Situated Grounded Word Semantics

Luc Steels
SONY CSL - Paris
VUB AI Laboratory - Brussels
steels@arti.vub.ac.be

Frederic Kaplan
SONY CSL - Paris
LIP6 Universite Paris VI - Paris
kaplan@csl.sony.fr

Abstract

The paper reports on experiments in which autonomous visually grounded agents bootstrap an ontology and a shared lexicon without prior design nor other forms of human intervention. The agents do so while playing a particular language game called the guessing game. We show that synonymy and polysemy arise as emergent properties in the language but also that there are tendencies to dampen it so as to make the language more coherent and thus more optimal from the viewpoints of communicative success, cognitive complexity, and learnability.

1 Introduction

The goal of studying natural language semantics is to determine the systematic relations between language utterances, their meanings and their referents. Speakers must *conceptualise* reality to find an adequate meaning, and they *verbalise* this meaning to yield an utterance transmitted to the hearer. The hearer must *interpret* the utterance to reconstruct the meaning and *apply* the meaning in this particular context to retrieve back the referent. One possible framework for studying these relationships is the theory of formal (denotational) semantics and its application to the treatment of natural language [Montague 79]. In this framework, functions are defined for mapping natural language utterances into expressions in the predicate calculus (with suitable extensions) and for mapping logical expressions into their denotations. Such a framework has been the basis of much work in computational semantics and has been used to formalise the communication systems of autonomous agents.

Although this formal approach has many virtues, it makes a number of simplifying assumptions which are not valid for physically grounded evolving autonomous agents that are part of inhomogeneous populations operating in real world open environments. In such circumstances, it is first of all not possible to formalise the model which acts as the domain of the denotational semantics. The physical environment produces an infinite set of unforeseeable situations, independent of the agent and not delineatable by an outside observer. Because of the open-ended nature of the environment, the language cannot be designed (or pre-programmed) in advance but must expand or shrink to adapt to the changing task settings and evolving environments encountered by the agents. Second, the agents have to learn autonomously the language used in their community, including its underlying ontology (the denotational semantics of the predicates). Consequently, it cannot be assumed that this language is uniform throughout the population. There are going to be different degrees of detail in the different agents depending on the histories of interaction with the environment. Agents cannot inspect each other's brain states nor is there a central controlling agency that oversees additions to the language, so that polysemy (one meaning having different forms) and synonymy (one form having different meanings) are unavoidable. Finally, because the agents are situated and embodied in the environment, they must bootstrap their competence through strongly context and viewpoint dependent cases. For example, something that is to the left for one agent may be on the right for another one depending on their respective positions, which makes it hard to learn the meanings of ieft' and 'right'.

We have been developing a framework for studying situated grounded semantics that does not make the simplifying assumptions of classical formal semantics and is therefore more appropriate for studying human natural language communication or designing robotic agents. The framework consists of a theory on how agents could construct and acquire their individual ontologies, lexicons and grammars, and of tools for studying the macroscopic properties of the collective ontologies, lexicons and grammars that emerge in the group. These and similar efforts have been reviewed in [Steels 97]. We are also conducting experiments with physically instantiated robotic agents, both mobile robots moving around in their en-

vironment with relatively weak sensors [SteelsVogt 97] and immobile robots with active vision [Steels 98]. The latter experimental infrastructure is used in the present paper and is known as the Talking Heads experiment.

One of the major themes of our research is that language can be studied as a complex adaptive system. The agents are carriers of individual linguistic knowledge which becomes overt behavior in local interactions between agents. There is no central coordination nor access of one agent to the internal states of another. The local interactions shape and continuously reshape the language. The overall dynamics should exhibit certain emergent properties, which have been observed as universal tendencies in natural languages, such as the emergence of a globally coherent linguistic system or the damping of synonymy and polysemy. Given this broader framework, this research is strongly related to other efforts to understand how autonomous grounded agents may build up their intelligence [SteelsBrooks 96], and particularly how they may bootstrap language [Hurford, et.al. 1988]. It draws on other research in the origins of complexity in natural systems [Langton 1995].

The rest of the paper first briefly introduces the Talking Heads experiment. Then general emergent properties of languages are reviewed. Next, results from experiments in lexicon acquisition are shown and analysed with respect to the reaching of coherence and the damping of polysemy and synonymy.

2 The Talking Heads experiment

The experimental setup consists of two (possibly more) Sony EV1D31 pan-tilt cameras in which different agents can be loaded (figure 1). Agents can travel through the Internet and install themselves in robots in different locations. An agent can only interact with another one when it is physically instantiated in a body and thus perceive the shared environment. For the experiments to be reported here, the shared environment consists of a magnetic white board on which various shapes are pasted: colored triangles, circles, rectangles, etc. The interaction between agents takes the form of a language game, further called the guessing game.

The guessing game

The guessing game is played between two visually grounded agents. One agent plays the role of *speaker* and the other one then plays the role of *hearer.* Agents take turns playing games so all of them develop the capacity to be speaker or hearer. The objects located on the white board at the beginning of the game constitute the *context.* Agents are capable of segmenting the perceived image into objects and of collecting various characteristics about each object, specifically the color (decomposed in RGB channels), grayscale, and position in pan/tilt coordinates. The speaker chooses one object

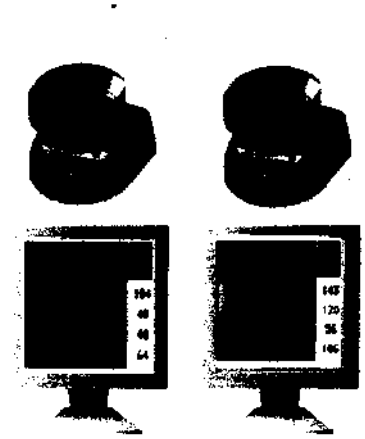

Figure 1: Two Talking Head cameras and associated monitors showing what each camera perceives.

from the context, further called the *topic,* and gives a linguistic hint to the hearer.

The linguistic hint is an expression that identifies the topic with respect to the other objects in the context. For example, if the context contains [1] a red square, [2] a blue triangle, and [3] a green circle, then the speaker may say something like "the red one" to identify [1] as the topic. If the context contains also a red triangle, he has to be more precise and say something like "the red square". Of course, the Talking Heads do not say "the red square" but use their own language and concepts which are never going to be the same as those used in English. For example, they may say "malewina" to mean [UPPER EXTREME-LEFT LOW-REDNESS]. Due to space limitations, this paper only considers situations where the meaning consists of a single perceptually grounded category and the form consist of a single word.

Based on the linguistic hint, the hearer tries to guess what topic the speaker has chosen, and he communicates his choice to the speaker by pointing to the object. A robot points by transmitting in which direction he is looking. The game succeeds if the topic guessed by the hearer is equal to the topic chosen by the speaker. The game fails if the guess was wrong or if the speaker or the hearer failed at some earlier point in the game. In case of a failure, the speaker gives an extra-linguistic hint by pointing to the topic he had in mind, and both agents try to repair their internal structures to be more successful in future games.

The architecture of the agents has two components: a conceptualisation module responsible for categorising reality or for applying categories to find back the referent in the perceptual image, and a verbalisation module responsible for verbalising a conceptualisation or for interpreting a form to reconstruct its meaning. Agents start with no prior designer-supplied ontology nor lexi-

fore two synonyms and " tisame" is polysemous; it can meanboth [GRAY-0.0,0.25] and [HPOS-0.5,0.75]. Three words are used to refer to *object- 1* This kind of situation is typical in certain stages of our experiments and complexity rapidly increases when the same meaning is also used to denote other referents (which is obviously very common and indeed desirable).

As mentioned earlier, incoherence is not necessarily impinging on the communicative success of the language. The RMF-iandscape in 3 still leads to total success in communication whenever both meanings are equally adequate for picking out the referent. Even if a speaker uses "tisame" to mean [GRAY-0.0,0.25] and the hearer understands "tisame" to mean [HPOS-0.5,0.75], they still have communicative success. The goal of the language game is to find the referent. It does not matter whether the meanings are the same. The agents cannot even know this because they have no access to each other's brain states.

The degree of coherence of a language can be measured by observing the actual linguistic behavior of the agents while they play language games, more specifically, by collecting data on the frequency of co-occurrence of items such as the possible forms of a certain referent or all the possible meanings for a certain form. The relations are represented in competition diagrams, such as the RF-diagram in figure 5, which plots the evolution of the frequency of use of the Referent-Form relations for a given referent in a series of games. One co-occurrence relation will be most frequent, and this is taken as an indication how coherent the community's language system is along the dimension of the relation investigated. For example, if a particular meaning has only one form, then the frequency of that form in the MF-diagram will be 1.0, which means that there are no synonyms.

The remainder of this paper now looks at a particular case study performed with the Talking Heads as currently operational. To allow this investigation, we restrict the set of possible referents (by keeping the environment constant) so that we can indeed track the grounded semantic dynamics forming and deforming the semiotic landscape.

4 Damping synonymy and polysemy

Figure 4 shows typical results for an experiment in which 20 agents start from scratch to build a new communication system, both the ontology, by the growing and pruning of discrimination trees, and the lexicon, by creating new words and adopting them from each other. As communicative success is reached, there is an evolution towards a unique form for each referent, as illustrated in figure 5. This is expected because the agents get explicit feedback only about this relation, not about any other one. This diagram shows that there must be a damping

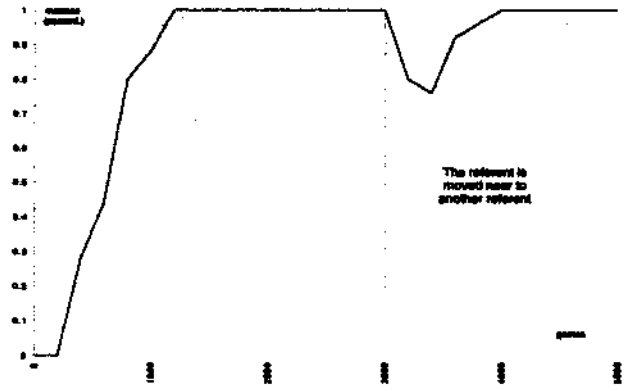

Figure 4: This graph shows the average success per 200 games in a series of 5000 games played by 20 agents. The agents evolve towards total success in their communication after about 1000 games. A change in the environment induced after 3000 games gives a decrease in average success which rebounds quickly.

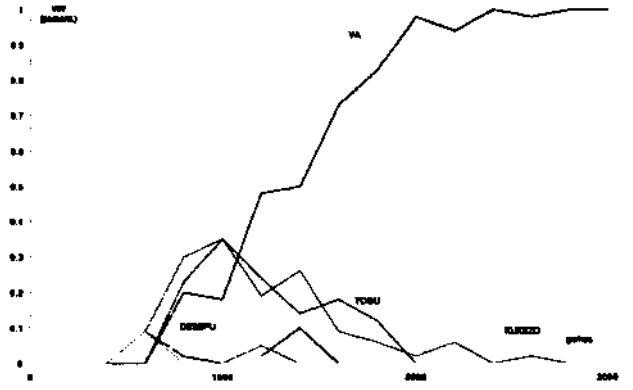

Figure 5: This RF-diagram shows the frequency of each referent-form co-occurrence in 3000 language games for a single referent. One word "va" comes to dominate.

of synonymy as well (and it is even clearer if we look at the RM-diagram).

When we inspect the different meanings of "va", through the FM-diagram (figure 6), we clearly see that even after 3000 games polysemy stays in the language. Three stable meanings for "va" have emerged: [RED-0,0.125], [BLUE-0.3125,0.3125], and [VPOS-0.25,0.5]. They are all equally good for distinguishing the topic "va" designates, and there are no situations yet that would have allowed disentanglement.

In game 3000, the environment produces a scene in which a category which was distinctive for the object designated by "va" is no longer distinctive. More precisely, we, as experimenters, have moved the object very close to another object so that the position is no longer distinctive. Figure 4 shows first of all that success drops (meaning there have been some failures in the game),

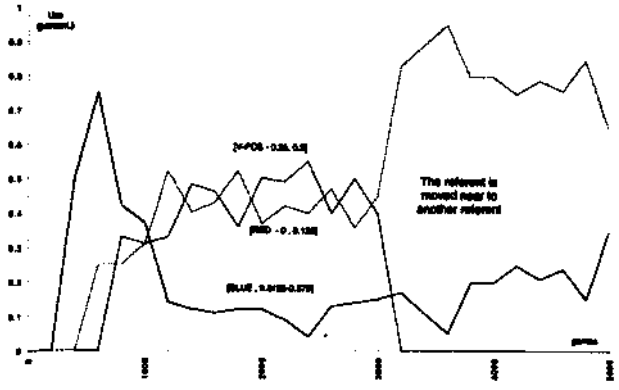

Figure 6: This FM-diagram shows the frequency of each form-meaning co-occurrence for "va" in a series of 5000 games. A disentangling situation arises in game 3000 causing the loss of one meaning of "va".

but that it rebounds quickly. The agent's language and ontology is adaptive. What has happened is that a large part of the agents still keep using "va", because the color-based categories are still applicable. But "va" no longer picks out the right object for those who believe that "va" means [VPOS-0.25,0.5], so they have to learn an alternative meaning for "va", compatible with the new situation. The FM-diagram in figure 6 shows that the positional meaning of "va" (namely [VPOS-0.25,0.5]) has disappeared. The other meanings, based on color, are still possible because they are not affected when the object designated by "va" moved its position.

This case study illustrates the main points of the paper. The overt selectionist force on the language system is success in the game, which is maximised if the agents use the same word for designating the same referent (in the same context). This does not in itself imply that there are no synonyms nor polysemy because agents could prefer different words which they mutually know from each other and they could associate different meanings with a certain word which are nevertheless compatible with the environments they have seen. We have shown that nevertheless damping of synonymy and polysemy occurs. Synonymy is damped because of the lateral inhibition of alternatives in lexicon use. This creates a positive feedback loop in which words that have a slight advantage will further gain in a winner-take-all process. Polysemy is damped because there are situations disentangling incoherent form-meaning relations.

5 Conclusions

The construction and acquisition of grounded languages poses specific difficulties for autonomous agents, causing their languages to exhibit partial incoherence. Agents have to develop their own categories for approaching their world, and they cannot know which meaning has been intended by a speaker, even if they know the referent through extra-linguistic means. We have shown that a particular agent architecture can achieve the bootstrapping of a language system from scratch and that the collective dynamics it generates dampens synonymy and polysemy.

6 Acknowledgement

This research was conducted at the Sony Computer Science Laboratory. We are strongly indebted to Angus McIntyre for creating the Babel tool that is the technical backbone of our experiments and for technical support in the low level vision interfacing.

7 References

[Hurford *et al.* 98] Jim Hurford, C. Knight and M. Studdert-Kennedy (eds.) *Approaches to the Evolution of Human Language.* Edinburgh Univ. Press. Edinburgh, 1997.

[Langton 95] Langton, C. [ed.] Artificial Life. An overview. The MIT Press, Cambridge MA, 1995.

[McLennan 91] McLennan, B. Synthetic ethology: An approach to the study of communication. In Langton, C. (ed.) *Artificial Life II,* Redwood City, Ca. 1991, Addison-Wesley Pub. Co.

[Montague 79] Montague, R. Formal Philosophy. University of Chicago Press, Chicago, 1979.

[Oliphant 96] Oliphant, M. The dilemma of Saussurean communication. *Biosystems,* 37 (l-2):31-38.

[Steels 97] Luc Steels. The synthetic modeling of language origins. *Evolution of Communication,* l(l):l-35.

[Steels 98] Steels, L. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103 (1998) 1-24.

[Steels *k* Kaplan 98] Steels, L. and Kaplan, F. Spontaneous Lexicon Change, In *Proceedings of COLIN G-ACL,* Montreal, August 1998, p. 1243-1249.

[Steels & Vogt 97] Steels, L. and P. Vogt Grounding adaptive language games in robotic agents. In Harvey, I. et.al. (eds.) *Proceedings of ECAL 97,* Brighton UK, July 1997. The MIT Press, Cambridge Ma., 1997.

[Steels & Brooks 95] Steels. L. and R. Brooks The Artificial Life Route to Artificial Intelligence. Lawrence Erlbaum, New Haven, 1995.