Mathematical Models of Human Language

by W. Garrett Mitchener

About human language

The human language faculty is remarkable in that children are able to learn their native language without formal education and in spite of the complexity of the task. Children listen to adult speech and extract words and their meanings, as well as the rules of grammar that assemble them into sentences. The general problem of determining a grammar from sample sentences is impossible without some constraint on the set of possible grammars. A widely accepted theory is that the human brain naturally includes a set of rules known as *universal grammar* or UG, which limits children to a narrow enough set of possibilities that they can reliably learn their parents' grammar.

The intent of my research is to explore mathematical models of how grammars and universal grammars interact and compete for carriers within a population. The results yield insights into how human languages change over time.

The model

We model a large population, each member of which is born with one of the N universal grammars $U_1, U_2, \ldots U_N$ and speaks one of the n grammars $G_1, G_2, \ldots G_n$. A UG consists of a list of which grammars it allows, and is accompanied by a language acquisition algorithm. Individuals are assumed to reproduce at a rate determined by their communication ability, passing their UG to their offspring genetically. Grammar is passed on through teaching and learning. The error rate of grammar acquisition is large enough that it must be accounted for directly in the population dynamics. The rate at which genetic mutation affects UG is small enough that each such mutation can be treated as an isolated event.

$x_{j,K} =$ fraction of the population which speaks G_j and has universal grammar U_K. The whole population is accounted for, so $\sum_{j,K} x_{j,K} = 1$. Since $x_{j,K} \geq 0$, the population state can be represented as a point on a simplex.

$A =$ grammar similarity matrix. $A_{i,j}$ is the probability that a sentence spoken at random by a speaker of G_i can be understood by a speaker of G_j.

$Q =$ learning matrix. $Q_{i,j,K}$ is the probability that a parent speaking G_i produces a child speaking G_j given that both have universal grammar U_K.

$F_j =$ communication ability of speakers of G_j, used as their reproductive rate.

$$F_j = \sum_K \sum_i \frac{A_{i,j} + A_{j,i}}{2} x_{i,K}.$$

$\phi =$ average reproductive rate of the population.

$$\phi = \sum_K \sum_j F_j x_{j,K}.$$

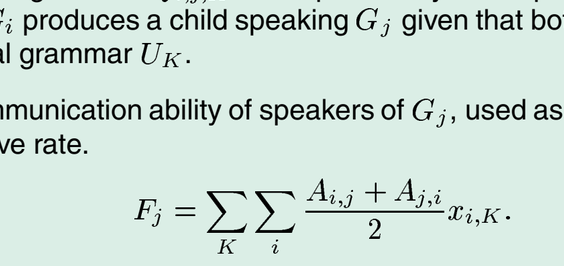
The language dynamical equation is as follows:

$$\dot{x}_{j,K} = \sum_i F_i x_{i,K} Q_{i,j,K} - \phi x_{j,K}.$$

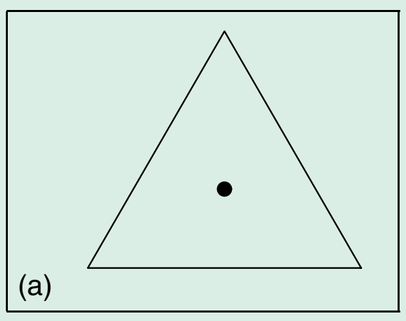
Each term in the sum represents reproduction by an individual with U_K, where some offspring learn G_j.

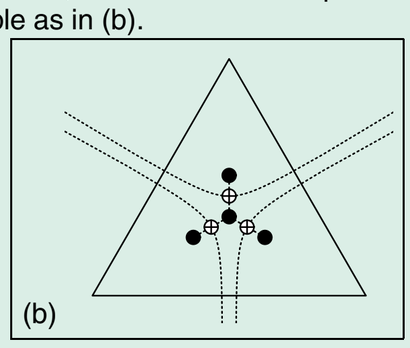
Results for one UG

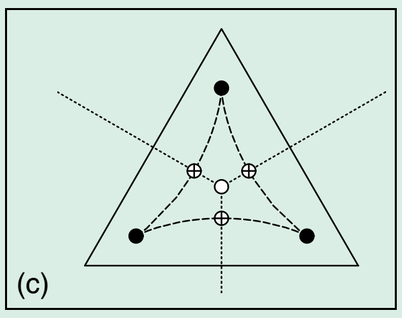
Take the case of a single UG and fully symmetric A and Q matrices. For three grammars, we can represent the population as a point on a triangle.



If learning is error-prone, the population settles into an incoherent equilibrium, where all grammars are present, as in (a). If learning is very reliable, it settles at a coherent equilibrium, where one grammar dominates as in (c). If learning is in between, both outcomes are possible as in (b).


(a)

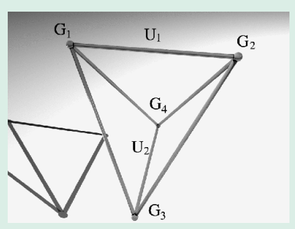

(b)


(c)

This model provides a mathematical foundation for understanding how languages change over time. For example, Old English was stable until England was invaded by Scandinavians. Their language, Old Norse, added linguistic "noise" and reduced learning reliability, thereby destabilizing the grammar. At the end of the invasion, the population settled into a new grammar, Middle English.

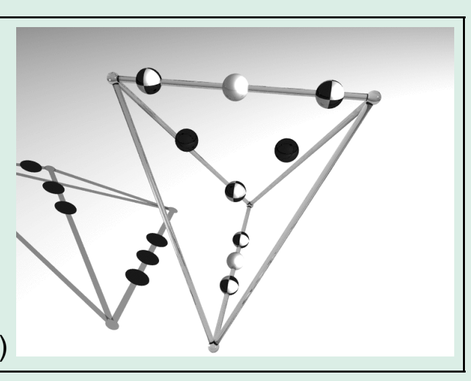
Results for several UGs

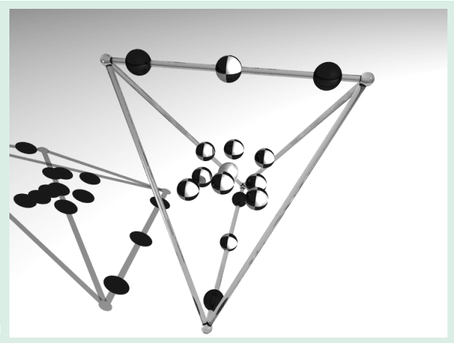
Consider a case with two UGs and four grammars, where $U_1 = \{G_1, G_2\}$ and $U_2 = \{G_3, G_4\}$. Populations in this case can be represented as points in a pyramid.



We would like to determine when one UG can take over from another by examining the stability of homogeneous states. These are the states with only one UG, and appear on the top and bottom edges of the pyramid.

In the case shown in (d), the two UGs admit similar grammars, and all homogeneous populations are unstable. The population always settles into a state in the middle, where both UGs coexist. Either one of these UGs can invade the other.


(d)


(e)

However, in the case in (e), the two UGs admit dissimilar grammars, and all homogeneous populations are stable. Neither UG can invade the other.

The important consequence of picture (e) is that genetic mutations resulting in drastic changes to UG are likely to die out quickly, because the individuals with those mutations are unable to communicate properly with the rest of the population. Picture (d) shows that if a mutation to UG results in small changes to the grammars, then it is likely to spread.

Things to read

♢ W. G. M., *Bifurcation analysis of the fully symmetric language dynamical equation,* J. Math. Bio., submitted. ♢ W. G. M. & M. A. Nowak, *Competitive exclusion and coexistence of Universal Grammars*, Bull. Math. Bio., submitted. ♢ M. A. Nowak et. al., *Computational and evolutionary aspects of language*, Nature 417:6889. ♢ *The Language Instinct,* by Steven Pinker.

PACM
The Program in Applied and Computational Mathematics
Fine Hall, Washington Road
(609) 258-3703; Fax: (609) 258-1735
www.math.princeton.edu/PACM

PrincetonUniversity