# COMPUTATIONAL MODELING OF LANGUAGE EVOLUTION: GAPS AND CHALLENGES

MONOJIT CHOUDHURY
*Microsoft Research Lab India*
*Bangalore, 560080, India*

ANIMESH MUKHERJEE
*Complex Systems Lagrange Lab, ISI Foundation*
*Torino, 10133, Italy*

## 1. Computational Models of Language in Tinbergen's Framework

*Language evolution* is arguably one of the hardest problems in science and is highly interdisciplinary in nature (Christiansen and Kirby, 2003). As evident from the exponential growth in number of publications, this area of research, like other scientific disciplines, has popularly adopted the use of mathematics and, more recently, computational techniques. Nevertheless, modeling seems to be quite a hard problem in language evolution and is still in its infancy (de Boer and Zuidema, 2009). In order to understand the difficulties faced by this emerging sub-discipline we advocate that one should not only look into the challenges and issues within language evolution, but rather *language* itself as the topic of research. In fact, by doing so, as we shall see, one can arrive at two important conclusions – (a) knowledge sharing between different communities studying language is currently limited even though cross-fertilization of ideas seems extremely necessary for the progress of the field, and, (b) there is a need for data creation and analysis before one jumps into modeling.

Language is a human behavior and therefore, it is reasonable to organize language research within the framework of *Tinbergen's four questions* (Tinbergen, 1963). In Table 1, we lay out the areas of language research within these four questions, and then list out the scope of mathematical and computational modeling within each division. We also note the current availability of data for each kind of research and ease of gathering more data.

Table 1. Understanding the interrelations between various disciplines that study "language" and its computational models within the framework of Tinbergen's four questions for ethology.

| *Tinbergen's questions* | *Research Areas* | *Computational Models/Methods* | *Availability of data* |
|---|---|---|---|
| Causation | a) Neuropsychological processing of language<br>b) Physiology of speech production and perception | a) Cognition aware models of NLP (esp. parsing)<br>b) Speech technology (esp. parametric models) | Moderate to high;<br>Easy to gather |
| Ontogeny | a) Language acquisition<br>b) Development of other communicative traits | a) Machine learning<br>b) Models of human learning | Little to moderate;<br>Moderately easy to gather |
| Adaptation | a) Uniqueness of human language<br>b) Language universals and typological studies<br>c) Relevant biological adaptations | a) Quantitative and statistical properties of languages<br>b) Findings from NLP might be of use | High;<br>Quite easy to gather |
| Phylogeny | a) How have languages evolved<br>b) How do languages change?<br>c) Evolution of brain and speech apparatus | a) Models of language evolution & change: math and simulation based<br>b) Cladistics and computational phylogenetics | None to little;<br>Very hard or almost impossible to gather |

Computational Linguistics (CL) or Natural Language Processing (NLP) also study and develop computational models of language. However, the focus is on human language technology, e.g., machine translation and speech processing (Jurafsky and Martin, 2009), and therefore, they do not fit into Tinbergen's framework. On the other hand, some of these research areas can really benefit and get benefitted by the research in other areas of language (e.g., speech technology and physiology of speech production and perception, machine learning for NLP and language acquisition). We do not see, however, much exchange of ideas between the language evolution or cognitive science communities and the NLP community.

In general, a careful study of Table 1 reveals two important points – (a) there is very little sharing of data and knowledge between the various communities studying computational models of language; and (b) there is very little data for directly modeling the problems within "phylogeny". However, other areas have access to large amount of data or at least possibility of gathering data. We argue that these factors, along with the fact that there are not many known or accepted ground truths to base the successful models of language evolution constitute the major hurdles in the area of language evolution. In the following three sections we discuss these issues briefly.

## 2. Lack of Knowledge Sharing

While there is a very vibrant and huge research community working on NLP, there is very little sharing of knowledge between them and the language evolution community. This is apparent from the facts that (a) there are hardly any cross-citations between these communities, and (b) we hardly see works on language evolution being published in ACL conferences. Fragmentation of the community has not only ceased cross-fertilization of ideas, but also made the communities very small. Any research area needs a *critical mass* of researchers to prosper (Shneider, 2009). We believe that lack of critical mass is one of the most detrimental factors in the area of language evolution. This also explains some of the other issues raised in (de Boer and Zuidema, 2009), such as why there has been a lot of modeling efforts only in few areas, even though there is data in the other areas. Indeed, there are not enough researchers in the community to work on many different data sets.

As an aside, it is interesting to note that in many other scientific disciplines, for example physiology and medicine, there is a strong sharing of knowledge. Physiological findings go into development of drugs and surgical procedures, whereas clinical data feed into the models of physiology.

## 3. Lack of Data

There is a lot of linguistic (e.g., corpora, treebanks, phonological databases) and psycholinguistic data (CHILDES, semantic relatedness of words) available for research. In certain areas, where data is not available, it might not be hard to gather more data using advanced techniques and sophisticated instruments. Nevertheless, we have very little access to data for directly validating models of language evolution and it is highly unlikely that we will ever be able to gather large quantities of such data. Therefore, one has to be content with indirect validation. Since insufficient data cannot distinguish between the good and the bad models, limited access to data prevents direct validation of the models leading to a loss of credibility.

## 4. Lack of Ground Truths and Accepted Frameworks

There is hardly any consensus on theories of language evolution; neither do we have concrete theories of language acquisition and processing. In fact, the most celebrated "Chomskyian" framework of *principles and parameters* is also contested. Therefore, it is hard to come up with a set of constraints or a framework which a model should follow. In absence of any such principles, the

space of possible models become infinite and lack of data for validation makes it harder to make strong claims based on computational modeling.

Note that, let alone physical sciences, even biological sciences have well accepted frameworks (Darwinian Theory), ground truths (central dogma of molecular biology) and quantitative laws (Junck, 1997).

## 5. Conclusions

Every scientific discipline goes through several stages of evolution (Shneider, 2009), and we believe that language evolution and especially the modeling approach is in the *first stage*, which is the time for spelling out the agenda, defining the vocabulary and coming up with the basic principles. Therefore, this area of research will benefit from some of the suggestions, such as identification of relationship between models and gaps therein, put forward by de Boer and Zuidema (2009). Nevertheless, we do feel that there is an urgent need for – (a) knowledge sharing with related communities and especially the NLP community which already has gathered lot of data and useful techniques; this can be achieved through organization of conferences and popularizing the work in various venues, (b) concentrating on creation and analysis of data, rather than jumping into synthetic and explicatory models. There are numerous examples of synthetic models in this area, which fall flat if the data is investigated a little deeper, and, (c) laying out the principles of modeling, such as the constraints that any model should satisfy, and the minimum needs for validation. The community should be encouraged to report the failed models as well.

## References


Christiansen, M. H., and Kirby, S. (2003) Language Evolution: The Hardest Problem in Science? In *Language Evolution: The States of the Art*. OUP.

de Boer, B., & Zuidema, W. (2009). Models of Language Evolution: Does the Math Add Up? *ILLC Preprint Series* (PP-2009-49)

Junck, J. R. (1997). Ten Equations that Changed Biology: Mathematics in Problem-Solving Biology Curricula, *Bioscene*, **23**, pp. 1-36

Jurafsky, D., and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall.

Shneider, A. M. (2009). Four Stages of a Scientific Discipline; Four Types of Scientist. *Trends in Biochemical Science* **34**(5), pp. 217-23

Tinbergen, N. (1963). On Aims and Methods in Ethology. *Zeitschrift für Tierpsychologie*, **20**, pp. 410 – 433