



语言的起源及建模仿真初探*

王士元 柯津云

语言的起源是研究语言的一个中心问题,也是研究人类的一个中心问题。因为人类最显著的特点就是有语言。没有别的动物发明过语言,也没有其他的动物学得会一个完整的语言。我们可以认为,研究语言的起源,是研究现代人的起源的一个最重要的部分。研究人类的起源,与研究语言的产生是密不可分的。

过去的一百多年里,考古工作已经发掘出很多古人的化石。如中国周口店发现的北京猿人,还有据考察距今二十万年的大荔人。原来人们认为我们就是这些古人的后代。可是近年来随着研究的深入,考古学家们认为这些古人不可能都是现在生活在地球上的人类的真正的祖先。如我们从动物的演化史知道的,很多动物的物种都没能通过代代相传保留到现在,他们在演化的过程中被淘汰掉了。这个论断从基因学的研究成果中可以得到支持。

基因研究是我们追溯人类的起源的一个非常有力的工具。关于现代人的祖先,这方面的工作也许可以从 Cann 等人 1987 年的一篇文章^①算起。他们采用母系传递的线粒体 DNA (MtDNA) 材料进行研究,他们得到一个谱系树,发现最早的分支是非洲人种与其他地方的人种分开。这个研究给出的结论是现代人的祖先,即智人(Homo sapiens),是从非洲迁徙到世界各地的。他们认为人类最近的共同老祖宗可能生活在十至二十万年以前。

Cann 等人的成果给研究人类的起源带来了很大的推动。近一二十年,人类基因学进展神速,几乎可以说是日新月异。值得一提的是最近的一个报告,2000 年 6 月发表在美国科学院学报的一篇论文,^②他们采用的材料是父系传递的 Y 型染色体,得到一个比 Cann 等人推算的更近的时间,他们认为人类最近的共同老祖宗距离我们只有五万多年。

这个最新的结论跟考古学家的一些已有的发现和推论非常吻合。考古学家们认为在五万年前,在人类文明演化的过程中出现一些很重要的里程碑,例如一些山洞里的壁画,埋葬死人的坟墓以及墓里陪葬的花草。从这些可以认为那时的古人已经有了艺术和宗教的萌芽。^③同时大约在五万年前,有证据证明人类第一次在航海的技术上有了突破,能从亚洲的南部迁徙到澳大利亚。所以基因学在发现人类最近的共同老祖宗距离我们大约五万年,而考古学也同样有很多证据证明那是个重要的时期,那时的人类在文明上已经有了一些很显著的突破。

是什么原因使一些古人在那个时候有了这些突破而有个大跃进呢?我们猜想那些古人能发达到那个程度,最主要的原因是他们发明了语言。语言能帮助他们开始有系统的、复杂的、

* 此文是基于王士元在广州举行的第八届当代语言学全国会议(2000 年 10 月)上的主题报告。报告的研究课题得到香港城市大学的研究项目 9010001 的资助。我们同时感谢课题的合作者 Stephen Smale、李行德、郑锦全、姚远和于江生。

抽象的思维;语言能使一群人团结起来,互相沟通,传递知识,积累知识。一些个人无法做到的事情可以通过语言的沟通进行有效的组织而成为可能。

那么“语言”这个奇物到底是怎样发生的呢?它是在一个地方起源然后逐渐散播到其他地方去的,还是同时在世界几个地方由不同的人群分别发明的呢?以前语言学家一般的看法认为语言是单源发生的。他们觉得语言是这样一个复杂而深奥的东西,发明语言一次已经是需要很大的巧合了,要能被发明几次,这个可能性则一定是很小很小的了。可是我们认为这个想法有两个基本的错误:一个是当代的语言的确非常复杂,每个语言都有成千上万的词和一些极为奥妙的结构。可是这些应该是几万年演化的结果。我们可以想象得到,最实际的原始语言,一定是非常简单的。而后来由于各种需要逐步逐步变得越来越复杂了(请参看 Schoenemann 1999^④)。就像数学的开始也只不过是发明数字来数东西而已,而当代的数学却是奥妙无比了,单源说的另一个错误可以从概率上分析。我们认为单源说忽略了一个很重要的变数。我们知道,五万年前一定有很多人,很可能有成千上万的不同人群。即使我们假设语言在一个地方产生的概率很小,但如果人群的数目足够多,我们可以用概率模型来证明多源说的发生可能性会比单源说的可能性大(详细的证明请参看^⑤)。

五万年这个时间距离在我们讨论语言的起源上还有一个重要的意义。近年来,很多语言学家把语言起源看作一种基因突变的结果。在文献中我们常常能看到诸如语言器官(language organ),语言生物程序(language biogram),语言本能(language instinct)等等的字眼。很明显,我们说话时需要呼吸和听觉等器官,我们的词汇需要记忆,我们造句的时候需要有种种分析和组合的能力。可是这些不同的能力是建立在生命的最基本的材料上以适应最基本的生存的需要,不可能是要产生语言而从基因变异特殊地演化出来的。而且,因为生物演化往往是要经过几百万或几千万年才能起作用的,所以语言是绝对不可能在短短的五万年以生物演化的形式由基因突变而来(这个观点的详细阐述见王士元曾发表于1978在印度海德堡的一个演讲,后收录于^⑥)。

根据基因学和考古学的一些发现,我们认为原始语言很可能是在五万多年前起源的。并且我们认为这个起源是多源的,也就是说,是在几个地方,分别独立地发生的。那这件事究竟是怎样发生的呢?这是学术上的一个大问题,很早以前就有智者思考过。比如古希腊的柏拉图(Plato),他借用了 Homogenes 的对话,在这个方面发表过意见:“Any name which you give is the right one, and if you change that and give another, the new name is as correct as the old.”(无论你起什么名字,都是对的,如果你想改名,起一个新的,也和旧的一样是对的)Homogenes 的意思恰好跟我们中国古代的荀子的看法很相近。荀子认为“名无固宜,约之于命,约定俗成,谓之宜。”“名”指的就是我们今天所说的“词”。特别有趣的是专家们推断柏拉图的那本书是公元前370年左右写的,而荀子是公元前323年出生的,所以他们可以说是同一时代的。这两个伟大的思想家远隔万里,在地球的两端,在同一个时代,得出非常相似的结论,这种巧合让我们惊讶之余也得到一点启示。

最近我们在语言起源这个问题上作了一些初步的探索。激发我们思路的主要有三方面的观察。第一个是上面提到的荀子所说的“约定俗成”。这究竟是什么意思呢?又是怎样实现的呢?我们想用仿真的方法做实验看看。词是语言的基本的单位。先从词的形成着手,我们认为这是个很适合的出发点去研究语言的起源。第二方面,我们注意到近年来其他学科的一些发展。有些学者正在把好几个学科分别考虑的问题归纳到若干个很基本的原理上,提出了混

混沌、复杂论等理论。其中著名物理学家、诺贝尔奖获得者 Murray Gell-Mann 对这些新发展很有兴趣,也很有信心。^⑦这些新思想中的一个中心概念就是非常复杂的现象可以从一组非常简单的现象中产生。而这个产生的过程并不需要任何超级的智慧来设计或制造,只需要适当的起始条件,以及一些偶然的因素和选择的机制。我们认为语言就是一种很典型的复杂系统,已有的研究复杂系统的理论和工具对研究语言的起源会有很多的帮助和可供借鉴的地方。关于复杂现象复杂系统的产生,前人大概都抽象地考虑过,但可以说是只有纸上谈兵。可是现在有了先进的计算机,有庞大的计算能力。我们可以把种种的假设用计算机程序来实现,建立模型,模拟现实中的问题,这就是仿真的研究方法。在研究复杂系统的产生时,仿真是个非常重要的方法。John Holland 在这个方向做了一些开创性的工作。^⑧在语言学研究上近年来也有一些建模仿真的前沿的工作,其中 Briscoe(2000)^⑨是个很好的尝试,他利用语言习得模型来研究语法的习得,探索语言与大脑的共同进化的可能性。

我们思路的第三个方面的提示来自动物学的研究成果。动物的一个常见的本能行为就是模仿。比如,小猴子会模仿它的母亲,只吃猴子能吃的果实。模仿是动物生存的一个必要条件。有些动物是特别能模仿声音的,我们熟知的譬如鸟类中的八哥和鹦鹉,还有哺乳类的海豚。很多灵长类的动物,包括人,也都是具有模仿的天性的。大家都知道小孩子出生不到几个月就会开始模仿身边的大人的语言了。

我们推想的语言产生的最初始情形是:在原始人群里,一开始时人们发出声音,只是对环境的一种潜意识的不自觉的反应;后来人们偶然地意识到可以用一些简单的声音,来指示身边的一些事物。这样人们就可以进行最简单的交流。(Keller 虚构了一个原始人 Charlie 的故事来构拟语言的开始,^⑩虽然这只是个故事,但对我们很有启发性)我们假设一开始时不同人可能用不同的声音来表示同一个事物,各自表达的方式不一样,但后来经过彼此长期的交流,互相模仿,最后在人群里形成一个统一的信号系统,这就是“约定俗成”,语言的开始。

我们假设在还没有语言的时候,有一群原始的人,有几个概念对这群人特别重要,比如“小心,狼来了”,“我们做个朋友好不好”等等。这一群人会发出一些整体不可分的信号,去表达这些概念,同时他们也能辨别和模仿这些信号。“整体不可分的信号”的意思是指不管声音信号的长短,一整段信号只表示一个概念,把信号拆分开则其中的部分不能表示任何意义。我们假设有这样一组概念, $\Phi = \{m_1, m_2 \dots m_M\}$, 有一组信号 $\Psi = \{u_1, u_2 \dots u_U\}$ 。一开始时每个人都有自己的一套方法,从 Ψ 中任意选一个信号 u_i 去表达一个概念 m_j 。所以对每一个人,我们可以给出他/她在某一时刻的 Ψ 与 Φ 的对应表。比如在一个人数为 P 的人群里, A 和 B 两个人的 $\Psi - \Phi$ 表如下:

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	...
$\Psi(A)$	u_2	u_4	u_1	u_3	u_2	u_{10}	u_5	u_7	u_6	u_8	...
$\Psi(B)$	u_5	u_3	u_1	u_7	u_9	u_3	u_{10}	u_2	u_4	u_8	...

从上表可以看出,一个人可以用一个信号表示不同的概念,例如 A 同时用 u_2 来表示 m_1 和 m_5 , B 同时用 u_3 来表示 m_2 和 m_6 。这相当于我们现代语言学里所说的同音词,同音词越多,传达意思的确定性就越差。另一方面,不同的人 Ψ 和 Φ 的对应会不一样。两个人可能用同一个信号来表达同一个概念,比如, A 和 B 都用 u_1 来表示 m_3 , 用 u_8 来表示 m_{10} , 但并不是所

有的对应都是一样。在一个人群里,各人的 $\Psi - \Phi$ 对应越一致,这个人群内部的交流就越有效,协调就越好。我们设计了两个指标,分别作为观察以上两方面的变化,也就是(1)一致性,标记为 C,(2)确定性,标记为 D。两个指标的计算方法如下面两个公式:

$$(1) C = \frac{1}{M} \sum_{i=1}^M C_i, \text{ 其中 } C_i = \frac{\sum_j \binom{S_{ij}}{2}}{\binom{P}{2}}, S_{ij} \text{ 是使用信号 } j \text{ 来表示第 } i \text{ 个概念的人数, } \sum_j S_{ij} = P, \text{ 另}$$

$$\text{外显然,当 } S_{ij} \leq 1 \text{ 时, } \left\{ \begin{matrix} S_{ij} \\ 2 \end{matrix} \right\} = 0$$

$$(2) D = \frac{1}{P} \sum_{i=1}^P D_i, \text{ 其中 } D_i = \frac{\sum_{j=1}^{N_i} P_{ij}}{N_i}, N_i \text{ 是第 } i \text{ 个人所使用的信号的个数, } P_{ij} \text{ 是第 } i \text{ 个人能正确理解信号 } j \text{ 的概率。}$$

我们假设起始状态时在一个人群里,每个人的 $\Psi - \Phi$ 对应都是随机的。当人们开始进行交流时,会有一些模仿现象发生。比如 A 和 B 在交流时,B 可能模仿 A 的声音,用 u_1 表示 m_2 。我们假设了以下五种不同的模仿策略,通过仿真,可以看出不同的模仿策略会得出不同的结果。我们希望能从仿真中看出某些原始语言发展的可能途径,还有发展所需的环境,比如人群的大小,所需的时间等等。

策略 1:随机模仿,模仿的方向随机而定。

策略 2:模仿大多数,当两个人的 $u_i - m_j$ 对应相同时,此对应得到加分。当两个人的 $u_i - m_j$ 对应不一致时,对应分低的向分高的模仿。

策略 3:在不增加同音词的条件下模仿。

策略 4:既要跟随大多数,又不要增加同音词。策略 2 和 3 的条件同时符合,模仿才会发生。

策略 5:模仿大多数,或者不增加同音词。策略 2 和 3 的条件符合其中之一,模仿就会发生。

我们把模型简化,假定概念的数目(M)是个定量,每次两个人对话,只交流一个概念,其中可能会有一个人模仿另一个人的声音。我们观察一段长时间的模仿的过程,并分别改变人数 P 和可用信号的数目 U,分析不同 P 和 U 的情况下的 C 和 D 的一些特点。我们进行仿真,每个策略的实验进行 20000 次对话,每个策略重复实验 100 次。人数 P 和信号数目 U 分别从 10 增加到 50。我们可以观察到以下几种现象:

a) 策略 1 随着人群 P 的增大,C 趋向 1 的速度减慢;信号数目 U 增加对 C 的变化影响不显著;最难达到 $C=1$ 的是 $P=50, U=10$ 。

b) 策略 2 随着人群 P 的增大,C 趋向 1 的速度减慢;信号数目 U 增加,C 能达到 1 的可能减小;最难达到 $C=1$ 的是 $P=10, U=50$ 。

c) 策略 3 如果信号数目 U 少,几乎不可能达到 $C=1$;随着人群 P 的增大,C 趋向 1 的速度减慢,人群 P 越大,达到稳定时的 C 越小;随着信号数目 U 增加,C 增大的速度越快。

d) 策略 4 如果信号数目 U 少,人群 P 越大,达到稳定时的 C 越小;如果信号数目 U 多,人群 P 越大,达到稳定时的 C 越大。

e) 策略 5 随着人群 P 的增大,C 趋向 1 的速度减慢;随着信号数目 U 增大,C 达到 1 的速度也稍微减慢。

f) 除了策略 3 的 D 可以一定达到 1 以外,其他策略的 D 都不能全达到 1。信号数目增大时,100 次实验的平均值增大,D 的变化范围减小。人群的大小 P 对 D 的影响不大。

我们尝试用严谨的数学模型帮助解释仿真的实验结果。从随机模仿的策略(策略 1)的仿真结果可看到在人数 P 少,声音数目 U 小的时候,C 总是很快地达到 1,也就是人群很容易地形成统一的信号系统。我们可以用马尔可夫链(Markov chain)来从数学上证明这种统一的必然性^①。(此文附录表 1,五种策略的 100 次实验结果的 C 和 D 的平均值,图 1,五种策略的四种参数条件下的 100 次实验的 C 和 D,本刊限于篇幅,未便刊出,谨此说明——编者)。

有研究认为人类祖先的群体中人数不会太多,我们的仿真说明在人能使用的信号数目不多的情况下(即 U 不大时),无论采取什么模仿策略,人越多都越难以达到统一,形成必要的交流。但如果能使用的信号数目大时,则不然,不同的策略有不同的结果。不过我们认为在人类语言萌芽的时期,人还不懂得使用信号的组合,只会使用整体不可分的信号,所以能使用的分辨的信号数目不多。因此我们的仿真实验结果也支持人类祖先的群体中人数不会太多的观点。另外从仿真结果可看到,同音词是很难避免的(除了策略 3 以外),即使在假设 U 与 M 比例较大时(实验中 U:M 最大为 5)。如果我们假设人能使用的整体性信号数目足够大(U:M) 2),那么我们可以认为策略 3(即尽力避免增加同音词的策略)和策略 5(即为了跟随大多数或者减少同音词而模仿对方)是假设中两个最优的模仿方法,可达到最大的一致性和确定性;如果 U 不大,那么策略 5 是目前假设的最优方案。

我们认为从没有语言演化到现代的语言,主要是要跨越两个大门槛。第一个是词汇的形成。我们报告的实验就是对这一过程的一个初步探索。我们目前把模仿行为分成五种,但是可以很容易想象词汇形成的实际情况会是复杂得多,比方说人跟人之间的关系,一定会影响模仿的方向,很可能是弱者模仿强者,而且接触的时候不会总是一对一的。再者,模仿的时候也不会一学即会,偏差或错误很可能会发生。另外我们报告的这个实验还没有考虑旧词的消失和新词的发明。另一方面,这个模型也可用来研究词汇扩散。我们可以在一个一致性已达到 1 的稳定的群体里加入另一种不同的对应系统(模拟现实中有移民迁入,语言接触这种情况),观察这个群体的词汇结构的变化。

第二个门槛是语法的形成。人类语言区别于动物的语言的一个特征就是具有语法。我们希望能用模型模拟语法中的层次结构、循环体系、不连续成分等人类不同语言具有的普遍语法的种种特征,^②这些特征产生的过程及其产生的必要性或必然性。还有在人类语言中普遍存在的歧义现象,也是我们希望研究的一个方向。在上面报告的实验中出现的同音词,就是歧义现象中的语义歧义的一种主要原因。我们也希望用模型仿真的实验来研究这些语义和语法中出现的各种歧义产生的原因,以及解决歧义的方法。

在很多科学研究的领域,要分析一个复杂的现象,了解现象的本质,科学家们往往采用分而治之的方法,先把问题分解简化,进行抽象,建立局部简单的模型来研究。而语言正是这样一种非常复杂的现象,我们如果要了解语言的本质,重构语言产生的过程,建立模型进行模拟是一种必要的有效的方法,把语言的各种重点因素作为参数逐步加入模型,进行分析。本报告的仿真实验是这方面的一个初步的尝试。同时我们借用数学的工具来帮助分析仿真实验的结果,我们认为这是一种很好的跨学科进行研究的尝试。希望能随着我们对语言的各种现象的更多的了解,提取更多语言基本的因素,丰富我们的模型,仿真实验能帮助我们揭示语言的本质,了解语言这个人类特有的现象,也是决定我们能演化成人类的一个关键因素。

附 注

- ① Rebecca L. Cann, Mark Stoneking & Allan C. Wilson, 1987. Mitochondrial DNA and human evolution. *Nature*, 325, 31 - 6.
- ② R. Thomson, J. K. Pritchard, P. Shen, P. J. Oefner and M. W. Feldman, 2000. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Nat. Acad. Sci.*, 97. 7360 - 7365.
- ③ Richard G. Klein, 1999. *The Human Career: Human Biological and Cultural Origins* (second edition). The University of Chicago Press.
- ④ P. Thomas Schoenemann, 1999. Syntax as an emergent characteristic of the evolution of semantic complexity. *Minds and Machines*, 9. 309 - 346.
- ⑤ David A. Freedman & William S - Y. Wang, 1996. Language polygenesis: a probabilistic model. *Anthropological Science*, 104. 2. 131. 138.
- ⑥ William S - Y. Wang, 1996. *Explorations in Language*. 105 - 131. Taipei: Pyramid Press.
- ⑦ Murray Gell-Mann, 1994. *The Quark and the Jaguar: adventures in the simple and the complex*, 316. New York: W. H. Freeman.
- ⑧ John H. Holland. 1995. *The Hidden Order: How adaptation Builds Complexity*. Massachusetts: Perseus Books. (中文译本: 约翰·H·霍兰著, 隐秩序—适应性造就复杂性。上海科技教育出版社, 2000)
- ⑨ E. J. Briscoe. 2000. Grammatical acquisition: inductive bias and coevolution of language and the language acquisition device. *Language*, 76. 2.
- ⑩ Rudi Keller, 1994. *On Language Change: the Invisible Hand in Language*. London: Routledge.
- ⑪ 这个想法最初是 Manuel Blum 教授跟我们交谈时提出来的, 后来又在 Partha Niyogi 教授较具体的建议下推导出来的。
- ⑫ Herbert A. Simon. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106: 467 - 482.

(王士元 柯津云 香港城市大学)

浙江大学汉语史研究中心《中古近代汉语研究》创刊

《中古近代汉语研究》是浙江大学汉语史研究中心的主要刊物, 刊登国内外研究中古、近代汉语方面的论文、译作、书评等, 包括文字、音韵、训诂、语法、修辞、校勘等各方面的考据性和理论性、综述性成果, 尤其欢迎把中古汉语、近代汉语打通起来研究以及把它们和上古汉语或现代汉语作比较研究的论文。中古汉语研究旨在加强汉语史研究的各个时段和环节的衔接, 注重古今汉语的沟通。本刊已经出版一辑。欢迎从事相关研究的学者惠赐佳作。稿件采用匿名评审的办法。投稿者请用单独一页标注作者姓名及详细通信地址。

(浙江大学汉语史研究中心《中古近代汉语研究》编辑部)