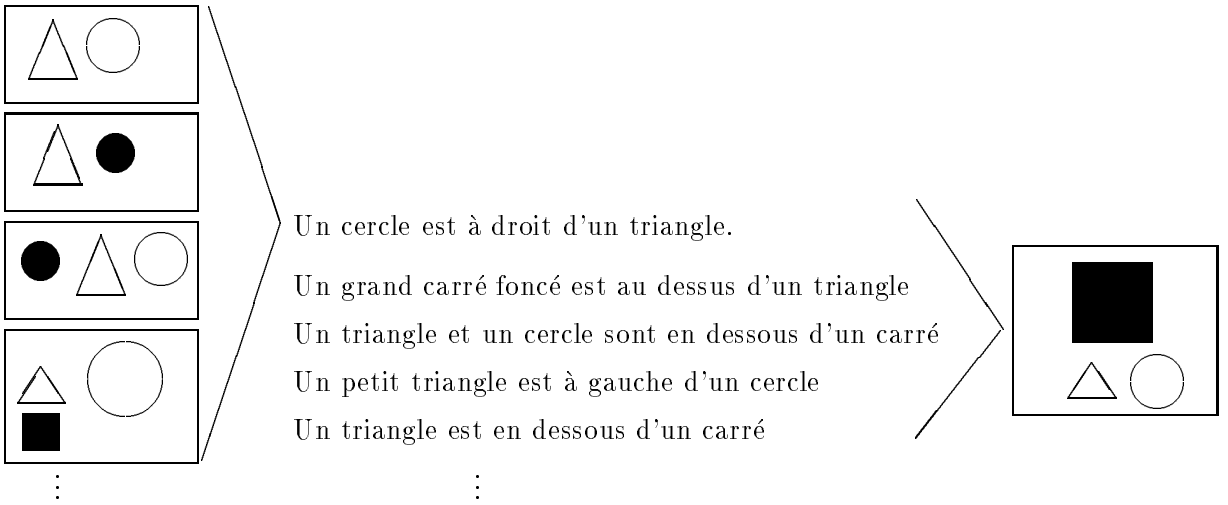

Figure 2: A given picture (right) has a large but finite number of applicable descriptions (shown here in French, center). Similarly, any given description is consistent with a very large, but also finite, set of scenes (left).

3 Variants

One advantage of the highly specific formulation of the L_0 task is that it focuses attention on what we consider to be the essence of the problem: acquiring syntactic descriptions of a limited but grounded semantics. The acquisition strategy should of course be extensible to a broader semantic range with the same grounding.

In addition to the base L_0 of Figure 1 we are looking at small variations. A specification that can be derived from Figure 1 by adding up to two words and two grammatical constraints is an acceptable *variant* of the task.[2] One would not be happy with a system that worked for exactly L_0 but totally broke down for one of these minor variants. We think of a solution as *robust* if its designer can revise it to accommodate any single L_0 variation in one day. We are only interested in robust solutions, namely, those that are easily modifiable for an enormous range of minor variants in any natural language. This should guarantee that any robust solution must be doing quite a few things right. As we shall see, adding even one minor variation can produce a great deal of complexity.

Note that the task as stated does not explicitly entail that linguistically significant generalizations be learned. We speculate that the robustness condition will guarantee a reasonable level of significant generalization. If a system can be extended simply to deal with any one of a very large number of variants, it would most likely have to have generalized pretty well.

Some of the L_0 variants that we have found useful to work with are:

1. Synonyms: e.g., have 'big' and 'large' used interchangeably.

[2] Again we would like to emphasize that English syntax is used here merely as a convenient (if somewhat arbitrary) instrument to indirectly specify the intended semantic scope, thus avoiding a specific formalization of those semantics.

2. Abstraction: add 'thing' to the definition of OBJ. This shows that one cannot simply identify an object with its shape.

3. Predicate negation: add "is not" and "are not" to the definition of VI. Negative sentences say much less about a picture, e.g. "A dark circle is not touching a square".

4. Sentence inversion: Add "PP VI NP" to the definition of S, e.g. "Above a circle is a small square". This forces the system to use grammatical cues in figuring out role assignment, rather than merely using relative position.

5. Plurals: add 'circles' and 'triangles' to the definition of OBJ, e.g. "A circle and a square are below large triangles".

6. Verb conjuncts: add "VT 'and' VI PP" to the definitions of VP, e.g. "A circle touches and is above a square". This makes explicit the requirement of allowing multiple references to a given object, e.g. "A circle touches a square and *the same circle* is above *the same square*".

7. Conjunctive attributes: add "REL1 'and' REL1" to the definition of REL. Conjunction can be subtle. For example, the sentence "A circle is above and to the left of a square" does not require that the circle be either 'above' or 'to the left of' the square.

8. Relative sizes: Add 'larger' and 'smaller' to definition of SIZE, e.g. "A circle is above a larger square". Expanding the scope of relational properties to comparatives highlights the necessity of adopting a visual representation that can handle abstract shape features and relations. One candidate is Ullman's [1984] visual routines. As Ullman points out, shape properties (e.g. connectedness) and relations (e.g. inside) can be computed by routines but are very hard to capture in a propositional semantics. Another L_0 example of this is the modifier 'far'.

9. Definite article: add 'the' to the definition of DET. Compare Figure 4 (c) with "The smaller circle is above the larger square".

10. Over and Under: add 'over' and 'under' to the definition of REL1. This is a far from trivial extension, since the term 'over' has dozens of related spatial senses [Lakoff, 1987]. For example, in Figure 3 (a), the circle is over (above) the square, while it is not in the topologically identical situation shown in Figure 3 (b). Another interesting asymmetry between 'under' and 'over' is depicted in Figure 4 (c).

11. Quantifiers: add 'every' and 'no' to the definition of DET, e.g. in Figure 3 (a), "No square is under a circle".

Another possible set of variants involve motion and time and entail a whole range of new sentences and representations and inference issues. The six variants listed below should provide some of the flavor of the additional considerations.

1. Single object motion: add 'moved' to the definition of VT,

   and/or add "is now" to the definition of VI,

   and/or add 'was' to the definition of VI.

   Temporal variants raise the issue of object identity. It seems reasonable to assume that the system is given the inter-scene object correspondences.

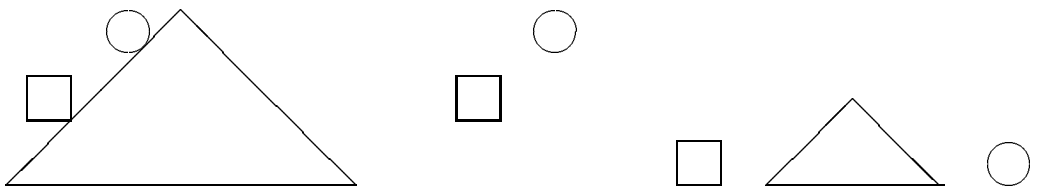

Figure 3: An interesting situation that highlights the complexity of the L_0 domain. (a) with the triangle as a frame of reference, the circle is seen to be 'over' the square; (b) without the apparent support of the triangle, the circle is no longer 'over' the square. Dialects differ on this. (c) The circle is over the triangle from the square.

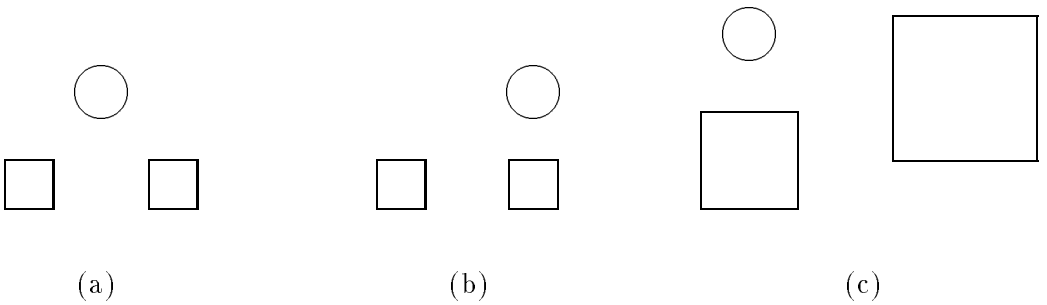

Figure 4: (a) the circle is 'over' the two squares, but the two squares are *not* 'under' the circle. (b) the circle is not over two squares (it is only over one of them). (c) The effect of context: "A smaller circle is above a larger square" is an acceptable description only with the indefinite article.

---

2. Single object change: add "turned into" to the definition of VT.
3. Single object contact: add "bumped into" to the definition of VT.
4. Temporal non-change: add 'remains' to the definition of VI.
5. Pronoun: add 'it' to the definition of NP, e.g. "It is now above it". Multiple scenes permit pronominal reference.
6. Trajectories: add "went between" to the definition of VT.

The examples involving motion and time suggest the need for an extended range of semantic primitives. For a variety of reasons [Feldman, 1988], we postulate that explicit trajectories will be an important primitive. It turns out that the trajectories are also useful in understanding some static scenes such as Figure 3 (c).

References

[Badler, 1975] N. I. Badler, "Temporal Scene Analysis: Conceptual Description of Object Movements," Technical report, Technical Report TR-80, Department of Computer Science, University of Toronto, 1975.

[Casad, 1982] Eugene H. Casad, "Cora Locational and Structured Imagery," Technical report, Doctoral Dissertation, University of California at San Diego, 1982.

[Casad and Langacker, 1985] Eugene H. Casad and Ronald Langacker, "'Inside' and 'Outside' in Cora Grammar," *International Journal of American Linguistics*, 51:247–281, 1985.

[Feldman, 1988] Jerome A. Feldman, "Time, Space and Form in Vision," Technical report, TR-88-011, International Computer Science Institute, Berkeley CA., 1988.

[Harris, 1989] Catherine Harris, "A connectionist approach to the story of 'over.'," Technical report, Proceedings of the Berkeley Linguistics Society, 15, University of California, Berkeley CA, 1989.

[Hershkovits, 1986] Annette Hershkovits, *Language and Spatial Cognition; An Interdisciplinary Study of the Preposition in English*, Cambridge: Cambridge University Press, 1986.

[Herzog *et al.*, 1989] G. Herzog, C. K. Sung, E. Andre, W. Enkelmann, H.-H. Nagel, T. Rist, W. Wahlster, and G. Zimmermann, "Incremental Natural Language Description of Dynamic Imagery," In W. Brauer, editor, *Proceedings of the Third International GI Congress '89*. New York: Springer, 1989.

[Hildreth and Ullman, 1989] Ellen C. Hildreth and Shimon Ullman, "The Computational Study of Vision," In Michael I. Posner, editor, *Foundations of Cognitive Science*, pages 581–630. Bradford Books, MIT Press, Cambridge MA, 1989.

[Jackendoff, 1983] Ray Jackendoff, *Semantics and Cognition*, MIT Press, Cambridge MA, 1983.

[Janda, 1986] Laura Janda, "A Semantic Analysis of the Russian Verbal Prefixes ZA-, PERE-, DO-, and OT-," *Slavistiche Beitrage, Munich: Sagner*, 192, 1986.

[Lakoff, 1987] George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press, 1987.

[Langacker, 1987] Ronald Langacker, *Foundations of Cognitive Grammar Volume 1*, Stanford: Stanford University Press, 1987.

[Miller and Johnson-Laird, 1976] George A. Miller and Philip Johnson-Laird, *Language and Perception*, Harvard University Press, Cambridge MA, 1976.

[Pinker, 1989] Stephen Pinker, "Language Acquisition," In Michael I. Posner, editor, *Foundations of Cognitive Science*, pages 359–400. Bradford Books, MIT Press, Cambridge MA, 1989.

[Posner, 1989] Michael I. Posner, *Foundations of Cognitive Science*, Bradford Books, MIT Press, Cambridge MA, 1989.

[Rudzhka-Ostyn, 1988] Brygida Rudzhka-Ostyn, editor, *Topics in Cognitive Linguistics*, Philadelphia: John Benjamins, 1988.

[Rumelhart, 1989] David E. Rumelhart, "The Architecture of Mind: a Connectionist Approach," In Michael I. Posner, editor, *Foundations of Cognitive Science*, pages 133–160. Bradford Books, MIT Press, Cambridge MA, 1989.

[Siskind, 1990] J. M. Siskind, "Acquiring Word Meanings," forthcoming thesis, MIT, 1990.

[Smith, 1989] Edward E. Smith, "Concepts and Induction," In Michael I. Posner, editor, *Foundations of Cognitive Science*, pages 501–526. Bradford Books, MIT Press, Cambridge MA, 1989.

[Sopena, 1988] Josep Maria Sopena, "Verbal Description of Visual Blocks World Using Neural Networks," Technical report, Technical Report, Departament de Psicologia Basica, Universitat de Barcelona, 1988.

[Stolcke, 1990] Andreas Stolcke, "Learning Feature-based Semantics with Simple Recurrent Networks," Submitted to the 12th Annual Conference of the Cognitive Science Society, MIT, July 1990.

[Talmy, 1972] Leonard Talmy, "Semantic Structures in English and Atsugewi," Technical report, Doctoral Dissertation, University of California at Berkeley, 1972.

[Talmy, 1983] Leonard Talmy, "How Language Structures Space," In Herbert Pick and Linda Acredolo, editors, *Spatial Orientation: Theory, Research, and Application*. New York: Plenum Press, 1983.

[Talmy, 1985] Leonard Talmy, "Force Dynamics in Language and Thought," In *Papers from the Parasession on Causatives and Agentivity at the Twenty-First regional Meeting of the Chicago Linguistic Society*, pages 293–337, 1985.

[Ullman, 1984] S. Ullman, "Visual Routines," *Cognition*, 18:97–157, 1984.

[Vandeloise, 1984] Claude Vandeloise, "Description of Space in French," Technical report, Doctoral Dissertation. University of California at San Diego, 1984.

[Waltz and Boggess, 1979] David L. Waltz and Lois C. Boggess, "Visual Analog Representations for Natural Language Understanding," In *IJCAI-79*, pages 926–934, 1979.

[Weber and Stolcke, 1990] Susan Hollbach Weber and Andreas Stolcke, "L_0: A Testbed for Miniature Language Acquisition," Technical report, TR 90-010, International Computer Science Institute, Berkeley CA., 1990.

[Wexler and Culicover, 1980] Kenneth Wexler and Peter Culicover, *Formal Principles of Language Acquisition*, Cambridge, Mass.: M.I.T. Press, 1980.

[Winograd, 1971] Terry Winograd, "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language," Technical report, MAC-TR-84, MIT, Cambridge, MA, January 1971.