在人类生境约束下思考语言的设计原理和运作机制*

袁毓林

（澳门大学　人文学院中国语言文学系　澳门　999078；
北京大学　中文系 / 中国语言学研究中心 / 计算语言学教育部重点实验室　北京　100871）

提　要　文章述评了霍凯特提出的人类语言的 13 种设计特征，并重点讨论其中的语言符号离散性和二元构型特征。从自然语言处理的角度看，离散性使语言符号的形式与意义之间存在着语义鸿沟，需要把自然语言的词向量化为连续性数值才可以进行计算，这种运算的结果可以从语言学上进行解释。由于语言系统和生物系统在二元构型上存在相似性，可以将自然语言处理的一些模型运用于生物分子研究领域。根据托马塞洛关于人类沟通的起源和合作性心理平台的学说，任何诉诸语言本身有某种繁复而自足的深层结构的幻想都是不切实际的，语言学家应该把语言置于人类生境（即人类进化与生存的现实境况）下，来思考语言的设计原理和运作机制，从而对语言的结构方式和功能效用有更加切实的了解，进而开辟一种更具人文主义情怀的语言学研究进路。

关键词　人类生境；语言的设计特征；离散性；二元构型；合作性心理平台
中图分类号　H002　**文献标识码**　A　**文章编号**　2096-1014（2022）06-0085-12
DOI　10.19689/j.cnki.cn10-1361/h.20220607

On the Constraint of the Human Habitat on the Design Principles and the Operating Mechanism of Language

Yuan Yulin

Abstract　Based on a critical review of Hockett's (1960) discussion of underlying design features of human language and Tomasello's (2010) psychological focus on human communication, this paper attempts to argue for the importance of humanizing and historical dimensions in language computerization study. This paper examines what Hockett (1960) proposed and referred to as 13 design features by which human language is characterized with a focus on two of these features: discreteness and duality of patterning. I first discuss three different ways of defining the discreteness of linguistic signs and the respective focal points of these three definitions and introduce methods to integrate the discrete attributes of linguistic signs into continuous vectors as well as computational applications of these methods. Secondly, I look into the similarity between the linguistic system and the biological system in terms of their duality of patterning, including how relevant algorithms and models in the field of natural language processing are applied in molecular biological studies. Finally, this paper turns to discuss Tomasello's (2010) theory regarding the origination of human communications and his ideas about the cooperative psychological platform. While showing that any effort to delve into an underlying complicated and self-sufficient structure is bound to be unrealistic, this paper argues that linguistic researchers should locate language in the human habitat (i.e., the real situation in which human beings exist and evolve) thus to reflect upon the design principles and the operating mechanism of language. In so doing, a greater prospect can be obtained into the structural patterns and functional utilities of language with an aim of opening up a linguistic approach with prominent humanistic considerations.

Keywords　human habitat; design features of language; discreteness; duality of patterning; cooperative psychological platform

作者简介：袁毓林，男，澳门大学教授，主要研究方向为理论语言学和汉语语言学。电子邮箱：yuanyl@pku.edu.cn。

*　国家科技创新 2030“新一代人工智能”重大项目“以自然语言为核心的语义理解理论、模型与方法”（2020AAA0106701），国家社会科学基金专项项目“新时代中国特色语言学基本理论问题研究”（19VXK06）。

一、引言：语言研究受人类生境的约束

语言研究面临的第一道门槛是问题的提出和方法的抉择，而问题和方法的确定又往往取决于研究者对于语言的结构与功能、来源与演化等根本问题的认识。为了检讨我们秉持的基于人文主义的语法研究道路的合理性，本文首先介绍和简评霍凯特（Hockett 1960）提出的人类语言的 13 种设计特征，然后讨论对于语言符号离散性的 3 种理解及其各自的侧重点，再介绍在自然语言处理领域中，怎样用“词嵌入”模型，把离散性的语言符号向量化为连续性数值，以及这种词向量在有关计算任务中的应用与效果；接着介绍和讨论语言系统和生物系统在二元构型（双重分节性）上的相似性，特别是由这种二元双层编码的相似性而引发的、自然语言处理的有关算法和模型在生物分子领域中的运用及其效果；最后介绍和评论托马塞洛（Tomasello 2010）关于人类沟通的社会起源和语言奠基于其上的合作性心理平台的学说，主张语言学家应该把语言置于人类进化与生存的现实境况（简称“生境”）下，来思考语言的设计原理和运作机制，从而对语言的结构方式和功能效用有更加切实的了解，进而开辟一种更具人文主义情怀的语言学研究进路（approach）。也就是说，语言研究受人类生境的约束，语言理论只能戴着现实语言生境的镣铐跳舞，要丢掉任何不切实际的企图为语言建造一座宫殿的幻想（袁毓林 2019）。

二、人类语言系统的设计特征

霍凯特（Hockett 1960：90 ～ 92）首次提出人类语言有 13 种设计特征，包括：[①]

（1）口耳通道（vocal-auditory channel）[②]。它区别于手语的姿势、蜜蜂的跳舞、刺鱼的求爱仪式；它的好处是可以解放手脚等身体部位，便于人类在交谈的同时从事其他活动。

（2）四散传播与定向接收（broadcast transmission and directional reception）。一个语言信号可以被一定范围内的任何听觉系统听到，并且声音来源可以用双耳锁定。

（3）迅速消失（rapid fading）。这意味着语言信号不会为了听者的方便而多停留一会儿，不同于动物的足迹和臭迹会保持一段时间。所以人类发明了书写记录，这是人类非常晚近的文化进化的成果。

显然，（2）（3）两点是由声音的物理性质决定的，也是（1）的不可避免的结果。

（4）互换性（interchangeability）。说话人可以产出任何他可以理解的语言消息。但是，雄性刺鱼和雌性刺鱼各自特有的求爱示意动作却是互不相同的，双方都不能使用对方的合适动作。另外，人类母婴之间进行交际时，双方都不适合发出对方特有的信号，或者做出对方典型的回应表达。

（5）完全反馈（total feedback）。当人类说话时，他会注意听跟他说话相关的一切事物；而雄性刺鱼并不会看它自己的眼睛和腹部的色彩，尽管它主要以此来刺激雌性刺鱼。反馈是十分重要的，因为它使得所谓的交际行为的内化成为可能；而这种内化的交际行为至少构成了思维的主要部分。

显然，这（4）（5）两点是通过跟其他交际系统进行比较才得以明确的。

（6）专门化（specialization）。它说的是：身体努力和发出言语声波只是让它成为一种信号。一只狗喘着气吐出舌头，给自己降温和保持合适的体温，但这只是一种生理动作。在吐舌喘气的同时，它

① 对这些设计特征的名称的翻译，参考了王士元（2017：9），但不完全相同。对于这些特征的解释、举例和说明，我们加入了自己的认识。如果要引用，务请核对原文。

② vocal-auditory channel 可以有“发声–听觉通道、口叫–耳听通道”等多种翻译。

可能偶尔附带着发出一些声音，从而会让其他狗（或人）知道它在哪儿和感觉如何；但是，这种传递信息的方式并不是专门化的。

（7）语义性（semanticity）。指在语言中，一段消息触发了特定的结果，因为消息中的构成成分（比如词）跟我们周围世界中反复出现的特征或情境有一种相对固定的联系。例如，英语单词 salt 指盐，而不是糖或胡椒粉。据此，上面（6）中狗的吐舌行为不具备语义性，它不是一个意指狗很热的信号，而只是狗很热的一个部分（一种表现）。长臂猿的呼叫则具有语义性。长臂猿有一种表示危险的叫声，其意义并不比我们叫喊“火！”更宽泛和模糊。

（8）任意性（arbitrariness）。在一个语义交际系统中，有意义的消息成分跟其意义之间的联结可以是任意的或非任意的；但是，在语言中这种联结是任意的。比如，英语单词 salt 并不是盐，dog 并不是狗；whale（鲸）形体短小却表示一种很大的物体，而 microorganism（微生物）形体较大却表示一种很小的物体。相反，图画看上去就像其所画的事物。如果一只蜜蜂要报告它发现的蜜源地很近，它会跳舞跳得很快；如果很远，就跳得很慢。“任意性”这种设计特征有任意武断这种不利之处，但是也有其巨大的优势：对于要交际的内容没有什么限制。

（9）离散性（discreteness）。尽管人类的发音器官可以发出许多不同的声音，但是任何一种语言却只使用其中很少的一部分声音；并且，这些不同的一部分声音之间的差别在功能上是绝对的（不受限制的）。比如，英语单词 pin 和 bin 对于耳朵来说只在清浊这一点上有差别。如果说话人在说 pin 时跑了音，朝着 bin 的发音方向去了，带着噪声说了 pin（或 bin），但听话人很可能基于语境仍能明白说话人说的是什么单词。这种语言的基本的、构成信号的单元中的离散性特征，不同于通过嗓音示意的方式来进行的对声音效果的使用。后者存在一种实际上是连续的程度等级，比如，人们在表示愤怒的时候会提高声音，而在表示信任时会降低声音。

（10）超越时空（displacement）。显然人类在这一点上几乎是唯一的：可以谈论在空间或时间（或两者）上距离交谈当下及地点遥远的事物。这种超越时空特征在人类近亲的发声打信号行为中无疑是缺乏的，尽管它倒是出现在蜜蜂的跳舞打信号行为中。

（11）能产性（productivity）。指语言有这样一种性能：说出以前从来没有说过或听过的话语，并且能够被操这种语言的其他人理解。如果一只长臂猿发出任何叫声，那只是一个小型的由数量有限的熟悉的叫声组成的库藏中的这一种或那一种。长臂猿的呼叫系统是封闭的。而语言是开放的，或者说是能产的，人们可以创造新的话语，把在旧的话语中熟悉的片段放在一起，按照在旧的话语中熟悉的配列模型来组装。

（12）传统传授（traditional transmission）。人类基因中带有获得语言的性能，也许还有一种很强的获得语言的内驱力；但是，任何一种语言的许多具体而微的惯例却是通过教和学来代际传授的。这种“传统传授”在长臂猿的呼叫系统或其他哺乳动物的发声信号中到底起什么或多大作用，还不得而知；尽管在一些实例中，同一种系的动物（不管它们在世界的哪个地方）的发声的一致性，在很大程度上要归因于其基因。

（13）二元构型（duality of patterning）[①]。任何语言中有意义的成分，日常语言所谓的“词”，或者语言学家所谓的“语素”，其数量都是十分庞大的。然而它们却是由一组数量较少的具有区别性的语音经过数量不多的配列方式来表示的，并且这些语音本身是不具有意义的。这种二元构型可以用英语

① 王士元（2017：9）译作“二重层级性”，其实也可以译作“双层构型”，或者“构型的两重性”。

单词 tack、cat 和 act 来说明：虽然它们在整体意义上各不相同，但是它们都是由 3 个相同的基础的不表示意义的语音经过不同的排列组成的。其实，这种二元构型就是通常所说的“双重分节”。

霍凯特（Hockett 1960：92）指出：这 13 个设计特征并不都是各不相关的，其中有一些是互相依存的。特别是，一个系统不可能是任意的或非任意的，除非它是语义的（即只有语义性交际系统，才谈得上其形式与意义之间的关系是任意的还是非任意的——袁按）；同样，一个系统不可能具有二元构型特点，除非它是语义的（即只有语义性交际系统，才谈得上其形式表示意义的方式是否是二元构型的——袁按）。并且，这个列举也不企图囊括不同种系的交际行为的所有已经发现的特征，而只包含对于语言来说显然重要的特征。

根据霍凯特（Hockett 1960：93）的图示，陆地哺乳动物以下的爬行动物、两栖动物、脊椎动物、脊索动物不采用口叫–耳听式交际，其交际系统也不具备上述 13 个设计特征。大象之类的陆地热血哺乳动物，具有社会行为，会玩耍，其交际系统具有（1）～（5）特征，即发声–耳听通道、迅速消失、完全反馈、互换性、四散传播与定向接收；猴子等灵长类动物，具有杂食性，有可动的面部肌肉，拥有双眼视觉和双手，还能够手–眼协调，其交际系统除了（1）～（5）之外，还具有（6）～（8）特征，即专门化、语义性、任意性；古猿虽然可以双足行走，但不能直立，偶尔使用工具，其交际系统除了（1）～（8）之外，还具有 2 种特征，即（9）离散性和（12）传统传授；而人类会制造和携带工具，有喉咙和软腭，具有幽默感、元音色彩和音乐，其交际系统除了（1）～（9）和（12）之外，还具有 3 种特征，即（10）超越时空、（11）能产性、（13）二元构型。

霍凯特（Hockett 1960：92）指出，这 13 个特征中的 9 种已经出现在原始古猿的口叫–耳听式交际中；并且，这 9 种特征在今天的长臂猿和人类交际系统中可以得到证实。比如，长臂猿有一打左右不同的呼叫，每一种合适的发声反应都针对一种反复出现的、生物学上重要的情境类型：发现了食物，察觉到捕食动物，性兴趣，需要母亲照顾，等等。这样，探索人类语言的起源问题，就是要确定：这种交际系统是怎样发展出另外的 4 种特征（超越时空、能产性、充分发展的传统传授、最后发展出来的二元构型）的？从而回应作者在该文章标题之下的题记中所指出的：人类是唯一能够使用抽象符号来进行交际的动物。但是，这种能力跟其他动物的交际系统共享许多特征，并且正是从这些比较原始的系统中产生出来的。

霍凯特（Hockett 1960）随后对于人类语言形成这 4 个特征的条件、生存价值等进行了假设和说明。特别是从可区分的声音刺激的数量的有限性的角度，解释了对于人类语言这种复杂的交际系统来说，二元构型是必要的。这里不再赘述。下面两节，我们将重点讨论语言符号的离散性与二元构型特征。

三、语言符号的离散性特征和向量化表示

从文献上看，关于语言符号的离散性特点，有 3 种不同的理解。第一种是上文提到的霍凯特（Hockett 1960）所谓的：构成信号的单元（即语音）在区别性功能上的绝对性（不受限制性）。比如，英语等语言，辅音的清浊具有对立功能（能够区别词的语音形式，从而区别词的意义），但是清辅音的送气与否则不具有对立功能；而汉语普通话，辅音的清浊不具有对立功能，但是清辅音的送气与否则具有对立功能。换句话说，我们只能把语流中听到的某个音素，归类到该语言中具有区别性价值的、数量有限的一套音位的某一个音位之中，不同的音位之间不具有连续性，是非此即彼的。所以，当你听到一个介于 pin 和 bin 之间的英语单词的含混发音时，你必须断定它是 pin 还是 bin。显然，霍

凯特（Hockett 1960）所谓的语言系统的离散性设计特征，主要着眼于语言的声音形式及其类别的非连续性方面。具体指语言的基本的信号单元（音素或音位）之间的区别是绝对的、类别性的，而不是连续的。比如，现代实验语音学证明：不同元音之间的差别，主要体现在第二共振峰的不同上；并且，对于第二共振峰的一定范围内的实际音素，母语听话人要么听成 [o]，要么听成 [u]，要么听成……，等等；而不会听成介于 [o] 和 [u]……之间的某种在类别上两可的元音，如此等等。推而广之，对于一个语音片段，本地听话人要么听成甲词（如 pin）、要么听成乙词（如 bin）、要么听成……，而不会听成是介于甲词与乙词……之间的某种两可的东西。

第二种理解是指连续的语流可以切分成大小不同的分析单位。比如，哈里斯（Harris 1954：158）在讨论分布分析可以发现语言成分时指出："首要的分布事实是：可以把任何语流划分（切割）成一个个部分，循此我们就可以在特定的语流中，找到某一个部分相对于其他部分的若干出现规律。这些部分是离散性成分，它们在特定的语流中有一定的分布（一组相对的位置）；并且，每一段言语都是一些成分的特定的组合。"他所谓的"语言成分"包括音位、语素、词、短语以至于句子。与此相似，中国语言学界一般从语言结构可以逐层切分为大小不同的语言单元的角度，来定义语言符号系统的离散性特点。比如，冯志伟（2007：41）对离散性的描述，大意为：连续不断的语流却是由许多离散的单元所组成的，包括组合轴上的"段落—句子—短语—词—语素—音节—音素"及其各聚合类中的离散单元。

第三种理解是自然语言处理文献上的未加明确定义的用法，大意是指语言符号在形式线索上的疏离性，即语素、词等语言单位，其在意义上的相关关系通常得不到形式上的表征。比如，即使是"移动电话"和"手机"这样的同义词语，除非你已经知道它们所指相同，否则从这两个词语的形式本身，你是无从了解它们的意义关系的。结果，语素和词等语言单元成了一个个疏离（各自独立、没有连续性）的单位。这是用"离散性"来反映语言符号的这种象征性的符号学特点。显然，语言符号的这种离散性特点，是可以从语言符号的任意性上推导出来的。前者强调了单个符号的音义结合的武断性（arbitrariness，也译作"任意性"），后者强调符号之间语义关系在形式表征上的不透明性。这就解释了为什么索绪尔的《普通语言学教程》没有专门讨论语言符号的离散性特点。因为语言符号的音义结合的任意性，规定了语言符号之间的语义关系在形式表征上的不透明性。从数据科学的角度看，文本等自然语言是一种象征性的符号数据，[①] 只在某种语言共同体的人们的大脑中具有心理上的实在性。因为，正如索绪尔（1981：4）所指出的，语言符号的音义结合，在逻辑上是任意性的；什么样的意义用什么样的声音来表达，并没有必然的理由。于是，两个语言符号（比如，语素或者词）即使在意义上有关系（比如，同义、反义、类义、上下义、蕴含等），但是在形式上也未必表现出来。这就是自然语言处理文献上所谓的自然语言符号的离散性特点，及其在数值表示上的不连续性。[②]

其实，作为对数据的数学属性的刻画，离散是跟连续相对的。比如，一个 120 名学生的班级考试，如果按百分制计分，那么，学生的成绩可以从低到高画出一条曲线，[③] 这种连续的分数是一种数值型的连续属性。如果改成 5 分制，或者"优秀、良好、及格、不及格"之类的等级制，就是一种有序的离散属性。据此，上述 3 种对语言符号的离散性的认识都有一定的道理，都揭示了语言符号非连续性的一个侧面，只是侧重点有所不同罢了。

① 关于信号数据和符号数据的区别，参考赵军等（2018：58）。

② 语素、词等语言符号不容易用连续的数值来表示，即使用词表中的 ID 号码（编号）、甚至用独热向量（one-hot vector）来表示，也不能反映语义相关的词语之间的意义联系。

③ 这条曲线一般是中间高、两头低，能够反映分数的正态分布：高分段和低分段的人数少，中间分段的人数多。

从自然语言处理的角度看，语言符号离散性特点的结果是，语言符号的形式与意义之间存在着巨大的空档。这就是所谓的语义鸿沟现象，[①] 意思是从符号的形式（声音或者文字）上提取到的信息到符号所表示的意义之间有很大的距离。这种语义鸿沟，给自然语言处理的文本表示和计算处理带来了巨大的挑战。为了机器处理的方便，通常需要把自然语言文本的符号数据转化为数值数据。由于文本的基本单元是词，因而面向数值计算的词的表示问题，成为近年来自然语言处理领域的一个热点问题；并且，形成了一种用数值表示文本实值向量形式的“词嵌入”（word embedding）技术。这种技术根据哈里斯（Harris 1954）关于“意义相似的词有相似的分布（即出现在相似的上下文）”的思想，用神经网络来从文本语料上学习和发现两个或更多单词一起出现的概率，从而将意义相似的单词聚合在一起，在向量空间中形成一个聚类；并且，赋予它们各自独立但相似的向量。2013 年，Google 团队发布了可用以提取词向量的 word2vec 工具包，其目标是理解两个或更多单词一起出现的概率，从而将具有相似意义的单词汇聚在一起，在向量空间中形成一个聚类。word2vec 本质上是一种只有两层的浅层神经网络，其中主要包含两种语言模型：连续词袋（continuous bag of words，CBOW）模型和跳字（skip-gram）模型。前者基于上下文预测当前单词，将当前单词的周围单词作为输入来产生单词作为输出；后者将单词作为输入，理解单词的意思，并将其分配给上下文来预测单词周围的单词。打一个比方，前者是玩选词填空游戏，后者是玩词语接龙游戏。但是，两者的共同点是根据本地（附近）单词的上下文来预测单词。跟其他深度学习模型一样，word2vec 可以从过去的数据和过去出现的单词中学习；进而根据过去的事件和上下文，准确地猜测一个单词的意思，就像我们理解语言的方式一样。比如，我们听到或看到“男孩”和“男人”以及“女孩”和“女人”这几个单词，如果能够理解它们的意义，就能够在它们之间建立联系。同样，word2vec 也可以形成这种连接，并且为这些单词生成向量。这些单词被紧密地放在同一个簇中，以确保机器知道这些单词意味着类似的事情。一旦给了 word2vec 一个语料库，它就会产生一个词汇表；其中，每一个单词都有一个自己的向量。这就是所谓的神经词嵌入。简单地说，这个神经词嵌入是一个用数字写的单词。[②]

由于这种词向量是连续的数值，因而可以进行加减运算。并且，这种运算的结果可以从语言学上进行解释，从而具有语言学的意义。比如，Man（男人）和 Woman（女人）之间的词向量距离跟 King（国王）和 Queen（王后）之间的距离大致相同，方向也一样。结果，用 king 这个词的向量（记作：W_{king}）减去 man 的词向量（记作：W_{man}），再加上 woman 的词向量（记作：W_{woman}），得到的与结果最近的词是 queen。也就是说，在词向量空间里，诸如 $W_{king} - W_{man} + W_{woman} \approx W_{queen}$、$W_{king} - W_{man} \approx W_{queen} - W_{woman}$ 这样的等式关系基本成立。同样的等式关系还适用于形容词的原形和最高级，比如：$W_{biggest} - W_{big} + W_{small} \approx W_{smallest}$、$W_{biggest} - W_{big} \approx W_{smallest} - W_{small}$，等等。[③] 如果说在等式 $W_{biggest} - W_{big} \approx W_{smallest} - W_{small}$ 中，等号前后的向量值大概地表示了“英语形容词最高级”之类的概念或意义，那么在下列等式 $W_{Paris} - W_{France} \approx W_{Rome} - W_{Italy}$ 中，等号前后的向量值大概地表示了“首都”之类的概念或意义。这种结果是语言学家所始料未及的。

四、语言与生物类似的二元构型和编码模型

关于语言在构型上的双层性特点，袁毓林（1998）在前贤研究及其相关文献的基础上，进行了总

① 关于语义鸿沟，参考赵军等（2018：58）。

② 以上参考 Bokka et al.（2019）§1. 5，中译本第 13 ～ 16 页。当然，中间加入了我们的理解和发挥。

③ 参考 Goldberg（2017），中译本第 122 页；详见 Mikolov et al.（2013）。

结。现在择要简述如下。

语言是一个层级系统，它通过属于纯形式的音位层次的分级组合和属于音义结合体的符号层次的分级组合，产生无穷多的形式，来表示人类交际所需的无穷多的意义。这就是人类语言信息编码的双重分节原理。双重指语言由音位和符号两个大的层级构成，分节指在音位和符号层上分别都可以由较小的单位组成较大的单位。可以表示如下：

音位 → 音节 → 音节群 ⇒ 语素 → 词 → 词组 → 句子

双重分节的编码原理使语言成为一种极为经济而有效的信息系统，通过大约 50 个最基本的语音元素的多层次组合来表示无穷的意义。

袁毓林（1998）还在相关生物学文献的基础上，综述和构想了生物遗传信息编码与人类语言信息编码在双重分节方面的类同性。

生物体也是一个层级系统，可以表示为：

细胞 → 组织 → 器官 → 系统

比层级性更有意思的是，如果把生物体的性状看作一种信息或意义，把生物性状赖以实现或表达出来的生化物质基础看作一种信号或符号，那么可以发现：生物信息的编码（即生物性状跟其生化物质基础之间的表达或实现关系）明显地遵循了双重分节的原理。比如，人体的 10 万种生物性状是由 10 万种蛋白质决定的。奇妙的是，决定人体性状的 10 万种蛋白质是仅由 20 种氨基酸通过不同的排列来造成的。几个、几十个到几百个氨基酸以一定的顺序连接起来，组成一条条长长短短的多肽链。多肽链又可以盘旋折叠，形成蛋白质的高级结构。

概略地说，氨基酸是一种分子中同时含有氨基和羧基的有机化合物，是组成蛋白质的基本单位。氨是氮和氢的化合物，化学分子式为 NH_3；氨基是氨分子中失去 1 个氢原子而形成的一价原子团（$-NH_2$）。羧基是由羰基和羟基组成的一价原子团（−COOH），羰基是由碳和氧两种原子组成的二价原子团（=C=O），羟基是由氢和氧两种原子组成的一价原子团（−OH）。也就是说，通过氢、氧、碳、氮 4 种元素在不同层次上的分级组合形成数以万计的蛋白质，从而为实现或表示数以万计的生物性状提供了足够的生化物质。这种生物信息的编码方式，可以图示于下：

原子 → 原子团 → 小分子：氨基酸 ⇒ 大分子：蛋白质（⇔ 生物性状）

4 种　　4 种　　20 种　　几万种　　几万种

如果把生化物质跟语言形式做一个类比，那么这里的原子相当于音素或音位，原子团相当于音节，分子相当于音节群；它们都是用有限的基本形式，通过分级组合的方式来形成无穷多的复杂形式，用以实现或表达无穷多的信息。

现在，生物学家已经知道，组成 DNA 大分子的核苷酸都是由糖、磷酸和碱基组成的，它们的成分基本相同；其中的糖分子是脱氧核糖，所含的碱基有 4 种：腺嘌呤（A）、胞嘧啶（C）、鸟嘌呤（G）和胸腺嘧啶（T）。因此，不同的核苷酸链（即 DNA）的差异就在于碱基排列次序的不同。正是 DNA 分子中的这种碱基的顺序决定了组成蛋白质分子的氨基酸的顺序。也就是说，遗传信息是由 4 种碱基通过一定的排列次序来编码的。这种为氨基酸在蛋白质中的排列顺序编码的 DNA 上面的碱基顺序，就是著名的遗传密码。

自然界的生物千变万化，为什么仅靠这 4 个碱基就能蕴藏和表示这么多信息，创造出如此众多的生物呢？其中很重要的一点是采用了双重分节的结构原则：不是用一个碱基直接来表示一种氨基酸，而是用三个碱基组成的三联体来表示一种氨基酸；[①] 不是用一个氨基酸分子来实现一种生物性状，而是

① 因此，这种三联体被称为“密码子”（codon）。

用多个氨基酸组成的蛋白质大分子来实现一种生物性状。有了这样一种翻番增量的结构原则，再加上一个 DNA 上可以有上亿个碱基对给这样的物质材料做基础，生物的多样性问题也就不难理解了。

既然生物分子在功能性构造方面跟自然语言有以下的平行性：

最小的信号单位：A、G、C、T4 个碱基～ 30 来个音位 / 字母

最小的信息单位：20 种氨基酸 / 核苷酸链～几千个语素 / 几万个单词

复合的信息单位：蛋白质 / 基因片段 ～ 句子

全局的信息单位：蛋白质复合体 / 基因～段落

那么，自然会让人想到：处理自然语言卓有成效的有关算法，能不能运用到生物分子领域呢？毕竟，DNA 中有 31.6 个碱基对，三联码的起止有时不好判断。也就是说，DNA 链中处处有歧义。比如：……GAACATGATTCATAGAGTACGG……。这 TGA 看起来是个终止符，而那跟它部分交接的 GAT 看起来是个天冬氨酸。于是，只能把所有可能的排列全都统计一遍。其中，所统计的 DNA（或 RNA）中长度为 K 的子序列称为 K-mer。这种子序列的频率信息，可以应用到跟基因相关的诸多任务中。比如，基因组错配检测、致病基因检测、重复序列检测、重组点位检测、蛋白质生产速率控制、基因突变或多态性鉴定、人类线粒体单倍群分类、物种分类、物种丰富度估算，等等。尽管由于每 3 个核苷酸编码一个氨基酸，即 3 个核苷酸构成一个传递生物信息的密码子，因而，K = 3 是一个具有生物学意义的取值；但是，它也会导致特殊信息的丢失。比如，……ATGTGTGTGTGTGTGTG……，其实只是在复读。而且，1 个密码子最多对应 1 个氨基酸，那只是蛋白质的“字母”。如果要理解一段基因序列的功能，显然 K 需要取更大的值。也就是说，不同的 K 值有不同的作用。

Asgari & Mofrad（2015）首次将 Word2Vec 的思想运用到蛋白质分类领域，提出了 Protein Vector（ProtVec）和 Gene Vector（GeneVec）的概念。这种做法基于蛋白质“结构决定功能”的假说：蛋白质是由氨基酸排列而成后，凭借分子内和分子间作用力形成特定的空间结构，然后发挥功能的。具体地说，氨基酸序列形成蛋白质的一级结构，由氢键导致的折叠形成蛋白质的二级结构，由多个二级结构在空间中排列后的三维结构形成蛋白质的三级结构（单条肽链），一条以上的肽链相互作用形成的蛋白质分子形成蛋白质的四级结构。这样，当氨基酸的排列相似时，蛋白质的空间结构也会相似，最终功能就会相似。如果这个理论成立，那么蛋白质分类就能参考自然语言处理上比较文本相似度的办法来寻找模型。Asgari & Mofrad（2015）据此将氨基酸片段转换为向量，即 ProtVec。为了验证 ProtVec 有意义，他们用氨基酸向量之和来表示蛋白质，并利用二分类模型“支持向量机”（SVM）对长度相近的蛋白质进行分类。结果，在 7020 个蛋白质族中，平均达到了 93% 以上的准确率。这显示出，ProtVec 确实能够较好地区分不同类型的蛋白质。特别是对于“氨基酸排列不变，但没有稳定的三维结构”的无序蛋白质，ProtVec 的分类效果很好。这可能是因为 ProtVec 关注的是蛋白质的第一、二级结构所包含的信息。基因向量 GeneVec 跟蛋白质向量 ProtVec 的使用假设基本类似，目前它们主要用于：蛋白质分类、蛋白质结构可视化、蛋白质空间结构预测、蛋白质反应机理分析、蛋白质功能预测、基序提取、基因段功能检测、功能性基因检测，等等。

值得一提的是，自然语言处理模型正在不断演进，处理效果也在不断提升。2018 年，Google 团队在 Transformer 架构的基础上，开发了预训练语言模型 BERT（Bidirectional Encoder Representation from Transformers，基于转换器的双向编码表示模型），在多项自然语言处理任务上取得了当时的最好成绩。BERT 在各种自然语言处理任务上的运用越来越广泛，以至于有人喊出“万物皆可 BERT”的口号。于是，也有人尝试把 BERT 模型引入生物分子领域，进行分子功能预测。但是，至今在效果和合

理性方面都没有出彩的表现。

总之，基于自然语言和生物分子在信息编码方面的某种相似性，自然语言处理中的一些思想和模型是可以运用到生物分子研究领域的。但是，许多神经网络模型是针对自然语言数据的结构特点而设计的，它们在生物分子等研究领域的适用性问题，尚需做进一步的研究。当然，我们乐意看到将来有朝一日，有人发现（或发明）能够同时适用于人类语言和生物编码的通用模型。①

五、人类语言交际的起源和所依托的心理平台

至少从表面上看，使用有声语言进行交往沟通是人类跟动物的显著差别。因此，反过来说，观察和研究语言可以让人类更好地认识自己的本性。平克（Pinker 2007）指出：

> 语言与人类生活有密切不可分的关系。我们不仅用语言传递信息、游说他人，我们也用它来威胁、引诱他人，当然，语言还可以用来发誓赌咒。语言反映了我们对现实的领悟，不仅如此，它还是我们留在他人心目中的活生生的印象，是把人们紧密联系在一起的纽带。我希望你也能相信这个事实：语言是通向人性的窗口。（前言，第 II 页）
>
> 仔细观察我们的语言——人们的交谈、玩笑、诅咒、法律纠纷、为婴儿取的名字，能让我们对“我们到底是谁”这个问题有更加深刻的感悟。（前言，第 I 页）

那么，自然语言这种人类沟通方式是怎样产生的？或者说，它是建立在什么样的心智或心理基础上的呢？对此，托马塞洛（Tomasello 2010）提出了下列富有启发意义的语言演化假设：

> 人类最初的沟通模式，就是比手划脚（即自然的手势——引按），以手指物是人类独有的原始沟通形式。手势这种由社会认知及社会动机的基础结构所促成的新的沟通模式，便形成了一种心理平台。不同系统、各种规约的（conventional）语言沟通模式（总共 6000 种），就奠基在这层平台之上。比手划脚是人类沟通的演化史上最关键的过渡点，体现了人类独有的社会认知与社会动机形式，这些都是后来发展规约的语言所必备的。（中译本第 2 页。引文中有少量自己的改译，与中译本文字不尽相同，如果要引用，务请核对原文。下同。——引按）

为什么这种貌似简陋不过的以手比物、指指点点，居然能够成为人类沟通的肇始和标志，并且成为约定俗成的有声语言得以奠基于其上的心理平台呢？托马塞洛（Tomasello 2010）别具洞察力地揭开了一个人们通常熟视无睹的秘密，即人类手势直指具有一种利他性的社会化功用：

> 人类以手指物这个平凡的动作，从演化论的角度来看，还有个不平凡的方面，就是它的利社会动机（prosocial motivation）。我用手指一指图书馆边上那辆好像是你前男友的自行车，从而提醒你：他可能在里面，你还要不要进去；这是因为，我认为这可能是你想知道的事情。在人以外的动物界里，这种有效传递信息的沟通相当罕见，即使是我们的近亲灵长类也不会如此……。因此，当小黑猩猩呜咽地寻找妈妈时，邻近的其他黑猩猩也都会知道。但是，即使它们知道它的妈妈在哪儿，也不会特地伸出前臂指点或比划一下。（中译本第 4 页）

你看，人兽之间，就差这么一点点：能不能伸出友爱的小手指点一下下。显然，利他性的社会动机有助于滋养人类的团结与合作精神，培养更加社会化的主体（subject）与主体交互（intersubject）意识。托马塞洛（Tomasello 2010）特别强调人类沟通的合作性质：

① 以上关于将自然语言处理中的“词向量”等运用于生物分子的介绍，根据白鹤鸽（2020）。

人类的沟通动机基本上是合作性的，我们不仅会告知对别人有帮助的事，而且当我们对别人有所求时所用的主要方法之一，就是让别人知道我渴望什么，并期待他们会主动协助。所以我若想喝杯水，可以明说我要水（告诉你我想要的），我也知道多半情况下，你主动协助的倾向（我们彼此都知道的），会把我这个告知的举动，有效地转变成充分发展的请求。

人类的沟通行为本质上是一种合作的事业，在（1）彼此假定的共同概念基础下，（2）彼此假定的合作沟通动机下，以最自然且平顺的方式进行。（中译本第 4 页）

其实，也正是这种根深蒂固的合作精神及其在交际双方之间的不言自明性，培育了一种人类的主体间性（intersubjectivity）：我们对特定情境中事物的感觉、经验、认知、理解等，并不是专属于我们个人的，而是为我们的社团群体所共享的。这构成了我们可以互相交际、互相理解的基础。正是在这种心心相印的共享空间中，我们实现了人际交往和语言沟通。①

托马塞洛（Tomasello 2010）还尝试揭示人类沟通在精神和心理方面的条件：

共同概念基础（common conceptual ground）包括共同的注意力、共有的经验、相同的文化知识。这是人类沟通必备的重要条件。（中译本第 3 ～ 4 页）

人类合作行为以共享意图（shared intentionality）为前提条件，这种活动的主体一定是复数的“我们”：大家有共同的目标、共同的意念、共有的知识、共享的信仰——而且都是在具有合作动机的情境下进行。（中译本第 5 页）

人类的合作式沟通（不管用自然的手势，还是武断的语言规约）是人类独有的合作活动之一例。它同样以共享意图为基础。共享意图的［社会认知］技巧与［利社会的］动机（与常规），构成了人类沟通的合作性的基础结构。

对于人类沟通如何从自然的手势发展到规约的语言，托马塞洛（Tomasello 2010）勾画了如下这幅宏伟的草图：

以手指物（pointing）奠基于人类自然而然地会循着别人的目光凝视物品，比划示意（pantomiming）则基于人类会自发地解读别人的动作。这种自然的反应，让手势成为由人猿的沟通进步到武断的语言沟通之间的过渡点。

在互助活动的情境下，参与者间有共同的意图与关注，并借由自然的手势沟通来协调，演化史上武断的语言规约才会随之诞生。约定俗成的语言（先是手语式的，再来才是口说的）于是依附在已知的手势上，以共享的（而且众人彼此知道是共享的）社会学习经验，取代了自然的比手划脚。这个过程当然是由人类独特的文化学习和模仿技能所促成，让他们得以用独特的有利方式，从他人也从自己的意念状态学习。同样也是在演化过程中，人类开始创造并传递文化中由不同的语法规约组成的复杂语言结构，并将繁复的信息以不同的语言结构编码为不同的类别（types），以便运用在反复出现的沟通环境中。

对人类沟通及语言所持的这种观点，可以说推翻了乔姆斯基的言论，因为人类沟通中最基础的方面，是因应一般的合作与社会互动所产生的生理调适，而纯语言的沟通，包括语法方面，则是由文化建构，并经由个别的语言社群代代相传。（中译本第 7 ～ 8 页）

人类沟通的基本的社会意图 / 动机：分享、告知、请求。（中译本第 91 页）

托马塞洛（Tomasello 2010）还构拟了下面这个基于合作的语言交际的图示（中译本第 72 页）：

① 详见 Fultner（2012：216）。

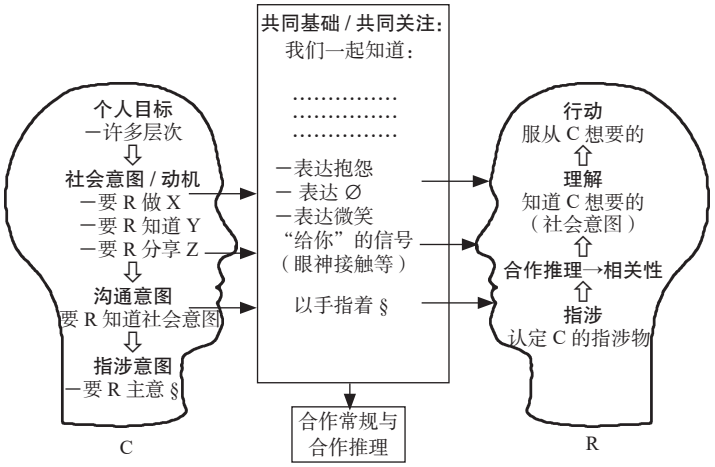

图 1　合作式人类沟通简图（C = 沟通者；R = 接收者）

这就是人类语言交际的现实生态，任何诉诸语言本身有某种繁复而自足的深层结构的幻想，都是不切实际的。要知道，目前我们对人脑的工作机理所知甚少。我们只知道不同物种的神经元数量有巨大的差别。据报道：蛔虫有 302 个神经元，果蝇有 10 万个神经元，老鼠有 7500 万个神经元，猫有 10 亿个神经元，黑猩猩有 67 亿个神经元。而人类有 860 亿个神经元，大脑神经元之间的连接约 150 万亿个。但是，人类对于自己大脑的工作机制充满困惑。神经科学家还没有办法详细解释：大脑神经元之间的电化活动交互作用，是如何变成我们脑海中的想法、情绪、记忆和推理活动的？也就是说，支撑语言生成和理解的人类神经系统是非常唯物和机械的；虽然神经元的数量极其庞大，但是神经元之间的作用方式只有简单的连接和断开两种状态。这是脑科学对语言学理论的一种刚性的约束。

六、结语：在人类社会互动和文化实践的视域下研究语言

我们相信，语言是现代人类最近 20 万年以来通过改造手势、叫声等沟通手段逐步演化出来的；虽然有声语言提高了人类交际的效率，但是在面对面交流时仍有高达 2/3 的语义依靠肢体动作、眼神表情乃至心理默契等非语言信号。[①] 因此，语言是一种不完善的“编码-解码”型信息系统，必然依赖于“示意-推理”等关联性合作机制。虽然我们赞成乔姆斯基的观点——儿童生下来头脑中并非白板一块，而是有各种先验的认知结构和语言能力，但是我们相信，在语言运用中，交际双方共享的基于经验的概念结构是认知结构和语言能力发挥作用的基础性认知资源，对于语句构成及其意义识解起着重要的作用；并且，各种认知模块之间有着广泛的交流和互动，语言官能并不是一种独立的认知系统。

正是在上述思想的启迪下，我们进行了几个基于社会互动和文化实践的语句意义识解的个案研究，来解释汉语、英语、日语和韩语中的相关现象。下面举 3 个案例。

案例一：基于接近心理和乐观原则的接近性副词及相关句式的句法语义研究。详见袁毓林（2013）和袁毓林、郑仁贞（2015）。

案例二：基于劳酬均衡原理的“白”类副词及其相关句子的语义识解研究。详见袁毓林（2014a）和朴珉娥、袁毓林（2015）。

① 出处失记，特此说明和致歉。

案例三：基于疑善信恶心理的“怀疑”类动词识解的跨语言比较研究。详见袁毓林（2014b）和朴敏浚、袁毓林（2016）。

通过这几个语义识解案例的研究，我们发现，人们对于特定词语和构式的语义理解是一个句法、词汇、语义、语用等多平面知识互动的过程；并且，期间还要援引“反通常性”的“疑善信恶”之类社会心理学原则。显然，这种语义识解是基于社会互动文化和实践经验的。

参考文献

白鹤鸽　2020 《NLP 太卷，我要去研究蛋白质了》，https://mp.weixin.qq.com/s/57S-NDBSrkpcWxcSw7z7-g。

冯志伟　2007 《论语言符号的八大特性》,《暨南大学华文学院学报》第 1 期。

朴珉娥，袁毓林　2015 《汉韩“白”类词的语义和语用特征对比研究》,《外语教学与研究》第 4 期。

朴敏浚，袁毓林　2016 《汉英日韩“怀疑”类动词的句法语义和语用对比》,《汉藏语学报》第 9 期。

索绪尔　1981 《普通语言学教程》，高名凯，译，北京：商务印书馆。

王士元　2017 《人类起源、语言的形成及其演化问题》，载张玉来主编《汉语史与汉藏语研究》第一辑，北京：中国社会科学出版社。

袁毓林　1998 《语言信息的编码和生物信息的编码之比较》,《当代语言学》第 2 期。

袁毓林　2013 《“差点儿”中的隐性否定及其语法效应》,《语言研究》第 2 期。

袁毓林　2014a 《概念驱动和句法制导的语句构成和意义识解——以“白、白白（地）”句的语义解释为例》,《中国语文》第 5 期。

袁毓林　2014b 《“怀疑”的意义引申机制和语义识解策略》,《语言研究》第 3 期。

袁毓林　2019 《为什么要给语言建造一座宫殿？——从符号系统的转喻本质看语言学的过度附魅》,《语言战略研究》第 4 期。

袁毓林，郑仁贞　2015 《汉英日韩接近性副词和相关格式的句法语义比较》,《汉日语言对比研究论丛》第 6 辑，上海：华东理工大学出版社。

赵　军，刘　康，何世柱，等　2018 《知识图谱》，北京：高等教育出版社。

Asgari, E. & M. R. K. Mofrad. 2015. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE* 10(11), 1–15.

Bokka, K. R., S. Hora, T. Jain, et al. 2019. *Deep Learning for Natural Language Processing*. Birmingham, UK: Packt Publishing. 中译本：《基于深度学习的自然语言处理》，赵鸣，曾小健，詹炜，译，2020，北京：机械工业出版社。

Fultner, B. 2012. Intersubjectivity in the lifeworld: Meaning, cognition, and affect. In A. Foolen, U. M. Luedtke, T. P. Racine, et al. (Eds.), *Moving Ourselves, Moving Others: Motion and Emotion in Intersubjectivity, Consciousness and Language,* 197–220. Amsterdam: John Benjamins.

Goldberg, Y. 2017. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers. 中译本：《基于深度学习的自然语言处理》，车万翔，郭江，张伟男，等译，北京：机械工业出版社。

Harris, S. Z. 1954. Distributional structure. *Word* 10(2–3), 146–162.

Hockett, E. C. 1960. The origin of speech. *Scientific American* 203, 88–96.

Mikolov, T., K. Chen, G. Corrado, et al. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Pinker, S. 2007. *The Stuff of Thought: Language as a Window into Human Nature*. New York, NY: Viking. 中译本：《思想本质：语言是洞察人类天性之窗》，张旭红，梅德明，译，2015，杭州：浙江人民出版社。

Tomasello, M. 2010. *Origins of Human Communication*. Cambridge, Mass.: The MIT Press. 中译本：《人类沟通的起源》，蔡雅菁，译，2016，北京：商务印书馆。

责任编辑：王　飙