The Emergence of Language

Abstract

Language is clearly an emergent phenomenon that is critical to societal development. In this paper I discuss how a consistent language can be developed by a population that experiences pair-wise communication interactions. Both vocabulary and grammar are shown to emergent in this model. The stability of the emergent grammar is also discussed.

Rob Putman

November 27, 2006

Introduction

Language is the fundamental building block of human communication and is hence at the root of much of human society. The means to communicate is crucial to our survival and societal development. A language is a combination of two things: lexicon and grammar. The lexicon, or vocabulary, is a list of sounds (or in the case of this paper symbols) that has a direct mapping to a meaning:

$$\text{Lexicon(sound)} \rightarrow \text{meaning}$$

for the purposes of this paper we'll assume that the meanings already exist in the brain. Grammar is a mapping between various meanings and the integrated meaning:

$$\text{Grammar}(\text{meaning}_1, \text{meaning}_2, \ldots) \rightarrow \text{integrated meaning}$$

Since words can only be spoken in sequences grammar is a mapping between the order of words and the way they form a meaning. Grammar is the reason that the sentences "You ate an apple" and "An apple ate you" have different integrated meanings, even though they contain the same root meanings(words).

Clearly for a language to be a successful means of communication both the lexicon and grammar must be understood and used by all the people who are trying to communicate. In this paper I will discuss various models for the development of a language and show how both a shared lexicon and grammar are emergent phenomenon.

The Emergence of a Language

Much of this paper will be based on a model by Tao Gong and William S-Y. Wang. I will briefly discuss various aspects of this model now. I will only introduce enough information to explain their results here. An interested reader is encouraged to read their paper[3].

The Setup

The setup of the model is basically this: there are a list of *Agents*(people) who try to communicate with each other. To do this the agents are grouped into pairs and within each pair one agent tries to communicate with the other. For concreteness let me say that agent A is communicating with agent B. One agent (say A) first selects a meaning to try and express. If A can assemble the meaning from the language rules already known A will select the *Utterance*(sentence) that has the highest weight. If A can't assemble the meaning it creates new rules when necessary and constructs the utterance using the new rules. A then shares this utterance with B. B then tries to reconstruct the meaning A has just shared. B also may receive some *Cues* from the outside environment which may or may not reflect the intended meaning. From the combination of language rules and cues B then tries to reconstruct a meaning and if B is confident enough of this meaning

B sends positive feedback to A[1]. If positive feedback is sent both A and B will increase the weight of the language rules used and if positive feedback is not sent then both A and B will decrease the weight of the language rules used. This is schematically shown in figure 1.

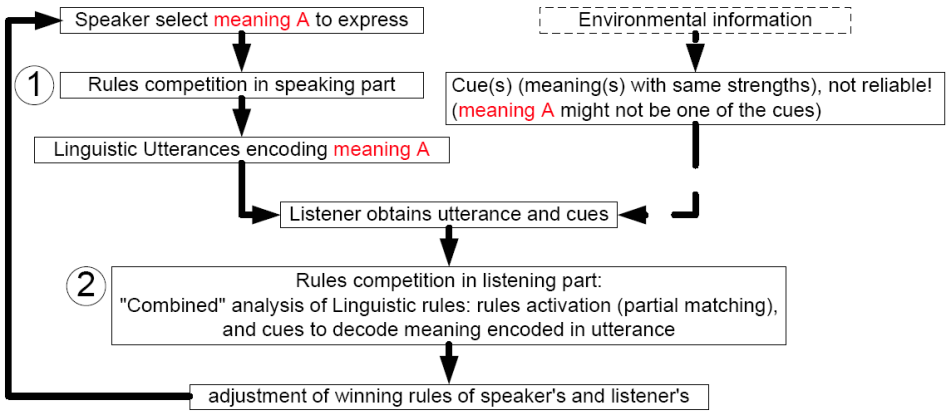

Figure 1: One round of communication

The language rules that the agents use come in three types: *holistic* and *compositional*. Holistic rules are a direct mapping between the total utterance and it's integrated meaning. The constituents of the utterance (i.e. individual words) may or may not be understood. Compositional rules are rules that are combined to form the integrated meaning. These include both knowing the individual words and knowing how to put them together. For example we could take

$$d(x) = \text{The dog ate } x$$

to be a rule that specifies if we have some object, to express that a furry animal ate it we say $d(object)$ = "The dog ate *object*". The last type of language rule is a word order rule. In this model only two and three concept sentences are used: Subject + Verb(SV) and Subject + Verb + Object(SVO). This is the order in English of course, but it is not in all languages. Word order rules tell us the order we say the constituents to form the integrated meaning. This is elementary grammar and it is why the sentences "You ate grapes." and "Grapes ate you." mean two very different things. In this model the word order rules for two concept sentences and three concept sentences evolve independently. Compositional rules are clearly more powerful for communicating. Compositional rules are the reason a person can understand a sentence that he or she has never heard before[5].

Performance

To track the progress of the emergent language we define two quantities: *Rule Expressivity*(RE) and *Understanding Rate*(UR). They are defined as

[1]Notice that B sends positive feedback to A even if agent B thinks the sentence means something different than A intended. Confusion actually aids in the symmetry breaking process and helps create the language because it can cause rules to gain weight and become used in the language.

$$\text{RE} = \frac{\sum_{\text{agents}} \#\text{ of meanings agent can express}}{\#\text{ agents}}$$

$$\text{UR} = \frac{\sum_{i,j} \#\text{ of understandable meanings between } i \text{ and } j}{\#\text{ of all possible pairs } i \text{ and } j}$$

There is a one-to-one correspondence between holistic rules and meaning but a one-to-many correspondence between compositional rules and meaning. RE therefore will increase more dramatically with compositional rules than holistic rules. RE measures how well the language is developing on an individual basis. UR[2] measures how uniform the language is becoming. Many agents may have fully developed vocabulary but not be able to communicate with each other, UR measures how the individual vocabularies are condensing into one.

Results

The results shown are those described in more detail in reference [3]. The population size is ten. Within each round of communication all the agents are paired up and one communicates several times with the other. There are twelve meaning constituents which can be built up to forty-eight integrated meanings[3]. At the outset all agents share six holistic rules which contain all twelve meaning constituents and have no dominant word order. The reliability of outside cues is 0.7 and the probabilities for creating a new rule or generalizing previous communications into a needed rule are equal: 50%.

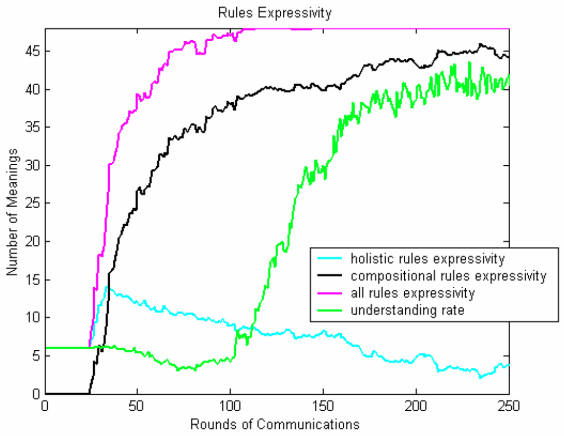

Figure 2: Evolution of RE & UR

[2]UR is measured in the absence of external cues. It is thus a true measure of the language, not the language + environment combination that contributes to learning the language.

[3]This model does not deal with the interesting case where meanings can be acquired from language. This is of course removes many interesting questions, but it is not an unreasonable assumption when modeling the development of a language from a series of individuals. Presumably the meanings preexist the words for them in the mind of each individual.

Figure 2 shows the results of RE and UR for one run of the simulation. At the beginning of the simulation each agent can express six meanings with 6 holistic rules. There are no compositional rules. As the simulation progresses each agent begins to create rules, both holistic and compositional(round 25). As more communications take place eventually random patterns repeat and get generalized into compositional rules. A transition occurs then between a holistic language and a compositional language(round 40). The number of compositional rules grows and the number of holistic rules dwindles. The word order then becomes important. This was not a factor before, so it causes the understanding rate to drop somewhat(round 50). The self-organizing mechanisms of rule competition and rule adjustment then cause certain word order rules to become shared among all the agents. When this happens the understanding grows dramatically and a consistent language emerges(rounds 100-150).

It is interesting to note that initially the language is a very limited holistic language and that self-organization causes a phase transition to a compositional language. What drives this change? In part is is because each compositional rule can be used in many integrated meanings. A rule with a high weight from one meaning has the same weight in another meaning, so many high-weight compositional rules can easily outweigh one holistic rule. Additionally cues provided by the environment will boost compositional rules; if even a constituent of a cue is shared by the inferred meaning this increases the chance of that compositional rule being affirmed. A holistic rule must receive exactly the same cue for there to be an increased chance of success.

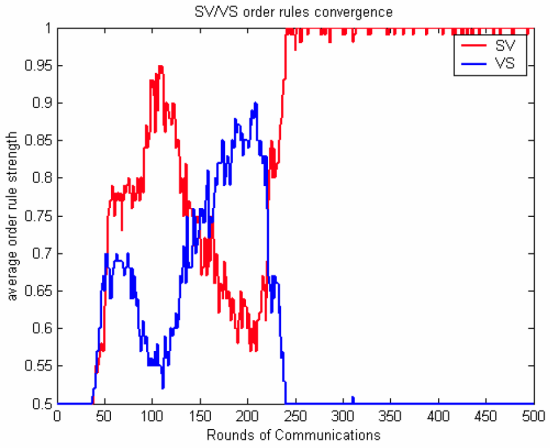

Figure 3: Evolution of two concept sentences

Figure 3 shows the evolution of the two-word order rule in the simulation. Note that the competition between word-orders doesn't begin until about round 40 when compositional rules start to be created.

Figure 4 shows the evolution of the three-word order rule in the simulation. Again the competition doesn't begin until compositional rules are created, but this rule reaches the ordered state more quickly. This ordered state is what drives the UR up much faster starting shortly after round 100. Note the process here; first the vocabulary is spread between each agent via the self-organizing methods discussed above. Once a common

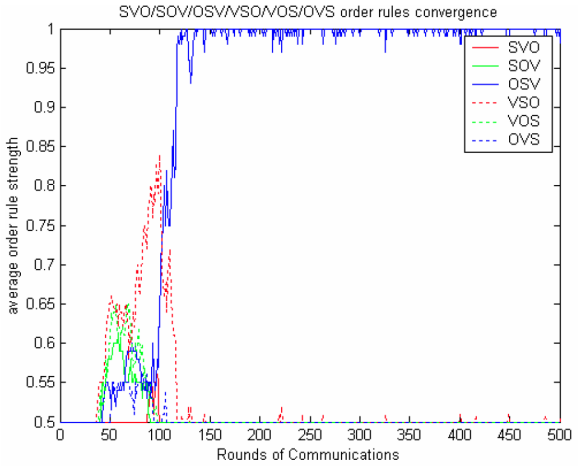

Figure 4: Evolution of three concept sentences

vocabulary emerges, the grammar also emerges via the same self-organizing methods. The emergence of the vocabulary is necessary for the grammar to emerge. Once the grammar has spread between all the agents it can then be successfully used to communicate.

The Stability of a Language

Now that we have seen that language can be emergent there are many other interesting questions we can ask. How stable is this language? What if we put a flux of agents into the system? I now turn to models that address these questions.

The Stability of Grammar

The previous model assumed that the evolution of the word order rules was independent. This is a rather unrealistic assumption though, it is hard to believe that any language would have the order SV for two concept sentences but VOS for three concept sentences. A paper by Minett, Gong, and Wang[6] addresses this issue by considering the word-order competitions for a language that already has a developed vocabulary.

We can construct grammar by treating the words as having local interactions: when dealing with two concept sentences there is only one competition of word orders: SV/VS. When dealing with three concept sentences there are three: SV/VS, VO/OV, and SO/OS. Suppose that the first competition settles as VS. There are three languages that are consistent with this state: OVS, VOS, and VSO. Of these three the first two are of the form OS and the last is SO. If OS wins the competition the language is likely to fluctuate between the two languages that satisfy this requirement. The two grammars are called *imprecise*. If SO wins the competition the language is uniquely specified (*precise*), but it is unstable; probability dictates that chance will pull up more OS combinations than SO combinations so the language will tend to evolve into one of the other two that satisfy OS. Generically what tends to happen is rules that specify a unique syntax tend to give way to rules that generate an imprecise syntax[6].

Running the Model

This model has 10 agents which have a pre-specified lexicon of 4 nouns, 4 single noun verbs and 4 double-noun verbs. The grammar is specified by two local syntax rules. In the first experiment agents had an equal probability of speaking two and three-concept sentences. They were initialized with the precise local syntax SV+VO. The simulation was run for 5000 communication rounds. A typical run of data is presented in figures 5-7.

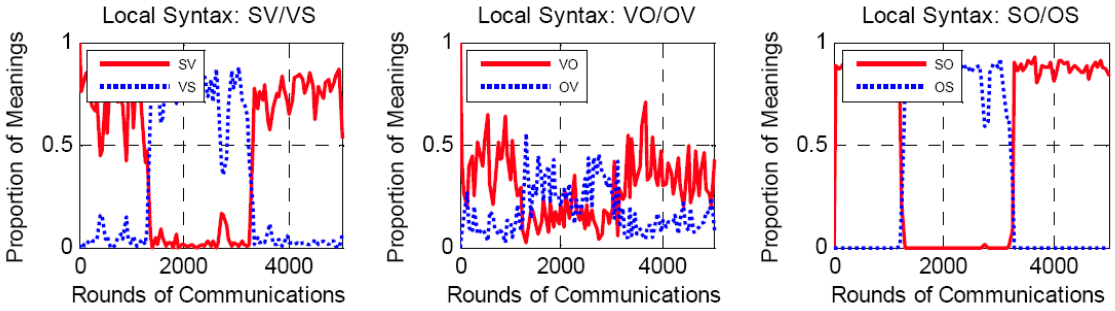

Figure 5: Local syntax competitions

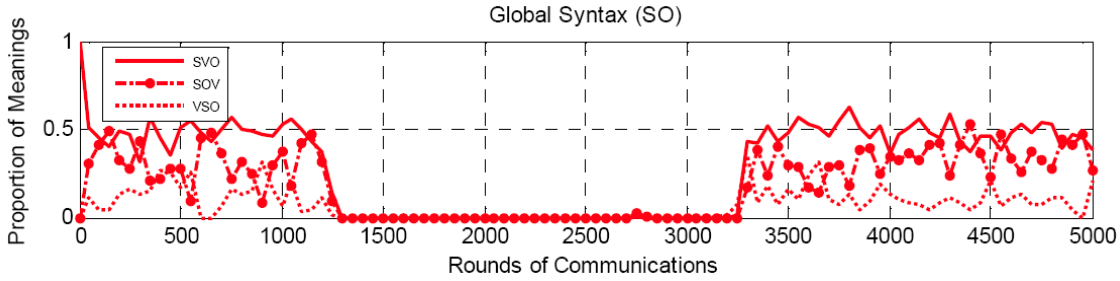

Figure 6: Time evolution of the global syntax with SO ordering

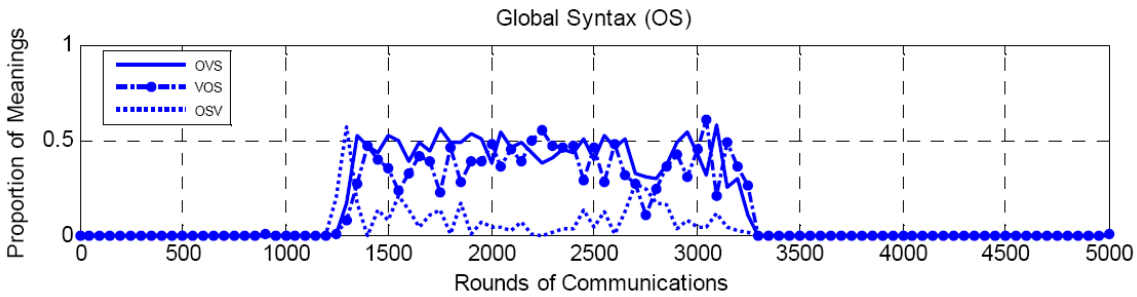

Figure 7: Time evolution of the global syntax with OS ordering

The first thing that happens is the unique global syntax SVO quickly drops from the universal syntax and SOV begins to compete with it. The local syntax is SO + SV and

the two imprecise syntaxes compete for the global syntax. This competition continues until shortly after round 1000 when the the VO/OV competition shifts to favor OV. OVS makes a brief appearance in the language, but the syntax then quickly shifts to a competition between OVS & VOS - the local syntax has shifted to VS + OS. The local syntax has completely inverted itself. This competition then unfolds for about 2000 more rounds and then then reorganizes back to the competition between SO + SV.

Notice that the SO/OS competition is much more stable than the other two local syntax rules. This is because the other two competitions are between a noun and a verb. The SO/OS competition is between two nouns and so it causes the most confusion when it is not decided. For example if I know that the subject comes first I can decode

I Nigel emailed

meaning I sent an email to Nigel because I know the subject of the sentence comes before the object or verb and I know that Nigel is a noun emailed is a verb. However if all I know that the verb comes first I can't decode the sentence. There is no way to know if

Emailed I Nigel

means Nigel sent me an email or I sent him one. The other two local syntax competitions may sound awkward but they don't hinder the communication; the SO/OS competition does. Thus fluctuations in the SO/OS competition are highly penalized and this combination will tend to be stable.

The Real World

In 1963 Greenberg[4] observed that of the existing languages most were SVO, SOV or VSO. It is interesting to notice that these are all SO languages. This model predicts that if a population has a secondary population that branches off such that both languages evolve independently the two languages are most likely to keep the same SO/OS order but the other pairs are much more likely to change. This model suggests it is not unreasonable that many of the currently spoken languages had a common ancestor.

Consider for example that Portuguese, Spanish, French, Italian and Romanian all descended from Latin. Latin has the order SOV, whereas all the romance languages[4] have order SVO, except when a pronoun is used as the object; then the order is SOV[1]. The languages seem to have kept a remnant of their ancestor language but fluctuated to the other imprecise grammar. This fits nicely in the framework of this model.

The Stability of Vocabulary

Another question we can ask is how does a flux of agents affect the development of a language? Clearly we would expect that introducing new agents into the system will change the language at least initially, but will the language emerge the same or will it be

[4]Languages derived from Latin are called *Romance Languages*.

altered dramatically? A paper by Bodìk and Takàč[2] asks this question. In this paper they model language emergence by putting agents on a map and having them talk about objects. Agents are replaced at a given *Flux Rate*[5]. The idea is to then look at how the flux rate changes the development of the language.

In their model Bodìk and Takàč found that the changes in the language depend drastically on the flux of agents. A flux rate that is too high can destroy the language entirely, but languages with low flux rates are relatively stable. For example figure 8 shows the relative weight for all the words used for just one object over 11300 communication games. Over the course of 10000 rounds 35 words we invented for that object and only two survived to the end. These were both created in round 53500 and were replaced by new words shortly after this snapshot of the simulation.

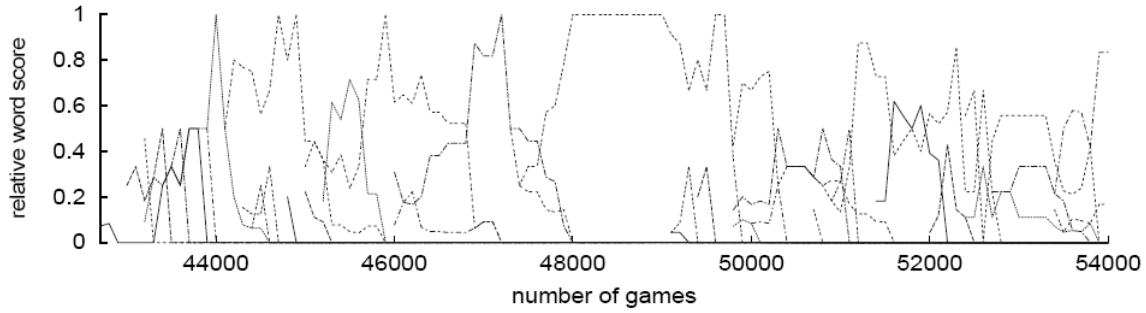

Figure 8: Relative word scores for one object with flux rate 50

Clearly this model is very turbulent and the language is neither stable nor useful for communication. Even though at many times a word score reached one (the vocabulary became uniform for that word) it was short lived and soon to be replaced. Contrast this with the flux rate of 2000 shown in figure 9.

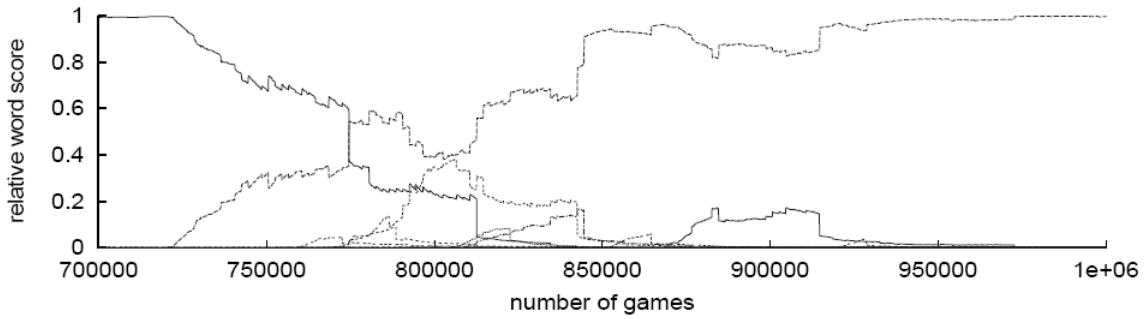

Figure 9: Relative word scores for one object with flux rate 2000

Of 5 million games played with the flux rate of 2000 words only changed 13 times. Figure 9 shows one of these word changes. Notice the change of word took about 250000 games or 125 agents were replaced before the word had changed. The entire population was switched out 12 times before the new word became uniformly spread to all the agents!

[5]A flux rate of 50 means one agent is replaced by a new agent every 50 communication games.

Also notice that as a word becomes less uniformly used more words are created for it. This is because agents aren't consistently being told the same word for an object and it is thus harder for existing words to gain weight in the lexical rulebook. The reinforcement mechanisms break down when the system isn't in the ordered state, making it more chaotic.

The stability of the language depends largely on how quickly agents are removed from the system. In their paper Bodìk and Takàč show that a flux rate of 500 permanently lowers the success rate to about 0.6. The language is never able to reach a uniform state, but it is not wholly destroyed either. Instead it is in a state where it is constantly evolving, never able to reach the ordered state.

Conclusions

The origins of language is a subject of much controversy among linguists. There are many who believe[6] that language can be a wholly emergent phenomenon and that as long as the intelligence to recognize patterns and associate sounds with ideas exists a language needs no further biological background. The aim of this paper has been to show that this is at least plausible. Language has been shown to emerge from a very simple model of agents attempting to communicate with each other. The desire to successfully communicate creates reinforcement mechanisms that drive the system into an ordered state. This ordered state has specific properties that we observe in the real world: the grammar tends to fluctuate between two imprecise grammars and the lexicon is mostly stable provided the flux of agents is low enough although it does slowly evolve.

Although the models so far seem to indicate the right trends, there are many more interesting questions to ask. The dynamics of two developed languages interacting with each other is very common among people groups and has not yet been addressed. Abstraction as well has not been dealt with. This is certainly one of the most important features of language. Additionally the learning of new semantic meanings often occurs through language, this has not been considered. Computational modeling of language is a rather new field and we will have to wait.

References

[1] Peter Bodik and Martin Takac. Formation of a common spatial lexicon and its change in a community of moving agents. In *Frontiers in AI: Proceedings of SCAI'03*. IOS Press, 2003.

[2] Tao Gong and William S-Y Wang. Computational modeling on language emergence: A coevolution model of lexicon, syntax and social structure. *Language and Linguistics 6.1*, 2005:1–42, 2005.

[3] J. H. Greenberg. *Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements*. Cambridge, MA.: MIT Press, 1963.

[6] this author among them

[4] T. M. V. Janssen. *Compositionality*. Handbook of Logic and Language, Elsevier, Amsterdam and MIT Press, Cambridge, J. Van Benthem & A. ter Meulen (Eds), 1997.

[5] James W. Minett, Tao Gong, and William S-Y. Wang. A language emergence model predicts word order bias. *Proceedings of the 6th evolution of language conference*, 2006.

[6] Wikipedia: Subject Object Verb. http://en.wikipedia.org/wiki/subject_object_verb.