第20卷 第3期　　四川师范大学学报(社会科学版)　　Vol, No. 3
1993年7月　Journal of Sichuan Normal University (Philosophy and Social Sciences)　July., 1993

论语言信息和汉语词汇系统双音节化的关系

李　恕　豪

在古代汉语的词汇系统中，单音词占绝对优势，而在现代汉语中，却以双音词为主，这说明双音节化是汉语词汇发展的一条重要规律。与此同时，汉语的音位系统经历了一个由繁到简的过程。对于这两条规律之间的关系，前人已经作了不少论述。我们认为，使用信息论的观点可以更加合理地说明这两条规律之间的内在联系，因为语言就其本质来说，就是一种信息。陈原说："信息是由物理载体(按一定方式排列的信号序列)与语义两者构成的统一体。在这个意义上，语言本身也可以称为信息"[①]。

一

语言是人类特有的社会现象，交际即传递信息是语言最重要的功能。随着社会的发展，新的事物和概念不断地产生，这就要求人们在交际中不断地创造和使用新的词语。在语言的发展过程中，既有新词的产生，也有旧词的消亡，但从总体上看，词语增加的趋势是主要的。词汇量的增加实际上就是词汇系统所提供的信息量的加大。

但是，假如汉语音位的总数不变，词语仍然采用原来的以单音节为主的形式，那么，词汇系统所提供的信息量就必然会受到很大的局限。我们知道，语言是一种两层的分层装置，它的底层是一套音位，这一层就叫做音位层。音位一般只有几十个。语言中的上层是音义结合的符号以及符号与符号相组合的序列。这一层叫符号层，其中又可以分为若干级。第一级是由音位组合成的音节。由于在古代汉语中单音词占优势，因此这一级一般也就是词；现代汉语以双音词为主，这一级在多数情况下是语素。第二级是音节与音节组合成的符号。这一级在古汉语中一般是词组，在现代汉语中一般是词。再往上的组合就是更复杂的词组或句子了。由音位层组合为符号层，以及在符号层内部由第一级组合为第二级，由第二级组合成第三级，这就是语言层级装置的运转情况，其中高一级的单位比低一级的单位在数量上要多得多。一般说来，语言中音位层上的单位只有几十个，而符号层中第一级的单位则上千，第二级可以上万，第三级以上就无限了。由此可见，古代汉语的词一般处在符号层的第一级上，可以用来区别语义的符号只有几千个。而现代汉语的词大多处在符号的第二级上，可以用来区别语义的符号在数万个以上。在汉语中，如果仍然使用单音词来表示新的意义，创造新词，那么语言符号中的形式部分语音，其选择余地就必然很小，不能满足语言符号中的内容部分意义日益增长的需要。在原有的构词模式下，为表达新的意义，汉语的词汇系统可以采用以下两种方式。

一是增加同音词。大量的同音词必然会增加歧义，影响语言交际的效果。当然，在多数情况下，歧义可以被语境(context)所排除，但是当同音词多到一定程度时，语境有时就不太容易排除这些歧义了。从信息论的观点看，同音词的出现必然会使词所负载的信息量(information content)减少。这是因为词所提供的信息量与词在一定的语境中所出现的概率(Probability of occurence)成仅比。如果在某个语境中，只可能出现 a，b 两个词，而且它们出现的可能性一样大，那么 a 的概率 Pa 和 b 的概率 Pb 都等于 1/2。如果在某个语境中有 n 个出现概率一样的词 $a_1, a_2, a_3, \ldots a_n$，那么，每一个词的概率都是 1/n。概率相同的词所负载的信息量完全一样。如果在某个语境中只有 a，b 两个词，当 a 平均出现两次时，b 平均出现一次，那么，Pa＝2/3，Pb＝1/3。由于词的出现概率与词的信息量成反比，因此，a 的信息量必然小于 b 的信息量。这就是信息论的基本原理。信息量一般以“比特”(bit)为单位[②]。出现概率为 1/2 的单位所负载的信息量为 1 比特，出现概率为 1/4 的单位所负载的信息量为 2 比特，出现概率为 1/8 的单位的信息量为 3 比特。也就是说，信息量等于以 2 为底数的出现概率分母数字的对数，即 $\log_2 2=1$，$\log_2 4=2$，$\log_2 8=3$，……。可见，同音词的增加，使同一种声音形式出现在同一语境中的可能性相应增加，加大了出现概率，造成词的信息量的减少，因而不利于语言交际中的信息传递。

二是使词由单义向多义发展。具体步骤是一个词在原有意义的基础上，通过各种方式引申出新的意义。例如，“牧”本是一个单义词，根据《说文解字》的解释，意思是“养牛人”。这个最早能考察的意义我们称之为“本义”。由这一本义可引申出下列各个意义：

养牛人 ┌牧人
　　└牧牛——畜牧——牧场——远郊
　　　　└统治——地方官

例如，《诗·小雅·无羊》：“尔牧来思，何蓑何笠。”《孟子·公孙丑下》：“今有受人之牛羊而为之牧之者。”《诗·邶风·静女》：“自牧归荑，洵美且异。”《尔雅·释地》：“邑外谓之郊，郊外谓之牧。”《管子·牧民》：“凡有地牧民者，务在四时，守在仓廪。”《后汉书·刘焉传》：“太仆黄琬为豫州牧，宗正刘虞为幽州牧。”引申可进一步分为直接引申和间接引申。多次间接引申以后的意义和本义相去很远，往往看不出它们之间意义上的联系，清代学者戴震说，本义“展转引申为他义，有远，有近，有似远义实相因。”[③]上面“牧”所有的“养牛人”和“地方官”的关系就是这样。如果人们不知道它们在语义上的联系，也可以把它们看作是同音词。由此可见，多义词与同音词一样，能够使出现于同一语境中相同的声音的概率加大，从而使词的信息量减少，因而不利于语言的交际。

总之，在原有单音节的构词模式下，汉语要表达新的意义，无论是采用创造新的同音词还是通过词义的引申方式，结果都会导致词的信息量的减少。

于是，汉语不得不采用双音节的构词方式来解决交际内容的扩大与词语的信息量日益减少的矛盾。双音节是音节与音节的再组合，最初是词与词的组合，应当看成是词组，后来凝固下来才逐渐变成了一个词，这时，原来的词就成为语素了。这个过程，就叫做“粘合”。索绪尔说：“粘合是指两个或者几个原来分开的但常在句子内部的句段里相遇的要素互相熔合成为一个绝对的或者难于分析的单位。”[④]无论是上述的哪一种情况，双音词都处在符号层的第二级上面，其数量比第一级多得多，这有利于语言中同音词和多义词的减少，从而导致词语信息量的增加。

例如，“白”是一个多义词，主要表示“黑”的反义。由这个意义引申出：(1)清楚；(2)没有效果；(3)无代价或报偿。“白”也是一个同音词，因为它还可以表示：(1)说明，陈述；(2)戏曲唱腔以外的

语句。当我们使用“白”构成双音词“白色”、“明白”、“白费”、“表白”、“道白”的时候,“白”原来的各个意义就以新的双音形式出现,不再使用共同的语音形式,避开了同音和歧义,从而使词语所负载的信息量得到了增加。

双音词对语义的选择和确定是通过语言中有意义的单位的组合来实现的。如上所述,一个单独的“白”至少有五个意义。当“白”与“色”组合时,只能选择与“黑”相反的这一意义,而排斥其他的四个意义;而当“明”和“白”组合时,“白”只能选择“清楚”这一意义。同样,在“白费”、“表白”、“道白”中,“白”的意义的选择也要受它与“费”、“表”、“道”组合时的语义的限制。我们也可以把出现在“白”的前面或后面的“色”、“明”、“费”、“表”、“道”看作是“白”的各个意义所出现的语境或上下文。当“白”出现在“色”的前面时,只允许“白”表示与“黑”相反的这一意义出现,而不允许其他四个意义出现,当“白”出现在“明”后面时,只允许“白”的“清楚”的意义出现,而排斥其他意义,余此类推。无论是提供组合成分或者提供语境和上下文的制约,都会突破原来词语的单音节模式。

由此可见,汉语的词汇系统从单音节向双音节转化的原因,主要在于社会的发展,交际的增加,从而要求词语具有更多的信息内容。正如荀子所说:“单足以喻则单,单不足以喻则兼”[5]。

二

从单音节到双音节,从组合的层次上讲,是进入到了更高的一级,而从音位与音位的组合上看,则是增加了组合长度(syntagmatic length)。组合长度的增加,必然导致音位系统的简化。

在任何语言中,组合关系和聚合关系是最重要的两种关系,它们有密切的联系。由较低层次的聚合单位所组成的较高层次单位的组合长度,与聚合群中各个单位的数量成反比[6]。我们假定在一个系统中只存在着两个较低层次的聚合单位0和1,在另一个系统中有八个较低层次的聚合单位,就是从0到7的数字。我们还要提出另一条假设,就是在这两个系统中,所有的聚合单位都可以不受限制地与其他聚合单位任意组合。如果使用第一个两位系统的单位,通过组合构成八个不同的较高层次的单位,那么每个单位必须得有三个数字,它们的组合长度为三:000、001、010、011、100、101、110、111。如果使用八位系统的单位,组成八个单位,每个单位只需要一个数字,即0、1、2、3、4、5、6、7,其组合长度为一。如果要组成64个较高的单位,两位系统的组合长度必须是六位数,即$64=2^6$,八位系统的组合长度则仅仅是两位数,即$64=8^2$。于是可以得到下面这个公式:$N=n^m$。其中,N表示较高单位的数量,n表示聚合群中单位的数量,m表示组合的长度。这个公式也可以写成$m=\log_n N$。

当组合长度增加一倍时,情况将是这样:

$$m=\log_n N \Rightarrow 2m=2\log_n N=\log_n N^2=\log \sqrt{n}\, N$$

这时,如果聚合群中单位的数量不变,组合后的较高单位的数量就是原来数量的平方($2m=\log_n N^2$)。如果较高单位的数量保持不变,聚合群中单位的数量则是原有数量的平方根($2m=\log \sqrt{n}\, N$)。可见,组合的长度与组合后较高单位的数量成正比,而与聚合群中的单位数量成反比。

汉语词语由单音节转化为双音节,音位的组合长度增加了一倍,即$m\Rightarrow 2m$,这就使新词的产生在语音形式上拥有比过去多得多的选择余地($N\Rightarrow N^2$)。假如词汇的总量(N)仍然保持不变,音位的数目(n)必然会大大减少($n\Rightarrow \sqrt{n}$)。

实际上，语言中所发生的情况并不是这样简单。

首先，在一种语言中，不是任何一个音位都可以不受限制地与另外的音位组合，例如舌根辅音在现代汉语中便不能与齐齿呼、撮口呼的韵母相组合。有些音位不能出现在某些组合的位置上，例如舌根鼻音 ng 在现代汉语中便不能出现在一个音节的开头，而双唇鼻音 m 则不能出现在一个音节的末尾。因此，音位与音位组合而形成的单位在数量上必然比上述情况要少得多。这样就可以得到下面这样一个公式：

$$N=P_1\times P_2\times P_3\cdots\cdots P_m$$

其中，N 仍然表示组合以后较高单位的数量，m 表示组合的长度，P_1 表示能出现在组合的第一位置上聚合单位的数量。在现代汉语中，P_1 可以是/n/、/s/、/t/等等。P_2、P_3、P_4 则分别表示能出现在组合的第二、第三、第四个位置上聚合单位的数量。在现代汉语中，P_2 可以是/i/、/u/、/y/等，P_3 可以是/a/、/o/、/u/等等，P_4 可以是/i/、/u/、/n/等等。这样，便可以组成/liau/这类的音节（语素或词）。但是，并非凡是能够出现在 P_1 位置上的音位，都可以与出现在 P_2 位置上的音位相组合，例如现代汉语中的/s/便不能与/y/组合。P_2 同 P_3，P_3 同 P_4 也有这种情况。我们假设不能组合的（或不能接受的）高一层次的单位有 x 个，因此，上述公式应当修正为：

$$N=P_1\times P_2\times P_3\cdots\cdots P_m-x$$

虽然如此，语言中音位的数量与组合长度之间仍然存在着反比关系。

其次，语言中各个词的组合长度并不完全相等。现代汉语虽然以双音词为主，但也存在着大量的单音词，而且单音词的使用频率还相当高，各个单音词中所包含的音位的数量也不相同。但为了比较方便地说明问题，我们仍然以上述公式为出发点，作进一步的阐述。

假如汉语的词汇系统在古代是绝对的单音节，在现代是绝对的双音节，而词汇总量又保持不变，根据上述公式，音位的数量将变成原有音位数量的平方根。这种减少是相当剧烈的。但是古代汉语并非绝对的单音节，现代汉语也不是绝对的双音节，因而在汉语的词汇系统中，音位组合长度的增加不到一倍[⑦]。另外，随着社会的发展，词汇总量应当逐步增加，不可能保持不变。因此，汉语音位的数目不会以平方根倍数那样剧烈的程度减少，但无论如何，汉语词汇的双音节化给音位的缩减施加了很大的压力。

由此可见，社会的进步和交际的发展要求语言必须扩大词汇量（N），但如果仍然采用单音节的构词方式，只能使词所担负的信息内容减少，这不利于交际。为了增加信息量，汉语的词汇系统不得不以单音节为主的形式向双音节为主的形式转化（$m\Rightarrow 2m$），从而引起汉语音位数量（n）的减少。

三

汉语词汇系统的双音节化，还有利于增加剩余信息（redundancy），克服噪音（noise）。所谓剩余信息，是指超过传递最小需要量的信息；而噪音，则是指信息传递时的干扰。这种干扰能够使信息失真和减少。由于剩余信息可以排除噪音的干扰，因此在一般情况下，为了保证理解，总要给出比实际需要更多的信息。可见，语言中有一定的剩余信息是完全必要的。

例如，在古代汉语中，“道”的本义是道路。《说文》：“道，所行道也。”通过引申形成的多义至少有：（1）途径；（2）方法；（3）方向；（4）技艺；（5）规律；（6）思想，学说；（7）道德。另外，“道”还有一些

同音词。如前所述,词的多义和同音现象必然会减少词的信息量,因此,古代汉语中单音节的"道"在交际中就处于信息量日益减少的过程之中。我们选择一个与"道"的本义相同的"路",来与"道"共同组成一个复合词"道路"。这时,双音节的"道路"一词不仅阻止了"道"所负载的信息量日益减少的趋势,而且还使这个新词有了剩余信息。因为"道"与"路"的意义相同,这两个音节的相继出现,就意义而言,是同义重复。《诗·秦风·蒹葭》:"道阻且长",在《古诗·行行重行行》中则是"道路阻且长"。从对于意义的理解来看,"道路"比"道"更为清楚明晰。

剩余信息虽然能够克服噪音的干扰,有利于信息的传递,但并不是剩余信息越多越好。在交际中,我们只需要适当的剩余信息就可以了,如果过量,就不合乎语言交际中经济简便的原则[8]。

现在我们回到我们前面的 $m=\log nN$ 的公式中去。如果组合的长度增加一倍,则是:

$$2m=2\log nN=\log nN^2=\log \sqrt{n}\,N$$

也就是说,这个时候聚合群中的单位是原有平方根的倍数($\sqrt{n}$)。假如原来聚合群的单位是64个,到这时只需要8个($\sqrt{64}=8$)单位便可以提供组合长度为m时64个聚合单位所提供的信息量。这是惊人的减少。从另外一个方面说,假如这个时候的聚合单位(n)仍然不变,那么,它可能提供的信息比原来的信息要多得多(应该多出 N^2-N;如果 $N=1000$,则 $N^2-N=999000$)。这大大地超出了必要的剩余信息。汉语词汇系统的双音节化,总的说来,组合长度并没有增加到一倍,音节内部各个音位的组合也要受到种种限制,语素和语素(一般体现为音节与音节)的再组合更要受到语义的制约,而从词的数量上看,总的趋势是增加的,并不维持原来的数目。总之,它远远不能达到上述公式那样极端的程度。但有一点是相似的,就是假如汉语音位的数量不变,双音节组合后所提供的剩余信息将大大超过排除噪音干扰的需要,为了符合经济的原则,这就必须缩减音位的数量。这就是汉语在发展的过程中音位数目减少的根本原因。

四

汉语中个别的双音词虽然早在殷商就已经开始出现,但汉语词汇复音节化的发生却应当在西周、春秋时代。在这之后,汉语中双音词的增加是相当迅速的[9]。双音词的大量出现,必然引起汉语音节结构内部的变化。黄志强、杨剑桥指出,在上古汉语中,音节结构比较复杂,不仅存在着大量的复辅音声母,而且在不少的阴声韵字中也具有辅音韵尾。那时汉语音节的一般结构公式是:

起首辅音+基本辅音+元音+辅音韵尾

从西周、春秋开始,直到东汉、魏晋,是汉语语音大大减化的时期。其具体表现是复辅音声母逐渐丧失殆尽,具有辅音韵尾的阴声韵字也脱落韵尾[10]。我们认为,复辅音声母的失落大致完成于东汉末年。因为反切起源于汉魏之际,假如在汉语中音节结构中仍然存在着大量的复辅音声母,那么,反切上字必然非常复杂,因而不利于反切的拼音方法。部分阴声韵所带有的辅音韵尾,其失落的时间可能要晚一些。黄志强、杨剑桥认为它们的完全丢失,大致在齐梁之际[11]。复辅音声母的丢失以及阴声韵的部分辅音韵尾的失落,在音位的数量方面一般并没有减少,但是在音位的组合长度方面,却是一种缩减。汉语音节结构的简化,音位组合长度的压缩,促使了同音词的增加,从而有利于减少语言交际中过多的剩余信息。由此可见,由双音节化而带来的过多的剩余信

息，是造成汉语音节内部结构的简化，同音词增加的重要原因。

魏晋以后，汉语复音化的速度加快，尤其到了唐代，“汉语双音节非常丰富了”，“它们是出乎意外的多，而且多数一直传到现代，丰富了现代汉语的词汇”⑫。大量的双音词给汉语带来了大大超过需要的剩余信息。这时，由于过去的复辅音声母和阴声韵的部分辅音韵尾已经丢失，因此，汉语一般只能使用缩减音位数目的办法来解决这一问题。王力说：“现在我们还不十分了解唐末和宋代的实际语音情况，但是有种种迹象使我们相信从八世纪起，实际语音要比《切韵》系统简化了一倍。”⑬从宋到元，双音词进一步增加。这时，汉语不仅继续沿着减少音位的方向发展，同时还进一步压缩音位组合的长度，这就是入声韵尾辅音的脱落。这种脱落，使入声韵与阴声韵合流，这又增加了一大批同音词，从而使汉语的语音系统更进一步简化。王力在论述这一时期的情况时说：“到了《中原音韵》时代（十四世纪）又比第八世纪的实际语音简化一倍以上。”⑭可以认为，直到今天，汉语音位系统的简化过程，仍然处在继续之中，因为汉语词汇的双音节化到今天仍然没有停止。

总而言之，汉语词汇系统由单音节向双音节转化的最主要的原因，在于社会的进步所引起的交际需要的扩大。双音词增加了音位的组合长度，从而提供了大大超出实际需要的信息。为了达到新的平衡，汉语起初是缩减二个音节内部音位组合的长度，后来则采用减少音位的方法。可见，无论是汉语的复音节化，还是汉语语音系统的简化，其根本的推动力量都是社会进步所引起的交际需要的增加。那种把汉语词汇双音化的原因简单地归结于对汉语语音系统的不自觉的简化的一种补偿，其所持的理由是不够充分的。

注释：

①《社会语言学》第 68 页，学林出版社，1983 年。

②bit 是“二进位数”binary digit 的英文缩写。

③《戴东原集》卷三《答江慎修先生论小学书》。

④《普通语言学教程》汉译本第 248 页，商务印书馆，1985 年。

⑤《荀子·正名》。

⑥确切地说，应该是成一种反函数。见赵元任《语言问题》第 226 页，商务印书馆，1980 年第 1 版。

⑦有人作过统计，在现代汉语中，词和字的比例大约是 1.7∶1。古代汉语中的一个字大致相当个一个词，但也存在着一些双音节的联绵词和复合词。因此，从古到今，汉语词汇系统中，音位组合长度的增加，不会超过 0.7。

⑧关于剩余信息在语言交际中的分寸问题，可参看钱冠连的《语言冗余信息的容忍度》，《现代外语》1986 年 3 期。

⑨⑩⑪见黄志强、杨剑桥《论汉语词汇双音节化的原因》，《复旦学报》1990 年第 1 期。

⑫王力《汉语史稿》中册第 345 页，中华书局，1980 年新 1 版。

⑬⑭同⑫，第 342 页。