WHY HAS AMBIGUOUS SYNTAX EMERGED?

STEFAN HOEFLER
Language Evolution and Computation Research Unit, University of Edinburgh
40 George Square, Edinburgh EH8 9LL, Scotland, UK
stefan@ling.ed.ac.uk

Ambiguity is a defining property of natural language distinguishing it from artificial languages. It would seem to be dysfunctional, and therefore its ubiquity in language poses an evolutionary puzzle. This paper discusses the implications of a typical iterated learning model on the conditions under which syntactic ambiguity emerges and stabilises in language. It contrasts the purely nativist stance that language imperfections such as syntactic ambiguity are artifacts arising from internal constraints of the genetically determined language faculty with the view that they are frozen accidents persisting because they are easily learnt.

1. Introduction

Ambiguity is a striking property of natural language distinguishing it from artificial languages. The mathematically well defined case of *syntactic* ambiguity is present in sentences that can be structurally analysed in more than one way:

(1) The officer watched the spy with the telescope.

(2) The word of the Lord came to Zechariah, son of Berekiah, the prophet.

(3) The passengers who left the boat first were old men and women.

In an optimal communication system, a feature like syntactic ambiguity would seem to be dysfunctional. Effective coding implies that one signal corresponds to one meaning and can therefore be interpreted deterministically. Syntactic ambiguity, one would assume, should not be found in a successful communication system like human language. Its ubiquity thus poses an evolutionary puzzle: why has such an apparent imperfection emerged in language?

One view of language imperfections is that they are artifacts arising from *internal* constraints of the innate language faculty (Chomsky, 2002). Nativists conclude from the poverty of stimulus argument that language acquisition is not primarily a matter of learning but rather of setting predefined, genetically determined parameters of the innate language faculty.

This view is contrasted by an approach taken in studies of language evolution which explain hallmarks of language by the fact that language undergoes cultural transmission (Hurford, 2002; Brighton, Kirby, & Smith, 2005). This idea is based on the observation that language acquisition represents a special class of learning problem as the output of the language learning of one generation is the input to the learning of the next generation. Properties of language are exhibited because language itself, as opposed to its users, adapts to be learnable. Brighton (2003) points out that in such an iterated learning framework, language imperfections reflect residues of linguistic evolution through cultural transmission.

This paper introduces syntactic ambiguity as an example of an imperfection of language and discusses the implications of a typical iterated learning model on two questions raised by the evolutionary puzzle that comes with its ubiquity. Under what conditions does syntactic ambiguity emerge? Despite its ubiquity, only a sparse number of grammatical rules in human language are actually involved in syntactic ambiguity. The second question must therefore ask why syntactic ambiguity has been prevented from becoming more pervasive.

The remainder of this paper falls into three sections. In the first, I present a computational model to study the emergence and transmission of syntactic ambiguity. The subsequent section describes observations made in our simulations and illustrates them with three example experiments. These results are summarised and discussed in the last section of the paper.

2. An Iterated Learning Model of Emergent Syntactic Ambiguity

Our intuitions about the behaviour of complex dynamic systems, and any verbal theorising built on them, tend to be faulty. On the other hand, the formalisation of analytical mathematical approaches for such systems (Nowak & Komarova, 2001) proves to be difficult. Computer simulations therefore offer a useful alternative to study the evolution of language (Cangelosi & Parisi, 2001). To this end, Kirby and Hurford (2002) have developed the Iterated Learning Model (ILM). Iterated learning models have been applied to simulate how pivotal properties of language such as recursive syntax or compositionality can originate from cultural transmission (Kirby, 2002). Roberts, Onnis, and Chater (2005) have presented a simplicity-based model to explain the emergence of quasi-regular constructions. They point out that in such a model, the transmission bottleneck is a necessary prerequisite for the emergence of linguistic idiosyncrasies.

In the model presented here, two agents are 'alive' at any one time: a speaker and a learner. Each agent is endowed with an induction algorithm to infer context-free grammars from linguistic data it has observed, and with the ability to produce sentences from that grammar. There is no biological evolution in the agents. The speaker produces a certain number of sentences from its internalised grammar. The learner uses its induction algorithm to infer a grammar on the basis of the speaker's output. The number of sentences the learner is exposed to constitutes the learning bottleneck through which language is transmitted. After one iteration, the learner becomes the new speaker, a new learner is created and the process is started again.

The algorithm for grammar induction applied in this model is described in detail in Kirby (2002) and ultimately based on Wolff (1982). However, apart from minor simplifications, two noteworthy modifications have been made to the original algorithm. First, our algorithm is not enriched with any *explicit* semantics. The notion of syntactic ambiguity is intrinsically rooted in the principle of compositionality, which says that the meaning of a complex expression is a function of the meaning of its constituents and the way they are combined. Different syntactic structures correspond to different complex meanings. The meaning distinctions caused by *syntactic* ambiguity are therefore implicitly expressed in the compositional structures assigned to a sentence, if we presuppose the principle of compositionality for a model.

The second major modification of Kirby's original algorithm is that grammar induction in our model is not necessarily deterministic. Multiple hypothetical grammars can arise in a learner where rule subsumption can be carried out in more than one way. This will be illustrated in section 3. The induction algorithm also produces permutations of the original order in which the linguistic data was presented to the learner and induces alternative grammar hypotheses from these. In multi-generation simulations, hypothetical grammars are either selected for simplicity or for maximal expressivity.

The notion of expressivity adopted in the model corresponds to the number of structurally distinct sentences a grammar can produce. Counting distinct syntactic structures, rather than distinct strings, entails that each interpretation of an ambiguous sentence contributes to the expressivity of its grammar separately. A different syntactic structure also yields a different compound meaning according to the principle of compositionality.

Simplicity as included in the model is based on a non-probabilistic notion of the Minimum Description Length (MDL) principle (Rissanen & Ristad, 1994). The MDL of context-free grammars is calculated according to the methods set up for regular grammars in Teal and Taylor (2000). In its MDL condition, the model is thus very similar to the one presented in Zuidema (2003).

Our model of the learner does not acquire vocabulary or induce lexical categories. We take them to be already learnt. Terminal symbols in the examples of the following section are therefore to be thought of as lexical (or basic syntactic) categories rather than individual words.

3. Example Experiments

This section describes the behaviour observed in the simulations. I will first discuss *single learning* simulations, which were carried out to study the conditions under which structural ambiguity *emerges* in a typical model like ours. In a second step, we will analyse the *stabilisation* of ambiguous grammars in *iterated learning* simulations, where language is transmitted through a bottleneck over generations.

3.1. *Emergence of Syntactic Ambiguity in a Single Learning Model*

Single learning simulations have shown that a range of factors influence the emergence of structural ambiguity during grammar induction. I will use a simple example language to visualise the behaviour of these simulations. The initial example input consists of the data *ba, bca, bab, bac*. After the encounter of the first two strings *ba, bca*, the learner has incorporated the rules $S \longrightarrow ba$ and $S \longrightarrow bca$ to its grammar. These rules can be compressed by subsumption in two different ways, and hence induction yields two different hypothetical grammars at this stage of the learning process:

$$\begin{array}{l|l}
S \longrightarrow Xa & S \longrightarrow bX \\
X \longrightarrow b & X \longrightarrow a \\
X \longrightarrow bc & X \longrightarrow ca
\end{array}$$

We track the further development of the left-hand side hypothesis grammar. The next string *bab* is generalised by replacing the substring *b* with the already established non-terminal symbol X. The new rule $S \longrightarrow XaX$ is incorporated and compressed by subsumption with the existing rule for S:

$$\begin{array}{l}
S \longrightarrow XY \\
Y \longrightarrow a \\
Y \longrightarrow aX \\
X \longrightarrow b \\
X \longrightarrow bc
\end{array}$$

When encountered, the last string *bac* is generalised to $S \longrightarrow XYc$. It is this generalisation and the subsequent rule subsumption which introduce ambiguity to the grammar:

$$\begin{array}{l}
S \longrightarrow XZ \\
Z \longrightarrow Y \\
Z \longrightarrow Yc \\
Y \longrightarrow a \\
Y \longrightarrow aX \\
X \longrightarrow b \\
X \longrightarrow bc
\end{array}$$

This final grammar can generate 12 structurally different sentences. The ambiguous strings *babc* and *bcabc* are produced in two different ways where the substring *abc* is either structured as *(a(bc))* or as *((ab)c)*. Such a grammar can account for ambiguous constructions like the ones in the English example sentences (1)–(3), if we replace its symbols with appropriate phrasal and lexical categories. If we track the right-hand side hypothesis above, we obtain an unambiguous grammar with a 6 sentences expressivity.

Running such experiments, we have been able to identify conditions under which syntactic ambiguity occurs in a model like the one presented. There is one necessary prerequisite for the emergence of syntactic ambiguity: generalisation during grammar induction. No ambiguous grammar emerges without generalisation during the process of its acquisition. But in a typical iterated learning model, learners have to apply generalisation because of the transmission bottleneck. In such a model, syntactic ambiguity is thus one possible implication of the cultural transmission of language.

The likelihood that an ambiguous grammar will be induced is dependent on an interaction of (1) the *properties of the input strings* and (2) the *properties of the induction algorithm*. Structural ambiguity is the result of several coinciding rules induced from the data, rather than something a single learning event would elicit.

We observe that in cases where the hypotheses for a set of linguistic data comprise ambiguous as well as unambiguous grammars, the ambiguous grammars are more expressive than the unambiguous grammars. Similarly, more expressive grammars are more likely to be ambiguous than their less expressive counterparts. Cursorily viewed, this could lead to the assumption that ambiguity would increase over generations in an iterated learning model in which hypothesis grammars are selected for maximal expressivity. We will see below that this is not the case.

3.2. *Stabilisation of Syntactic Ambiguity in an Iterated Learning Model*

The described learner is placed in an iterated learning simulation to analyse the stabilisation of syntactic ambiguity over generations. The number of sentences heard by each generation has proven to be critical in all simulations. For the simulations described here, the bottleneck size was chosen such that the example language is stable if the learners select their hypothesis grammar for MDL. Since the induced ambiguous hypothesis grammars are usually of higher MDL than their unambiguous counterparts, the grammars evolved under such conditions are mostly unambiguous.

The same conditions were then applied to simulations in which the learners select their hypothesis grammars for maximal expressivity. Fig. 1 charts the emergence and stabilisation of syntactic ambiguity in three such simulations, based on the initial example input data *ba, bca, bab, bac*. The grammars evolved under these conditions stabilise in either minimally ambiguous or unambiguous form.

If expressivity was the only pressure to have an impact on language evolution, the given conditions would lead to highly productive and ambiguous languages. Undoubtedly, such languages would be positively dysfunctional and impose insurmountable problems on communication. But, as Brighton et al. (2005) observe, a language is stable if it is expressive *and learnable*. Remember that in our model the pressure for expressivity is realised by the selection of the most expressive grammars. Learnability on the other hand is implemented by the bottleneck through which languages are transmitted.

The example simulations in Fig. 1 illustrate the observation that languages stabilise at a relatively low level of expressivity. Highly expressive languages

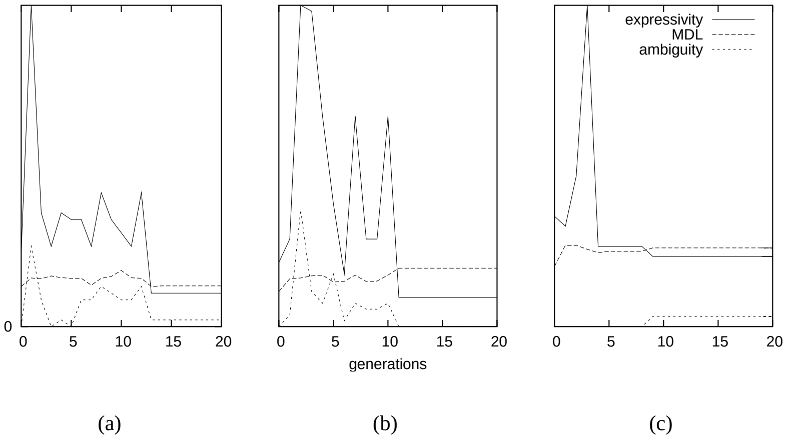

Figure 1. Emergence and stabilisation of syntactic ambiguity in three example simulations, relative to the expressivity and simplicity (MDL) of the evolved grammars. The dotted line denotes ambiguity, the dashed line MDL and the solid line expressivity. The y-dimensions of the lines are relative to each other and therefore not indicated in absolute values where they are > 0. Hypothetical grammars are selected for maximal expressivity in each generation. Due to the learning bottleneck, the evolved grammars stabilise on a low level of expressivity within the first 20 generations of 1000 in total. They are either minimally ambiguous or unambiguous. (a) The evolved stable grammar is minimally ambiguous. (b) The evolved stable grammar is unambiguous. (c) A minimally ambiguous stable grammar evolves from unambiguous unstable grammars.

cannot pass the learning bottleneck successfully and do not reach a stable state. We witness the impact of two competing pressures on the evolution of language: expressivity and learnability.

We can distinguish three types of behaviour in the simulations. The example case in Fig. 1(a) shows how a stable *minimally ambiguous* grammar emerges within the first 20 of 1000 generations of the simulation. The stable grammar emerges after a significant decrease in expressivity and ambiguity. This behaviour can also be observed in the example experiment in Fig. 1(b), where a stable *unambiguous* grammar evolves under the same conditions. We conclude from the equal distribution of these two types of behaviour that minimal syntactic ambiguity does not constitute an impediment to the successful transmission of a language. In cases represented by Fig. 1(c), a stable minimally ambiguous grammar evolves from unstable unambiguous grammars. Expressivity is slightly lowered at the moment of the introduction of ambiguity in generation 9. We have seen before that ambiguity tends to occur in more expressive hypothetical grammars induced from the *same* data. However, in this example it is introduced during the transmission of language from one generation to the next. Syntactic ambiguity occurs in a grammar that is less expressive than the one of the previous generation. This increases the learnability of the example language and stabilises it.

4. Discussion and Conclusion

We have found evidence that *in a typical iterated learning model*, the same phenomenon, the learning bottleneck, is responsible for the two evolutionary puzzles set out in the introduction of this paper. Given its dysfunctionality, why has syntactic ambiguity emerged? And given its persistence in human language, what has prevented syntactic ambiguity from becoming more pervasive? Our simulations suggest that the answer to both questions is the bottleneck through which language is transmitted in such a model.

In our *single learning simulations* illustrated by example experiments, syntactic ambiguity emerges during grammar induction due to coinciding properties of the data and the learning algorithm. The necessary precondition for its emergence is the process of generalisation. Learners need to generalise during language acquisition because they are only exposed to a limited set of linguistic data. The presented type of grammar induction thus implies that syntactic ambiguity reflects a residue of an accidental but not improbable coincidence in the evolution of language through iterated learning.

Does our model therefore oppose the nativist view of language imperfections? It seems that, like any iterated learning model, it takes an intermediate stance. It is potentially nativist in the explanatory emphasis it puts on the specifics of the learning algorithm. But at the same time, it adds a complementary explanatory process, cultural evolution through iterated learning.

The observations made in *iterated learning simulations* suggest that against a pressure for expressivity, the transmission bottleneck ensures that syntactic ambiguity does not become too pervasive in a language once it has emerged. The stable ambiguous grammars emerging from the simulations only exhibit sparse ambiguity. The evolutionary pressure on language to be learnable prevents it from becoming too productive and therefore also constrains ambiguity. Strikingly, the learning bottleneck seems to enable us to explain why syntactic ambiguity is not more pervasive despite its persistence in language. Under the presented assumptions, we can thus disregard such notions as ease of disambiguation or communicative dysfunctionality. The example experiments in this paper illustrate how the stabilisation of ambiguity can be subject to the fluctuation caused by two competing pressures on the evolution of language, learnability and expressivity. In a typical iterated learning model, both pressures are exhibited due to the learning bottleneck through which language is transmitted.

Acknowledgements

Stefan Hoefler was supported by Cogito Foundation Research Scholarship 101/04.

References

Brighton, H. (2003). *Simplicity as a driving force in linguistic evolution.* Unpublished doctoral dissertation, The University of Edinburgh, Edinburgh.

Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution.* Oxford: Oxford University Press.

Cangelosi, A., & Parisi, D. (2001). Computer simulation: A new scientific approach to the study of language evolution. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (p. 3-28). London: Springer.

Chomsky, N. (2002). *On nature and language.* Cambridge University Press.

Hurford, J. R. (2002). Expression/induction models of language evolution: dimensions and issues. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models.* Cambridge University Press.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models.* Cambridge University Press.

Kirby, S., & Hurford, J. (2002). The emergence of linguistic structure: An overview of the Iterated Learning Model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (p. 121-148). London: Springer.

Nowak, M. A., & Komarova, N. L. (2001). Towards an evolutionary theory of language. *Trends in Cognitive Sciences*, *5*(7), 288-295.

Rissanen, J., & Ristad, E. S. (1994). Language acquisition in the MDL framework. In E. S. Ristad (Ed.), *Language computation.* Philadelphia: American Mathematical Society.

Roberts, M., Onnis, L., & Chater, N. (2005). Acquisition and evolution of quasi-regular languages: Two puzzles for the price of one. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution.* Oxford: Oxford University Press.

Teal, T. K., & Taylor, C. E. (2000). Effects of compression on language evolution. *Artificial Life*, *6*(2), 129-143.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language and Communication*, *2*(1), 57–89.

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15.* Cambridge, MA: MIT Press.