THE EMERGENCE OF COMPOSITIONALITY, HIERARCHY AND RECURSION IN PEER-TO-PEER INTERACTIONS

JOACHIM DE BEULE

Artificial Intelligence Lab, Free University of Brussels,
Brussels, 1050, Belgium
joachim@arti.vub.ac.be

It is argued that compositionality, hierarchy and recursion, generally acknowledged to be universal features of human languages, can be explained as being emergent properties of the complex dynamics governing the establishment and evolution of a language in a population of language users, mainly on an intra-generational time scale, rather than being the result of a genetic selection process leading to a specialized language faculty that imposes those features upon language or than being mainly a cross-generational cultural phenomenon. This claim is supported with results from a computational language game experiment in which a number of autonomous software agents bootstrap a common compositional and recursive language.

1. Introduction

All human languages are compositional, hierarchical and recursive. These features, by combining words into hierarchical phrases which can then recursively be combined into larger phrases, allow the construction of an unlimited number of sentences using only a limited number of words. With only the words "red" and "ball", all of the phrases "red ball", "red red ball", "red red red ball" etc. can be formed. Compositionality refers here to the fact that the phrase "red ball" can be decomposed into two constituents "red" and "ball". Hierarchy refers to the fact that the resulting combination "red ball" is a genuine constituent itself, ready to be used as a single unit by other constructions. Recursion then refers to the fact that the rules for combining the units "red" and "ball" into a new unit can recursively be applied to the units "red" and "red ball" etc.

There have been three major hypothesis about how all languages came to share these features: nativism, iterated learning and cultural selectionism. The nativists claim that humans are endowed with a genetically encoded language faculty dictating us to develop recursive language (see e.g. (Hauser, Chomsky, & Fitch, 2002).) In the iterated learning model, compositionality emerges after many generations as the result of an iterated learning process in the presence of a learning bottleneck and when language learners are equipped with a generalizing learning mechanism (see e.g. Smith, Brighton, and Kirby (2003).)

In a previous paper (De Beule & Bergen, 2006) it was already argued that compositionality can also emerge in the course of only one generation when language is viewed as a complex adaptive system in which language users may refuse, adopt or propose new elements of language to improve their communicative skills which in turn depends on the skills of other agents *and on whether they collectively succeed in reaching a consensus*. Because compositional constructs can be used more frequently than holistic ones, communicative success improves when compositional constructs are preferred.

In this paper, we go one step further. We show that not only compositional but also recursive language emerges when agents are driven to use those language constructs that optimize their (estimated) communicative success, or, in other words, the (estimated) probability of conveying a message.

2. Experimental Setup

To support our claim we have carried out a language game experiment in which a population of social agents needs to bootstrap a communication system without there being any central control. At each time step, two agents are randomly selected from the population and are presented with a scene which one of the agents has to describe to the other. The second agent then signals whether he agrees with the description.

The experiment was carried out using the framework of Fluid Construction Grammar (FCG). This is especially being developed for doing experiments in the emergence and evolution of grammatical language. It is in line with the ideas advocated in usage-based cognitive approaches to language in general and construction grammar in particular, and it is designed to support highly flexible language processing in which conventions are not static and fixed but fluid and emergent.[a]

2.1. *The World Model*

We assume that all agents collectively are aware of 5 event types $(E_i\ ?e\ ?a\ ?p)$ and 5 possible event participants $(P_i\ ?o)$ which can each possess one or more of 5 features $(F_i\ ?f)$ (with $i = 1..5$.)[b] The scenes $S(e, a, p)$ about which the agents have to communicate during an interaction will consist of an event e with agent a and patient p. Their descriptions are generated according to the following prescription (with $i = 1, ..., 5$): (1) $S(e, a, p) \rightarrow A(a) \wedge A(p) \wedge E(e, a, p)$, (2)

[a]At this point the FCG system is ready for use by others, an implementation on a LISP substrate has been released for free download through http://arti.vub.ac.be/FCG/ and http://www.emergent-languages.org. We refer to these sites and Steels, De Beule, and Neubauer (2005) and De Beule and Steels (2005) for further information, the first paper explains how linking is done in FCG and the second shows how hierarchy and recursion are handled.

[b]Predicates are represented using prefix-notation, symbol-names starting with a question mark are variables.

$E(e,a,p) \rightarrow (E_i\ e\ a\ p)$, each with probability $1/5$, (3) $A(x) \rightarrow P(x)$ or $F(x) \wedge A(x)$, each with probability $1/2$, (4) $P(x) \rightarrow (P_i\ x)$, each with probability $1/5$ and (5) $F(x) \rightarrow (F_i\ x)$, each with probability $1/5$.

Agents do not always describe the entire scene to one another. Instead, the set of possible topic descriptions given a scene contains all of the $A(a)$'s, $A(p)$'s and $E(e,a,p)$'s present in the scene. It is easily verified that a scene can in principle contain an arbitrary number of predicates and that on average a scene and a topic description contain 5 and 2.75 predicates respectively. Every interaction a random scene and associated topic description are generated and presented to the speaker. The hearer is only presented with the scene, not the topic.

The variable equalities contained in a topic description always need to be expressed. This can be done using a holistic word or else with a grammatical construction (imposing an order on the words used, see e.g. Steels (2005).)

2.2. *The Language Model*

The meaning of a word may be any combination of predicates, possibly containing any number of equalities between the variables involved. Lexicon lookup results in a set of lexical constructions that are estimated to maximize the number of predicates being conveyed successfully. Word-Meaning probability scores and synonymy scores as defined in (De Beule, De Vylder, & Belpaeme, 2006) are used for this. In short, all agents associate probability and synonymy scores with word/meaning mappings. The probability score represents an estimate of the correctness of a mapping, and hence of the probability that other agents will understand the word as intended. This is needed because when a hearer adopts a new word, he may not be able to pin down the meaning of it without knowing the topic. As a result, and also because many different agents propose different words for the same meaning, after a while agents will know several words that all might be used to express the same meaning. Agents will of course prefer those words with maximum associated probability score, but after a while many such words will be known. Synonymy scores are then used to impose a further preference among them, they reflect the agent's estimate of the preferences of the other agents and are updated according to the well known lateral-inhibition mechanism.[c] The probability score of every word used is updated by the hearer after every interaction by combining the information provided in the speaker's utterance, the scene description and the knowledge that the topic is part of the scene (see De Beule et al. (2006) for details.)

The scores of different words can be multiplied to estimate the probability that

[c]In most previous experiments only one score was used, mostly resembling the synonymy score. The difficult nature of the problem with which the agents are faced in the current experiment (e.g. uncertainty of the hearer about the meaning of words or even of word boundaries, see later) forced us to split this score and use more complicated updating rules.

their combined meanings will be understood correctly. If one holistic word with high score covers the entire topic description then it might be used by the speaker. If several more atomic words that only together cover the entire topic description have a bigger combined score however, they will be preferred. Hybrid combinations containing both holistic and atomic words are also possible. If after the analysis some predicates contained in the topic description are left uncovered then the speaker simply proposes one new word with maximal initial scores covering all of them. Clearly, such a word can be holistic.

After this all predicates in the meaning are covered by a word. However, because of the way topic descriptions are generated, if multiple words are used then not all of the variable equalities will be covered and grammatical (linking-) constructions need to be applied (Steels et al., 2005; Steels, 2005). Such a construction maps a semantic pattern (basically a set of variable equalities between parts of the meaning covered by different words) to a syntactic pattern (word order) and vice versa. So assume that the topic is described by

$$(P_1\ ?x) \wedge (F_1\ ?x) \wedge (F_2\ ?x) \tag{1}$$

and that the speaker opted for using three words, one for every predicate in the description. Then he needs a construction imposing one of the $3! = 6$ possible word orderings to express the variable equalities. We assume a pre-defined one-to-one mapping between predicate types and semantic and syntactic categories as follows: event↔verb, participant↔NP, feature↔Adjective.[d] Furthermore, the overall type and hence syntactic category of a conjunction of predicates is determined as the smallest reduction given that every $A(x)$ is also of type participant (and hence NP) and every $S(e, a, p)$ is of type scene (syntactic category sentence.) So all of these 6 constructions can be described as rewrite rules of the form 'NP → Adjective NP Adjective' etc. All of them are recursive in the sense that after application the resulting construct could be used as a constituent of the same construction. They are however not minimal as defined shortly. Different agents might propose different constituent orderings. Therefore, agents also associate probability scores with every possible order. These reflect the estimated probability that other agents prefer the specific ordering.

If on an other occasion the meaning would only contain one feature and one participant predicate then the constructions should be of the form 'NP → Adjective NP' or 'NP → NP Adjective'. These are called minimally recursive. After that, this construction could also be recursively applied twice for expressing the longer meaning (1). It is easy to see that this can lead to conflicting analysis (i.e. conflicting word orderings.) Therefore, agents also associate preference

[d]This or a similar assumption has been made in almost all computational experiments involving the evolution of grammar because the nature and origins of semantic and syntactic categories and their relations are not yet very well understood (but see e.g. (Steels, 2004) or Steels (2002)) and because the current paper focuses on other aspects of language emergence.

scores with grammatical constructions, comparable to the synonmy scores of lexical constructions. The set of constructions to use for producing an utterance is determined in a similar way as how the set of words to use is determined: the probability and preference scores of all the constructions in a possible set are multiplied and that set with the biggest resulting combined score is selected. If the selected set leaves some of the equalities uncovered then a new construction (with random constituent order) is proposed covering all of them at once. Probability and preference scores are updated in the same way as in the lexical case.

As a third example note that a speaker might also opt for expressing the meaning (1) with one or two words. For example, one word might cover the participant predicate and another the two feature predicates. The latter word doesn't have a simple syntactic category (it is the combination of two adjectives which cannot be reduced to anything smaller, i.e. a single syntactic category.) Hence the required construction will not be recursive.

What is important is that the constituent structure and hence the result types (categories) of grammatical constructions (e.g. the number of adjectives they select for) is open ended. Agents propose new constructions on the fly as they need them and, depending on the specific set of words that need to be combined, no, minimally recursive, non-minimally recursive or non-recursive constructions will be proposed. But as the agents will be able to use the minimally recursive constructions more often, and since this allows the agents to reach a consensus about them more quickly, it will pay to use them. In the next section we will show that after some time agents indeed start preferring these over less re-usable constructions (i.e.constructions selecting for more constituents or non-recursive ones.)

Finally, but very important, a hearer is not aware of intended word-boundaries. So even if a speaker produces a phrase composed of several words and constructions, the hearer may still think it is a single word. He too combines his own probability and preference scores again to determine that decomposition (of the utterance into words and constructions) that is estimated to have the biggest combined score. So both producing and parsing an utterance make use of the probability estimates to determine the combination of words and constructions that together optimize the probability of having a (partially) successful communication and nothing else.

To sum up, care was taken that several solutions to the problem of successful communication are within reach to the agents, ranging from pure holistic lexical over hybrid compositional to full and minimally recursive encoding. The latter two solutions require that a consensus is reached about how to cut up meaning space in a compositional fashion and, on top of that, a set of grammatical constructions. On the other hand they require less words, and only a minimally recursive grammar requires no more than 2 grammatical constructions ('S→NP V NP' and 'NP→Adjective NP', or with other constituent orderings.) At all times, the only thing driving the agents towards one of the solutions is the estimated probability of having a successful communication.

3. Results

Figure 1 shows the evolution of the communicative success for different population sizes measured as a running average. A value of -1 corresponds to a failed game, 0 means that the hearer needed to learn something and +1 represents a successful game. From this graph it can be concluded that the agents in any case do

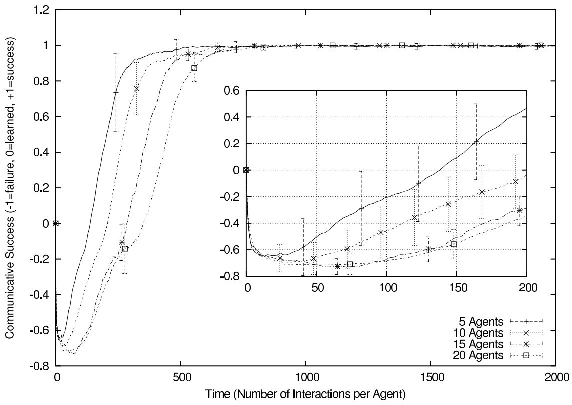

Figure 1. Evolution of the communicative success for different population sizes. (all graphs averaged over 10 independent runs with error bars 1 standard deviation wide.) Time was rescaled such that at time step t an agent has had on average $n_a t$ interactions with n_a the population size.

succeed in evolving a successful communication system.

Figure 2 shows the evolution of the number of predicates in a topic description divided by the number of words in the utterance, also measured as a running average. From this we can conclude that after about 100 interactions per agent no holistic words are used anymore (otherwise the curve would stay above 1.)

The top of figure 3 shows when constructions are used (curve 'Grammatical Construction Use') and also when constructions of different result types are used given that a construction is used. Together with what can be calculated about the structure of the world and of topics, these curves allow to conclude that after about 800 interactions per agent (for 15 agents, similar curves were obtained for different population sizes) only constructions of resulting type S(entence) and NP are used.

Finally, the bottom figure in 3 shows the number of constituents for the different types of constructions used. As the O(ther)-type of constructions are not used this curve can be ignored (curves are only updated when the constructions involved are used.) What is important is that the S(entence) and NP-type curves

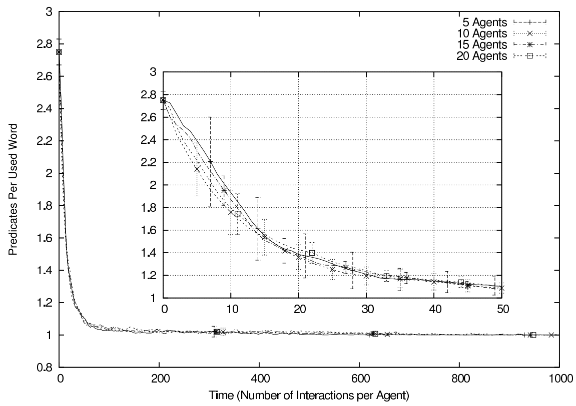

Figure 2. The number of predicates in the topic description divided by the number of words in the utterance. For a completely holistic language this would be 2.75 (the average number of predicates in a topic description.) A fully compositional language would give 1, which is the value to which all graphs converge.

converge to about 3 and 2 constituents respectively, meaning that the agents converge to the minimal recursive solution.

To conclude then, the experiment confirms the claim made: that the agents succeed in evolving a successful language that is purely compositional and almost completely minimally recursive. Further investigation shows that a minority of the agents continues to use holistic constructions. However, these are to the outside not distinguishable from the constructions needed in the minimal recursive solution used by almost all of the agents almost all of the time. Hence, we have shown that the languages to which the populations invariably converge are compositional, hierarchical and recursive, event hough there is no population turnover nor is there anything built into the agents that forces them to evolve such languages except for their drive towards successful communication.

References

De Beule, J., & Bergen, B. K. (2006). On the emergence of compositionality. In *Proceedings of the 6th international conference on the evolution of language* (p. 35-42).

De Beule, J., De Vylder, B., & Belpaeme, T. (2006). A cross-situational learning algorithm for damping homonymy in the guessing game. In L. M. R. et al. (Ed.), *Artificial life x* (p. 466-472). MIT Press.

De Beule, J., & Steels, L. (2005). Hierarchy in fluid construction grammar. In U. Furbach (Ed.), *Proceedings of ki-2005* (Vol. 3698, p. 1-15). Berlin: Springer-Verlag.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*, 1569-1579.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: the cultural emergence of compositional structure. *Artificial Life*, *9*(4), 371-386.

Steels, L. (2002). Simulating the evolution of a grammar for case. In *Presented at the 4th international conference on the evolution of language*. Harvard University.

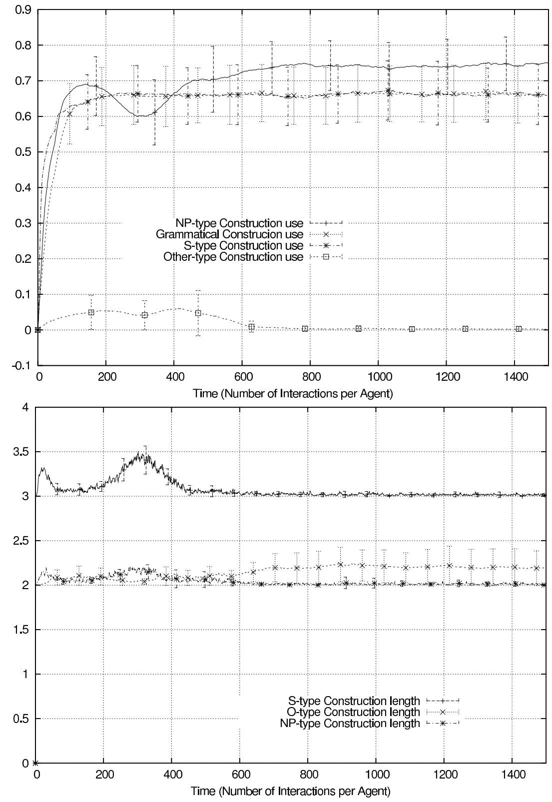

Figure 3. Usage (top) and number of constituents (bottom) of grammatical constructions used for different constructional result types in the case of 15 Agents (see text for further explanation.)

Steels, L. (2004). Constructivist development of grounded construction grammars. In *Proceedings annual meeting association for computational linguistics conference*. Barcelona.

Steels, L. (2005). What triggers the emergence of grammar? In *Aisb'05: Proceedings of the second international symposium on the emergence and evolution of linguistic communication (eelc'05)* (p. 143-150).

Steels, L., De Beule, J., & Neubauer, N. (2005). Linking in fluid construction grammar. In R. F. A. for Science & Art (Eds.), *Proceedings of bnaic-05*.