COMPUTATIONAL SIMULATION ON THE COEVOLUTION OF COMPOSITIONALITY AND REGULARITY

TAO GONG, JAMES W. MINETT, WILLIAM S-Y. WANG

Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, China

Compositionality and regularity are universals in human languages; in most languages, complex expressions are determined by their structures and their components' meanings. Based on a multi-agent computational model, the coevolution of compositionality and one type of regularity, word order, is traced during the emergence of compositional language out of holistic signals. The model modifies some questionable aspects in the Iterated Learning Model and Fluid Construction Grammar by considering the conventionalization in horizontal transmission and the gradual formation of syntactic categories which mirror the semantic categories. The model also implements a bottom-up syntactic developmental process, i.e., the global orders for regulating multiple arguments are gradually formed from simple local orders between two categories.

1. Introduction

There are two mainstream views on language emergence (Minett & Wang 2005). *Innatism* (e.g., Anderson & Lightfoot 2002) ascribes linguistic universals to certain innate mechanisms unique to humans (e.g., *Language Acquisition Device*). However, recent evidence from anatomical, biological and comparative studies of the human communication system and those of animals (Oller & Griebel 2004) suggests that language might have evolved from domain-general abilities, and these studies support *Emergentism* (e.g., Knight et al. 2000). Emergentism also gains support from many behavior-based computational models which have discussed the effects of some communicational constrains and certain domain-general competences of language user(s) on the formation of linguistic universals. For instance, Kirby's *Iterated Learning Model* (ILM) (2003) implied that the *bottleneck effect* (the restricted exposure of the previous generation's language to the next generation) in vertical transmission is sufficient to trigger a compositional language; Luc Steels' *Fluid Construction Grammar* (FCG) (2004) developed an artificial language with simple argument constructions through iterated description games between two agents. Despite the significant insights they provided, these models neglected some important processes that might affect language emergence, and built in some questionable assumptions.

In ILM, it is insufficient to claim that compositionality has emerged based on the limited facts that agents have acquired compositional rules and that these rules have high expressivity. Using compositional materials to exchange *integrated meaning* (which describes a complete event containing an action (*Predicate* (Pr)), its instigator (*Agent* (Ag)) and sometimes the entity undergoing such action (*Patient* (Pt))) also requires regulating these materials under similar mechanisms (e.g., syntactic or morphological operations); without regularity, the sharing of common compositional materials is useless. For example, without syntactic information, it is unclear "who are chasing whom" when hearing "dogs chase cats" or "cats chase dogs". The emergence of regularity occurs in the production and comprehension of integrated meanings based on linguistic and nonlinguistic information available during the communications. Due to the assumption of *direct meaning transference*, in which the speaker's intended meaning is always transparent to the listener, ILM has a simplified comprehension process, which makes it unsuitable for studying how agents develop regularity during the comprehension based on linguistic information. The inseparability of the lexicon and grammar has already been discussed in some empirical studies on language acquisition (e.g., Bates & Goodman 2001). Besides, cultural evolution covers both vertical and horizontal transmission. *Conventionalization* in horizontal transmission (during which an individual's language conforms acceptably to the language of the community) is important for shaping the language in one generation. If certain mechanisms during this conventionalization already incline toward compositional materials, the bottleneck effect shown in ILM might be weakened.

FCG, similar to the Construction Grammar theory (Fried & Östman 2004), assumed that syntactic structures mirror semantic categories, but it did not demonstrate how such matching was established. FCG also built in a preference for linguistic information — reference to visual events occurs after the linguistic comprehension fails. How this preference is built needs further discussion.

Regarding these limitations, based on our previous work in Evolang5 (Gong & Wang 2005, Gong et al. 2005), we present a multi-agent model to study the simultaneous emergence of compositionality and one type of regularity (word order) via horizontal transmission. This model adopts an indirect meaning transference, in which the comprehension is based on both linguistic and nonlinguistic information. Based on simple imitation ability, compositional materials (in the form of rules) emerge when agents identify recurrent patterns in the *Meaning-Utterance mappings* (M-U mappings) acquired during previous communications. Considering that some empirical study has shown that some primates can notice and manipulate simple orders (Hauser 1996), this model studies whether such simple sequential ability can lead to the complex syntax in human languages by adopting a categorization mechanism. Through categorization, rules which have the same semantic roles and which are used similarly in forming utterances are grouped into a category. The relative word orders of elements belonging to different categories are acquired as syntactic rules. Similar to the Verb Island hypothesis (Tomasello 2003), this categorization process traces how the semantics-syntax match becomes established. Meanwhile, complex orders that regulate elements from multiple categories can gradually form based on the simpler word orders between pairs of categories. This bottom-up syntactic development is consistent with the "tinkerer" view of evolution (Jacob 1977), which states that complex features can gradually develop out of simple available abilities. It also matches the "carpentry" theory of language evolution (O'Grady 2005), which points out that sentences can be resolved by simple processing forces driven by efficiency. In addition, similar developmental processes have already been traced in children's language acquisition, e.g., children progress from a "two-word" stage to a "multi-word" stage in the acquisition of lexical items and simple sentences (Clark 2003).

2. Model Description

2.1. *Linguistic Rules and Categories*

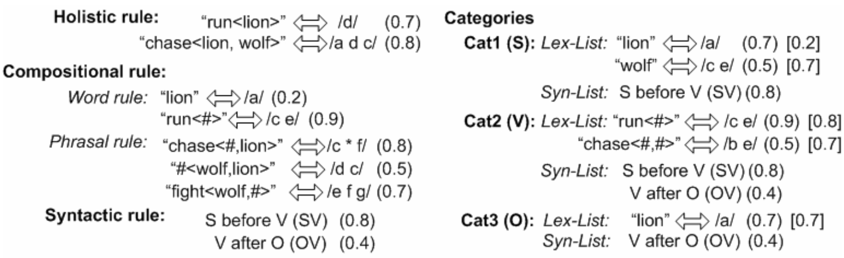

Figure 1. Linguistic rules and categories. "#" can be replaced by other semantic items, and "*" by other syllable(s). Numbers enclosed by ( ) denote the rule strengths, and those by [ ] denote the association weights of rules to their associated categories.

M-U mappings are used to represent language. The semantic space consists of a set of integrated meanings, each having an occurrence frequency. These integrated meanings are of two types: Type-I: "Pr_1<Ag>", e.g., "run<lion>"; and Type-II: "Pr_2<Ag, Pt>" e.g., "chase<wolf, lion>". *Utterances* (combinable syllables chosen from a signal space) map to either integrated meanings or some semantic item(s) (e.g., “lion” or “run<#>”). Different semantic roles (“Ag”, “Pt” and “$\text{Pr}_{1/2}$”) correspond to different syntactic roles (“Subject” (S), “Object” (O) and “Verb” (V)) in utterances.

Agents use linguistic rules to produce utterances encoding integrated meanings and comprehend these meanings from utterances. Linguistic rules (see Fig. 1) include *lexical rules* (M-U mappings plus strengths) and *syntactic rules* (local word orders plus strengths). Lexical rules are either *holistic*, encoding an entire integrated meaning, or *compositional*, encoding part of an integrated meaning; the latter can be further divided into *word* and *phrasal rules* (see Fig. 1). Every syntactic rule uses one of four simple orders—*before*, *after*, *middle* or *surround*—to regulate how lexical members from two categories with different syntactic roles may be combined. The global order that regulates members from multiple categories is the combination of these local orders. Sometimes the global order is *precise*, e.g., SV+VO results in the unique global order SVO; sometimes, it is *imprecise*, e.g., SV+SO results in either SVO or SOV.

A category consists of a list of lexical rules (*Lex-List*) and a list of syntactic rules (*Syn-List*). *Categorization* is a process of acquiring linguistic knowledge so that the local orders of lexical rules can be regulated. Agents can build up a category to associate some lexical rules with some syntactic rules, if these lexicons have the same semantic role and are found under the regulation of those syntactic rules’ local orders. In addition, agents can merge categories the having same syntactic roles if those categories share some lexical members that are regulated by a common syntactic rule. These categorization mechanisms are similar to those in the Verb Island hypothesis, but are not restricted to verbs.

Agents acquire linguistic rules and syntactic categories during communication. During production, similar to ILM, agents occasionally create holistic rule to express the whole integrated meaning when they have insufficient compositional rules to encode every semantic item in the chosen integrated meaning. If compositional rules are available but the associated syntactic rules are insufficient to regulate the utterance, agents occasionally create new syntactic rules in order to do so. In comprehension, based on the available M-U mappings acquired in previous communications (*previous experience*) stored in a *buffer*, agents can detect recurrent semantic item(s) and utterance syllable(s) that occur in two or more M-U mappings, and then map them as compositional rules that are stored in their *rule lists*. This detection of recurrent pattern is discussed in detail in Gong et al 2005. After that, based on the previous experience, agents may categorize some lexical rules into categories, or reorganize the available categories by merging some of them.

2.2. *Communication*

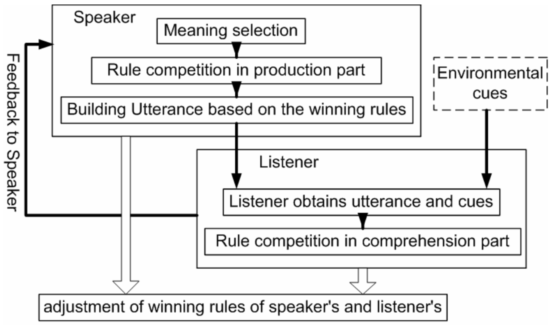

Figure 2. Information exchange in communications.

Each communication contains many rounds of information exchange (see Fig. 2). After choosing an integrated meaning based on its occurrence frequency, the speaker selects his winning linguistic rules based on their combined strength, builds up the utterance accordingly and sends the utterance to the listener. The listener receives the utterance and sometimes, some environmental cues. *Cues* are comprehension hints for the listener; the probability that one cue corresponds to the speaker's intended meaning is manipulated by the *Reliability of cues* (*RC*). Then, the listener selects the set of rules that allow her to comprehend an integrated meaning with the highest combined strength. The listener's calculation of the combined strength considers both linguistic rules and cues. Then, if the combined strength exceeds a certain *confidence threshold* (*CT*), the listener sends a positive feedback to the speaker and both of them reward their winning rules by increasing their strengths; otherwise, a negative feedback is sent and these rules are penalized.

There is no direct meaning check in the whole process; the strengths of linguistic rules that are rewarded sufficiently eventually exceed those of cues. Then, in order to comprehend utterances, agents will refer to these linguistic rules in comprehension, rather than the cues. Therefore, a preference for linguistic information is gradually established. At this stage, another feature of human language, *displacement* (Hockett 1960), is also established — agents are able to describe events not happening in their immediate space or time, and may still be understood. The model therefore simulates how linguistic communication turns into a reliable medium to exchange information, not just assistance to visual information to describe or discriminate simple concepts as shown in some models (e.g., Vogt 2005), though the assisting role may emerge earlier and might be a prerequisite for what we simulate here.

Syntactic categories play important roles for the emergence of global orders from local ones in communications, although categories themselves do not participate into the strength-based competition. In production, when choosing his candidate rules, the speaker first activates related compositional rules which can encode the chosen integrated meaning's semantic items. Then, considering these rule's syntactic categories, she activates some syntactic rules, the combination of which can resolve some global orders to regulate these compositional rules' syllables. Then, she judges whether these lexical and syntactic rules can win the strength-based competition. Similarly in comprehension, when choosing her candidate rules, the listener first activates some of her lexical rules whose syllables partially match the heard utterance. Some local orders can be detected based on these syllables' locations in the heard utterance. If some of these orders match the syntactic rules of categories to which these lexical rules belong, both the categories and syntactic rules are activated. Then, according to the categories, the semantic roles of the lexical items that are comprehended can be specified, particularly is the distinction of "Ag" from "Pt". Meanwhile, if some cue contains a meaning that corresponds to the comprehended meaning, its strength can participate into the calculation of combined strength of these rules. Then, the listener judges whether this set of rules can win the strength-based competition based on its combined strength.

3. Results and discussions

A semantic space containing 16 Type-I and 48 Type-II integrated meanings (consist of 4 "Ag"="Pt", 4 "Pr_1", 4 "Pr_2", the type ratio is 1:3) is used, in which the probability of choosing a Type-I or Type-II meaning are equal (by setting the token ratio of the two types to 3:1). Rule strengths are constrained to the interval [0.0, 1.0], the initial rule strength is set to 0.5 and the update step of strength is 0.1. 10 agents randomly communicate in pairs, and there are 20 information exchanges in one communication. The total number of communication is 3,000. Each agent initially shares 8 holistic rules but has neither categories nor syntactic rules. The buffer size is 40, and the rule list size is 60. After a certain number of communications, the *Rule Expressivity* (*RE*), *No. of common rules*, *Global Orders' Understandability* (average percentage of meanings comprehended using different global orders) and *Understanding Rate* (*UR*, average percentage of meanings accurately understood using linguistic rules only) are calculated.

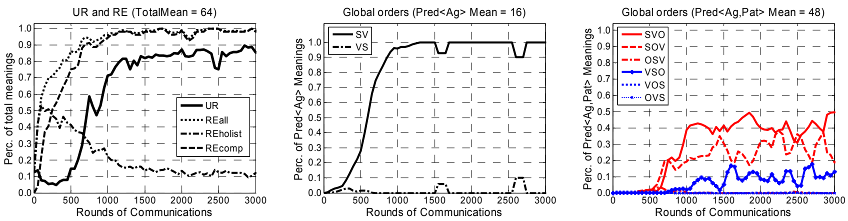

Figure 3. Coevolution of compositionality and regularity. RC=0.8, CT=0.75.

3.1. *Coevolution of compositionality and regularity*

The result of the simulation under the above parameters is shown in Fig. 3. In Fig. 3(a), the increase of the RE of the compositional rules is consistent with, but is not sufficient to prove, the transition from holistic language to compositional language. The understandability of the emergent language is traced by UR, which follows a U-shaped curve during the competition among holistic, compositional and syntactic rules and a sharp S-curve after some dominant orders are established and some lexical rules are shared. During the increase of UR, regularity emerges and some global orders become dominant (shown in Fig. 3(b)(c), agents use SV, SVO and SOV orders to comprehend most of the integrated meanings). The high value of UR (not RE) indicates the emergence of a compositional language which has both common compositional rules and consistent global orders. This emergent process occurs much later than the increase of RE, and indicates coevolution of compositionality and regularity, i.e., the acquisition of compositional rules and the formation of dominant orders boost each other, and are achieved simultaneously during language emergence.

4. Conclusions and future directions

This paper demonstrates the coevolutionary emergence of compositionality and regularity in a multi-agent system. A conventionalized language with common compositional materials and consistent syntax emerges out of holistic signals. Iterated communications in horizontal transmission provide opportunities for agents to get exposed to the utterances produced by others and to affect others with their own knowledge. Categorization shapes their categories to achieve a semantics-syntax match. This match can be imperfect, as still exists in modern languages, e.g., the Sex-Gender mismatch in German. The paper also implements a bottom-up syntactic developmental process in which complex syntax is formed based on local orders. How different sets of local orders affect the formation of global orders is discussed by Minett et al. in this volume and the linguistic or social factors that trigger the mismatch is under study.

Acknowledgements

This work is supported by grants from RGC Hong Kong: CUHK-1224/02H & CUHK-1127/04H. The authors thank Professors J. H. Holland, T. Lee, R. Chen and lab mates F. Wong & L. Shuai for useful discussions and suggestions.

References

Anderson, S. R., & Lightfoot, D. W. (2002). *The language organ*. Cambridge, MA: Cambridge University Press.

Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the lexicon: evidence from acquisition. In M. Tomasello & Bates, E. (Eds.), *Language development: the essential readings* (pp. 134–162). Oxford: Blackwell.

Clark, E. V. (2003). *First language acquisition*. Cambridge, MA: Cambridge University Press.

Fried, M., & Östman, J. (Eds.) (2004). *Constructional approaches to language*. Amsterdam, Philadelphia: John Benjamins Pub.

Gong, T., & Wang, W. S-Y. (2005). Computational modeling on language emergence: a coevolution model of lexicon, syntax and social structure. *Language and Linguistics*, *6(1)*, 1–41.

Gong, T., Minett, J. W., Ke, J-Y., Holland, J. H., & Wang, W. S-Y. (2005). A computational model of the coevolution of lexicon and syntax. *Complexity*, *10(6)*, 50–62.

Hauser, M. D. (1996). *The evolution of communication*. Cambridge, MA: MIT Press.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, *203*, 88–96.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196(4295)*, 1161–1166.

Knight, C., Studdert-Kennedy, M., & Hurford, J. R. (Eds.) (2000). *The evolutionary emergence of language: social function and the origins of linguistic form*. Oxford, UK: Oxford University Press.

Minett, J. W., & Wang, W. S-Y. (Eds.) (2005). *Language acquisition, change and emergence*. Hong Kong: City University of Hong Kong Press.

O'Grady, W. (2005). *Syntactic carpentry: an emergentist approach to syntax*. Mahwah, NJ.: Erlbaum.

Oller, D. K., & Griebel, U. (Eds.) (2004). *Evolution of communication systems: a comparative approach*. Cambridge, MA: MIT Press.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterative learning: a framework for the emergence of language. *Artificial Life*, *9(4)*, 371–386.

Steels, L. (2004). Constructivist development of grounded construction grammar. In D. Scott, W. Daelemans & Walker M. (Eds.), *Proceedings annual meeting association for computational linguistic conference* (pp.9–16).

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167(1–2)*, 206–242.