COMPUTER MODELING OF LANGUAGE EVOLUTION

- *a partial outline for presentation at COLING, Aug.,2002, Nankang*
- *draft submitted on July 20, 2002.*[i]

William S-Y.Wang and Jinyun Ke
City University of Hong Kong

[1]. Introduction.

Since language is the most distinctive aspect of our species, the origin and evolution of language has intrigued the human mind since ancient times. Earlier speculations on these questions were seldom fruitful because there was virtually no empirical foundation to build upon. It is well known that the linguistic societies in Paris and London banned such discussions in the 19th century. By the middle of the 20th century[ii], however, many of the disciplines relevant to these questions began to come together. Our ability to deal scientifically with these questions has been increasing at an accelerated pace.

These disciplines ranged literally from A to Z, from anthropological concern with the physical development of our remote ancestors, to zoological interests in animal communication and culture. More central here are the discoveries by linguists of universal tendencies found in all languages, by psychologists of the dynamics of language acquisition and loss, and by neuroscientists of how language is organized in the brain.

Over the last several decades, the range of disciplines has broadened in two major steps. First, genetics has come on board with important hypotheses regarding the age of our Most Recent Common Ancestor, and regarding the correlation between groups of peoples and groups of languages. This development started with the so-called classical markers, and has been successively refined to gender-specific materials, first the mtDNA for the maternal line, and then the Y-chromosome for the paternal line. A consensus is gradually emerging that although anatomically modern humans first appeared over 100,000 years ago, our Most Recent Common Ancestor may date to only some 50,000 years ago. Such a date correlates well with

the sudden burst of cultural achievements at many sites in the world, including the navigational skills to sail across large expanses of water.

It is reasonable to associate the origin of language with this date, since it is most likely that the power of language facilitated these cultural achievements. The more recent the emergence date certainly makes the question of emergence more tractable, since there has been less time to obscure the traces of our primordial language. Indeed, some bolder scholars have been prospecting for words that may have existed in that primordial language which have been preserved in most branches of the world's languages today. And other scholars have been exploring the possibility that the unique click consonants still extant in Africa were indeed part of the primordial language phonology which had become lost in the branch of humans that left Africa to populate the rest of the world.

Fascinating these explorations are, the fact remains that most of the pieces of evidence collected from the various disciplines are circumstantial, and that it is not possible to directly reconstruct the stages whereby our ancestors invented language dozens of millennia ago. This leads us to the second major step after genetics – the use of computational linguistics in the study of language evolution. The remainder of my remarks here will be devoted to this new area of research, which for convenience I will refer to as CSLE: computational study of language evolution. This is an area which has burst upon the scene with great vitality, attracting exciting research from a variety of viewpoints. This vitality can be seen from the many anthologies which have become available since 1998, including those by Hurford, Studdert-Kennedy, and Knight (1998), Knight, Studdert-Kennedy, and Hurford (2000), Paris conference (xx), Cangelosi and Parisi (2001), Briscoe (2001), etc.

To begin with, CSLE does not take the currently popular innatists position that there is literally an autonomous organ for language, or that language requires a special bioprogram, or that language is based on any instinct exclusive to it. Obviously, a very wide array of abilities must be in place before our ancestors were ready for language, ranging over sensory, motoric, memorial and cognitive dimensions, as well as social skills in courtship, forming alliances, collaborating in group activities, and strategizing against enemies. Many of these abilities are present to various extents in our ape relatives, even though it is clear our ancestors must have had more language readiness than they do. It is encouraging that some

recent studies are beginning to give us hints on the neurobiological bases of some of these abilities, such as the discovery of the so-called mirror neurons and their implications for the ability to imitate.

The basic assumptions CSLE makes are that numerous interactions among members of a community, as well as among members across communities, over a long span of time can result in behaviors and structures which are quite complex. Furthermore, the path leading to such complex structures often involves phase transitions, points in time at which there are abrupt nonlinearities, where the change seems to be more qualitative than quantitative.

We see such phase transitions in the physical world, for instance, when ice changes abruptly to water, and then abruptly to steam, even when heat is added gradually and by a constant amount. Similarly, we can perhaps identify some phase transitions in the cultural evolution of language, as in the emergence of segmental phonology, the invention of morphology and syntax, the use of recursion in sentence construction, etc. The points in time for such nonlinearities and the driving forces for change are not nearly as well-defined and uniform as in physical systems, of course.

The linguistic analog to the addition of heat which drives the change in water would be the set of communicative needs the early hominids felt as their world became increasingly complex, often a result of their own expanding consciousness as it interacts with the environment. Furthermore, given that by 50,000 years ago there were numerous communities scattered in many parts of the Old World in diverse environmental niches, it is very likely that the evolution of language proceeded at different rates in these communities, each community crossing the various linguistic thresholds in its own way.

[2] Imitation and the emergence of lexicon.

We will now consider three distinct cases of computational studies of language evolution, beginning with the fundamental symbol in language is the word, which pairs meanings with sounds in arbitrary ways. A modern individual typically has many thousands of words in his lexicon through which he sees his universe, and by means of which he communicates his needs and desires. At the outset, however, such symbols were much fewer. Zoologists tell us that no animal in its natural state has more than several

dozen symbols, be they vocal calls or facial expressions or body gestures. The question for CSLE is: how are we to understand the processes whereby the first words were formed and conventionalized, and whereby the words accumulate to large lexicons?

[. . . to be continued . . .]

[3] Optimization and the evolution of tones systems.

Now we turn to a second case of CSLE, of how tone systems evolve in phonological histories. While tones do not play a ubiquitous role in building words as consonants and vowels, they are widely distributed in the languages of Africa, Asia and Native America. The best known system is the 4-tone system of Standard Chinese, contrasting on monosyllables. Other Chinese dialects and minority languages of China vary in the number of tones, peaking at around a dozen. The exact number depends of course on the method of counting. Some years back, I attempted a feature analysis of tones, much has been done earlier for consonants and vowels.

[. . . to be continued . . .]

[4] Computational aspects of lexical diffusion.

The third case we will discuss concerns how changes are implemented in language. The hypothesis of lexical diffusion suggests that a change, whether phonological or syntactic, typically begins in a handful of words and then spreads both lexically from word to word and from speaker to speaker. Such diffusion processes can be modeled mathematically.

[. . . to be continued . . .]

[5] Discussion.

Lastly, we would like to offer some remarks on this exciting new area of CSLE, regarding the assumptions and limitations of the current methodology, as well as regarding the road that lies ahead.

While computational models can demonstrate how certain linguistic structures emerge and/or change, most of them, such as the first and third cases we reported above, have to simulate the interactions between

individuals under certain assumptions and constraints, with large degrees of idealization and simplification. It is actually an advantage of computational models that various assumptions must be made explicit and implementable, and thus can be examined, verified and compared. For example, in simulating the communication interactions among agents, the models have to clearly give the various details on how the meanings are represented in the agent, how meanings are sent by the speaker, and how the listener interprets the signals received. And in simulating language acquisition, the models have to be explicit on which properties the learners are assumed to be endowed with, such as the learning algorithm, if any, which determines how the learners construct their own language by memorizing and extracting the regularities from the linguistic input.

Currently most models make rather strong assumptions or great simplifications of the real situations. For example, in our model of lexicon formation, we assume the meanings are transmitted explicitly and listeners have no problem at all in knowing the meaning intended by the speaker. Many other computational models simulating the interactions between individuals adopt a similar assumption, especially in those where agents are represented by neural networks and they learn the meaning-signal mappings by some training process (eg. Batali 1998). However, the transparency of meaning in communication may not be true in many real situations as ambiguous interpretations are almost always possible.

Moreover, in the case of the acquisition of the first language, it is a well-recognized problem that how children can identify the intended meaning by the adult is not that straightforward. In fact there have been some studies addressing this problem by making the meaning transfer more realistic, such as sending the environmental information together with the signal, and the listener interprets the meaning from the environmental information (Smith 2001). When such realistic constraints and conditions are taken into account and embodied in the models, it may turn out that some of the previous assumptions are not necessary and more interesting results would be obtained.

A hypothesis supported by one model may not be supported by another model which is implemented on a different assumption. For example, Kirby (2001) demonstrates that a compositional language can emerge from a set of random meaning-signal mappings by an iterative-learning model. However, in his model the mapping is represented by a version of Definite

Clause Grammar and learners are assumed to have an induction algorithm which can look for common substrings and infer generalized rules generating them, which are highly biased toward language-like systems. He hypothesizes that a bottleneck effect, which says the learner is only exposed to a small subset of the possible language, is necessary for the emergence of compositional language. However, in a critique of this model by Tonkes and Wiles (2002), a neural network model is implemented, and no explicit rules or generalizations are required, and they show an explicit bottleneck hypothesis is unnecessary while the compositionality still emerges. It can be seen that the representation and assumptions are crucial in such models for examining such hypotheses.

Nonetheless, it is clear that we should be encouraged by what the new area has achieved so far. The knowledge base for research on language evolution must rest on what linguistics has to offer, regarding how the several thousand languages available to us are organized, what the common core of this organization is that is shared by all languages, extending to the most idiosyncratic features observed for a few languages, which mark the outer periphery of what a language can be like. This knowledge base grew tremendously in the 20th century, when linguists described a broad range of languages in many parts of the world which had not been studied scientifically before.[iii] This linguistic knowledge was joined by genetic knowledge since the 1980s in research on language evolution, and by computational studies since the 1990s.

As we look back on this decade or so of CSLE, it is clear that the achievements have been impressive and encouraging. At the same time, we see that there are many central topics on language evolution which await careful formulation and investigation. We will briefly touch upon three of these here, and they are: hierarchy, ambiguity, and heterogeneity.

While there have been exciting simulations on the emergence of the lexicon, and on the formation of phonological systems, not much is known on how hierarchical syntax emerged. Hierarchical structures are a hallmark of complexity, as Herbert Simon noted decades ago. When a chimpanzee takes off the top of a box to get at the banana inside, it presumably recognizes that the two parts of the box are discontinuous constituents of a single hierarchical unit which holds the banana. Cognitively it is comparable to separating constituents of language, such as taking apart ‘call up’ in ‘call him up’, or embedding large constructions within expressions

like ‘ what for’ , such as in ‘ what did you call him up for?’ Linguists have studied the dependency relations of constituents in great depth in a variety of languages – what can be moved, what can be deleted, what can cross over, etc. – and we can hope that computer simulations will soon be able to model such dependency relations within hierarchical structures.

Hierarchical structures are the bases of recursiveness, and recursion is the central mechanism that makes language infinite, via repeated conjoining and embedding. While it is undeniable there is no longest sentence, the fact remains that most utterances in everyday language are quite short, and statistical approximations to these utterances can be very useful in helping us understand the structure and function of such language.

Another question that has intrigued us a lot in recent years is that of ambiguity. It would seem that in an ideal code, one signal should correspond to exactly one message, and that ambiguities of one-many correspondences would cause miscommunication. Yet all languages are rife with such ambiguities at various levels, from polysemy to homophony to syntactic ambiguities. Indeed ambiguity was the most formidable barrier to computational linguistics since its start – from automatic abstracting, to machine translation, to speech recognition – and remains so today, even as methods of disambiguation are getting increasingly sophisticated and powerful.[iv]

From a CSLE vantage point, an interesting research topic would be to see at which points various types of ambiguities emerge as the most rudimentary languages with the simplest lexicon gradually grow toward the level of complexity of modern languages. Embedded in this topic are several questions concerning a typology of ambiguities in the languages of the world: are there universal ambiguities, how do we typologize them, and how do we predict them from the structures in which they reside? Since ambiguities are at once a robust phenomenon and probably unique to human communication, simulating their emergence can tell us much about the nature of language.

As our last point here, we would like to emphasize the tremendous heterogeneity of language. To get our computer simulations started, it is natural to have small and simple models, with a limited community of members who speak a homogeneous language. However, as the simulations continue, and as the members and generations multiply, and as the number

of interactions grows very large. we should expect the languages to become greatly diversified and the linguistic behaviors of the speakers increasingly heterogeneous.

The fact that two people are talking with each other by no means leads to the conclusion that they are completely understanding each other, or that they share the same grammar and linguistic representations. As communities become larger and more complex, their speakers become more diverse as well. Modern linguistics once claimed that the central focus of its research was on an ideal speaker-listener situated in a homogeneous community, an attitude that someone called ' monastic.' As the empirical foundations for linguistics grew, however, there is fuller and fuller realization of how much speakers differ from each other, even in the same family. It is such variability, of course, when amplified manifold across time and space, which produces dialects, and eventually distinct languages. It would be a worthy goal for CSLE to eventually be able to simulate such evolutionary processes with realism. Given that the area has been progressing at such an exciting pace, such a goal may not be too far away.

References:

Batali, J. Computational simulations of the emergence of grammar. In Approaches to the Evolution of Language, ed. by J.A. Hurford, M. Studdert-Kennedy and C. Knight, 405-426, Cambridge University Press, 1998.

Brisco, E.J. ed. Linguistic Evolution through Language Acquisition: Formal and Computational Models. Cambridge University Press, 2002.

Cangelosi, A. and D. Parisi. eds. Simulating the Evolution of Language. London; New York: Springer, 2002.

Cavalli-Sforza, L.L. Genes, Peoples and Languages. New York: North Point Press, 2000.

Cavalli-Sforza, L.L. and M.W. Feldman. Cultural Transmission and Evolution: a Quantitative Approach. Princeton, N.J.: Princeton University Press, 1981.

de Boer, B. The Origins of Vowel Systems, Oxford University Press, 2001.

Schwartz, J.-L., L.-J. Boë, N. Vallèe, and C. Abry. The dispersion-focalization theory of vowel systems. Journal of Phonetics, 25:255-286, 1997.

Freedman, D. and W.S-Y.Wang. Language Polygenesis: a Probabilistic Model. Anthropological Science, 104.2: 131-138, 1996.

Harnad, S.R., H.D. Steklis and J. Lancaster. Origins and Evolution of Language and Speech. New York Academy of Sciences, 1976.

Hockett, C.F. Reprinted in Wang 1991.

Holland, H.J. Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, 1975.

Hurford, J.R., M. Studdert-Kennedy, C. Knight. Approaches to the Evolution of Language, Social and Cognitive Bases. Cambridge University Press, 1998

Ke, J.Y., C. P. Au, J. Minett, and W. S-Y. Wang. Self-organization and selection in the emergence of vocabulary. Complexity, 7.3:41-54, 2002.

Ke, J.Y., M. Ogura, and W. S-Y. Wang. Optimization models of sound systems using Genetic Algorithms, to appear in Computational Linguistics.

Kirby, S. Syntax without Natural Selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, editor, The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form, 303-323. Cambridge University Press, 2000.

Liljencrants, J. and B. Lindblom. Numerical simulation of vowel quality systems: the role of perceptual contrast. Language, 48.4:839-862, 1972.

Lindblom, B., I. Maddieson. Phonetic universals in consonant systems. In Language, Speech and Mind, ed. by L.M. Hyman and C.N. Li, 1988.

Nowak, M. and D. Krakauer. The evolution of language. Proceedings of National Academy of Science, USA, 96:8028-8033, July, 1999.

Redford, M.A., C.C. Chen and R. Miikkulainen. Constrained emergence of universals and variation in syllable systems. Language and Speech, 44:27-56, 2001.

Rizzolatti, G. and M.A. Arbib. Language within our grasp. Trends in Neurosciences, 21.5:188-194, 1998.

Shen, Zhongwei. 1997. Exploring the Dynamic Aspect of Sound Change. Journal of Chinese Linguistics Monograph 11.

Smith, A.D.M. Establishing communication systems without explicit meaning transmission. In J. Kelemen, and P. Sosik, Eds. Proceedings 6th European Conference on Artificial Life, 381-390, Prague, 2001.

Steels, L. The synthetic modeling of language origins. Evolution of Communication, 1.1:1-34, 1997.

Tonkes, B. and J. Wiles. Methodological issues in simulating the emergence of language. In Alison Wray, editor, The Transition to Language. Oxford University Press, Oxford, 2002.

Wang, W.S-Y. Phonological features of tone. International Journal of American Linguistics, 33.2:93-105, 1967.

Wang, W.S-Y. Competing changes as a cause of residue. Language, 45:9-25, 1969.

Wang, W.S-Y. Language change. In Harnad et al, ed. 1976.

Wang, W.S-Y. The Emergence of Language: Development and Evolution: Readings from Scientific American Magazine. W.H. Freeman, New York, 1991.

Wang, W.S-Y. and J-Y. Ke. 言的起源及建 (A preliminary study on language emergence and simulation models, in Chinese), Zhongguo Yuwen, 2:1-5, 2002.

Wang, W.S-Y. and J-Y. Ke. Language and self-organizing consciousness, a commentary on Perruchet, P. and A. Vinter. The self-organizing consciousness. Behavioral and Brain Sciences, 25.3, 2002.

Wray, A. ed. The Transition to Language. Oxford University Press, Oxford, 2002.

Zuidema, W.H. Emergent syntax: the unremitting value of computational modeling for understanding the origins of complex language. ECAL01, 641-644. Springer, Prague, September 10-14, 2001.

Language Evolution and Computation Resources, http://www.canis.uiuc.edu/~junwang4/langev/bibliography/index.html

[i] Support for this research comes in part from the City University of Hong Kong, the Research Grants Council of the Hong Kong SAR, and the Chiang Ching-Kuo Foundation. Many friends have helped us improve our understanding of the issues discussed here. We would particularly like to thank L.L.Cavalli-Sforza, John Holland, James Minett, Merritt Ruhen, . .for their discussions.

[ii] Some good landmarks for the return to respectability to discuss these issues are the well known paper by Hockett (xxx), and the large conference anthologized by Harnad et al (xxx).

[iii] It is sad to note concurrently that the 20th century also marks the accelerated extinction of indigenous languages as these are replaced by a few international languages, empowered by economic and technological success. This development has a homogenizing effect which simultaneously expands the common core and shrinks the outer periphery of the space within which language locates.

[iv] Ambiguities sometimes serve various purposes in linguistic play – in puns, jokes, etc. – but these are surely developments which arose much later after ambiguities have taken root in languages.