《暨南大学华文学院学报》，2007 年，第 1 期, 37-50

论语言符号的八大特性[1]

冯志伟

（教育部语言文字应用研究所，北京，100010）

[关键词]： 语言符号，任意性，层次性，非单元性，离散性，递归性，随机性，冗余性，模糊性。

[摘　要] 索绪尔曾经提出语言符号有任意性、线条性两个重要特性，自然语言处理的发展，使我们对于语言符号的这些特性的认识和理解更为丰富、更为深刻了。这是自然语言处理对传统的理论语言学提出的挑战。本文针对信息时代语言学的新发展，分析了索绪尔“线条性”的不足，并提出了语言符号应当具有层次性、非单元性、离散性、递归性、随机性、冗余性、模糊性等七个十分重要的特性。这七个特性再加上索绪尔提出的语言符号的任意性，构成了我们对语言符号的特性的新认识，即语言符号具有任意性、层次性、非单元性、离散性、递归性、随机性、冗余性、模糊性等共八个特性。我们应当修正索绪尔对于语言符号特性研究的旧理论，而代之以反映当前人类对自然语言符号认识水平的新理论。

[中图分类号]H0-05　[文献标识码]A　[文章编号] 1671-5306（2007）01-0037-14

On Eight Characteristics of the Linguistic Sign

FENG Zhi-wei

(Institute of Applied Linguistics, Ministry of Education, Beijing 100010, China)

Key words: linguisitic sign, arbitrariness, stratification, non-mononialness, discreteness, recursiveness, stochastic, redundancy, fuzziness.

Abstract: Ferdinand de Saussure has proposed two characteristics of the linguistic sing. One is arbitrariness, the other is the linear nature of the signifier. The development of natural language processing not only enriches our knowledge and understanding of the linguistic sign but alos is a challenge for traditional theoretical linguistics. So a new theory of nature of the linguistic sign should be advanced. Besides the arbitrary nature proposed by Saussure, the author of this paper puts forward another 7 characteristics of the linguistic sign: stratification, non-monomialness, discreteness, recursiveness, stochastic, redundancy, fuzziness. In this way, we hope the old theory of the nature of the linguistic sign proposed by Saussure can be revised..

[1][作者简介]冯志伟(1939-)云南昆明人，教育部语言文字应用研究所研究员，中国传媒大学博士生导师，韩国科学技术院电子工程与计算机科学系教授，《中国语文》、《语言科学》、《语言文字应用》、《国际语料库语言学杂志 (International Journal of Corpus Linguistics)》（英文版）编委。主要研究方向为计算语言学和应用语言学，发表中外文专著 20 多部，论文 200 余篇。

一、问题的提出

自然语言处理(Natural Language Processing，简称 NLP)的研究已经有 50 多年的历史了，这门新兴的语言学科不仅在应用方面取得了巨大的成绩，而且还在理论方面强烈地冲击着索绪尔（De Saussure）以来的普通语言学基本理论，并以大量的新的事实和研究成果，严峻地考验着这些基本理论。这是自然语言处理对传统的理论语言学提出的挑战。可以预见，自然语言处理的进一步发展，必定会使我们对自然语言的本质获得新的认识，从而推动理论语言学的进一步发展（冯志伟，2007）[2]。

我们这里只是谈一谈关于语言符号的特性的问题。自然语言处理的发展，使我们了解到语言符号的许多重要特性，这些语言符号新特性的发现，从新的侧面进一步丰富了我们对语言符号本质的认识。

我在 1992 年发表的《计算语言学对理论语言学的挑战》[3]一文中曾经提到，在信息时代，自然语言的符号表现出一些我们过去未曾特别注意到的特点，这些特点加深了我们对于语言符号本质的认识。一些读者对于这个问题颇有兴趣，希望我进一步说明。这确实是一个值得进一步探讨的问题，本文也就是为了这个目的而写成的，希望能够产生抛砖引玉的效果。

索绪尔在他的《普通语言学教程》（索绪尔,1980）[4]一书中，曾提出语言符号具有如下两个重要的特性：

（1）符号的任意性：语言符号的能指和所指联系是任意的。索绪尔认为，符号任意性的原则“支配着整个语言学，它的后果是不胜枚举的，人们经过许多周折才发现它们，同时也发现了这个原则是头等重要的”。

（2）能指的线条性：索绪尔指出，语言的能指属于听觉的性质，只在时间上展开，而且具有借自时间的特征：（1）它体现为一个长度；（2）这长度只能在一个向度上测定，它是一条直线。索绪尔认为：“这是一个似乎为常人所忽视的基本原则，它的后果是数之不尽的，它的重要性与符号任意性的规律不相上下，语言的整个机构都取决于它。”

索绪尔提出的语言符号的这两个特性，当然是十分重要的。然而，索绪尔以后现代语言学的发展，特别是电子计算机出现以后自然语言处理的发展，严峻地考验着索绪尔的理论。

在我们看来，索绪尔提出的语言符号的任意性这一特征是无可非议的，但是，他提出的语言符号的第二个特征 —— 能指的线条性就未必是正确的了。因为新的研究结果表明，语言的能指并不只是线条性的东西。英国著名语言学家弗斯（J. K. Firth）提出“跨音段论”（prosodic）(Firth, 1948)[5]，他认为，在一种语言里，区别性语音特征不能都归纳在一个音段位置上，例如，语调就不是处于一个音段位置上，而是处于前后相续的线条性的音段之外，笼罩着或管领着整个句子的东西。如果我们把语调这样的跨音段成分算进去，语言的能指就不宜看作是线条性的东西，而应该看作是立体性的东西了。

索绪尔是一个天才的语言学家，他是名副其实的现代语言学的奠基人，他的语言学说，是语言学史上哥白尼式的革命，对于现代语言学的发展有着深远的影响。现代语言学的每一个领域，每一个流派，都直接或间接地受到了索绪尔语言学说的影响。他所说的语言符号的上述两个特性，是在当时的语言学和自然科学发展的水平下提出来的。在索绪尔的时代，还没有电子计算机，自然语言处理这样的新兴学科还远远没有形成，语言学主要是与语言教学、

[2] 冯志伟，《机器翻译今昔谈》，语文出版社，2007 年。
[3] 冯志伟，计算语言学对理论语言学的挑战，《语言文字应用》，1992 年，第 1 期。
[4] 索绪尔，《普通语言学教程》，中译本，商务印书馆。1980 年。
[5] J. R. Firth, Sounds and Prosodies, 1948.

文学、历史、考古学等学科有联系。在这种情况下，索绪尔当然不可能提出那些只有在电子计算机时代才能揭示出来的语言符号的新特点。

随着电子计算机的出现和发展，特别是在自然语言处理出现之后，普通语言学的理论也应该相应地发展。我们不能墨守成规，满足于旧有的结论，而应该站在前辈学者的双肩上，高瞻远瞩，吸取自然语言处理的新成果，从新的角度，用新的眼光，以新的方法来研究语言这个极为复杂的符号系统（Turing, 1950）[6]。正是基于这样的认识，我们觉得，语言符号除了索绪尔所指出的那两个不尽完善的特点之外，还有着如下七个十分引人注目的特点。

二、语言符号的另外七个引人注目的特点

1. 语言符号的层次性

前面说过，索绪尔关于语言符号线条性的观点，早就受到了语言研究新成果的严峻挑战。弗斯的“跨音段论”已证明，语言符号并不是线条性的东西，而是立体性的东西。

弗斯的“跨音段论”只限于音位学方面。其实，在语言的其它方面，语言符号也不仅仅是线条性的，而是立体性的东西。所谓立体性，就是说，语言符号具有分层结构，即层次性。

语言符号的层次性在句子结构方面表现得特别明显。

美国描写语言学派的语言学家早就指出，英语的“The old men and women stayed at home”（年老的男人和女人留在家里）这句话是有歧义的。如果我们把这一句话说给一些人听，很可能有的听话人会认为它的意思是“年老的男人和所有的女人（不论年龄大小）留在家里”，另一些听话人会认为它的意思是“所有年老的男人和所有年老的女人留在家里”，还有的听话人干脆不能做出决定，处于模棱两可的状态。

事实上，“old men and women”这个名词短语根据意义的不同有两种不同的层次结构。如果注意到层次的不同，那么，这种意义上两可的情况就可以得到解释。

一种层次结构是

old men and women

这时，这个名词短语的意义是：“年老的男人和所有的女人”。

另一种层次结构是

old men and women

这时，这个名词短语的意义是：“所有年老的男人和所有年老的女人”。

一般地说，如果要判断两个语言片段 A=a_1 a_2 ... a_n 和 B=b_1 b_2 ... b_m 是否具有同一性，至少应该满足三个条件：

① A 和 B 中对应的词形相同，词数相同。即有 $a_1 = b_1$, $a_2 = b_2$, ... , $a_n = b_m$, 且 n=m.

② A 和 B 中的词序相同。即：如果有 a_1==> a_2 , ... , a_{n-1} ==> a_n , 那么，则有 b_1 ==> b_2 , ... , b_{m-1} ==> b_m . 其中，“==>”表示前于关系。

③ A 和 B 中各个词之间的层次结构相同。

索绪尔主张语言符号具有线条性，他只看到了第 1 条和第 2 条，而没有看到第 3 条，这是他的局限性。今天，我们看到了第 3 条，发现了语言符号的层次性，应该说是一个很大的进步。

[6] A. M. Turing, Can a machine think? *Mind*, 50, 1950.

在自然语言处理中，常采用树形图来表示语言符号的层次关系。自然语言处理的理论认为，任何一个句子的线性序列的表层之下，都隐藏着一个层次分明的树形图。当一个句子的线性序列之下隐藏着两个或两个以上的树形图时，这个句子就会产生歧义，就会得到不同的解释。

树形图由结点和连接结点的枝组成。树形图的各个结点之间，有两种关系值得注意：一种是支配关系，它反映了上下结点之间的先辈和后裔的关系，一种是前于关系，它反映了左右结点之间前位和后位的关系。语言符号的线条性只是反映了前于关系，而没有反映支配关系，当然就有很大的局限。

树形图与自然语言处理中广为应用的短语结构语法有着明显的对应关系。乔姆斯基的短语结构语法，既能描述自然语言，也能描述程序设计语言，这种语法已经成为形式语言理论的重要研究内容。在形式语言理论中建立的短语结构语法与树形图之间的对应和联系，正是基于对语言符号层次性的认识的基础之上的。短语结构语法和树形图被广泛地使用于自然语言处理中，几乎每一个自然语言处理研究者天天都要与短语结构语法和树形图打交道，天天都要研究语言符号的层次关系。自然语言处理的发展，进一步加深了我们对于语言符号的层次性的认识，语言符号的层次性，确实是一个比索绪尔提出的语言符号的线条性更为深刻的特性。

2. 语言符号的非单元性

基于对语言符号的层次性认识基础之上的短语结构语法，在机器翻译和自然语言理解的研究中很快就暴露出了它的不少缺陷。这种语法分析能力不高，分析时难于处理歧义等自然语言中普遍存在的问题，常常捉襟见肘，进退维谷；这种语法生成能力过强，往往会生成许多歧义的句子或不合语法的句子，使人误入迷津，扑朔迷离。后来，自然语言处理研究者发现，引起这些缺陷的症结在于，短语结构语法是采用单标记来描述语言符号的，它把语言符号看成是不可分割的原子式的单元；如果把语言符号看成是可以分割的非单元性的东西，采用多标记函数或者复杂特征来描述，便可以从根本上克服短语结构语法的上述缺陷，大大地改善短语结构语法的功能，提高它过弱的分析能力，限制它过强的生成能力。这样，便提出了语言符号的非单元性问题。

其实，索绪尔早就认识到了语言符号的这种非单元性。他在《普通语言学教程》中指出：“语言可以说是一种只有复杂项的代数”。他举出德语中名词数的变化来说明这个论点。德语中名词 Nacht（夜，单数）：Nächte（夜，复数）这个语法事实可以用 a/b 这个符号来代表，但是，其中的 a、b 都不是简单项而是复杂项，它们分别从属于一定的系统之下。Nacht 有名词、阴性、单数、主格等特征，它的主要元音为 a，Nächte 有名词、阴性、复数、主格等特征，它的主要元音为 ä，结尾加了 e，ch 的读音从 /x/ 变为 /ç/。 这样，就可以形成许多对立，所以叫做复杂项。每一个符号独立地看，可以认为是简单项，但是从整体来看，则都是复杂项。索绪尔指出：“语言的实际情况使我们无论从哪一方面去进行研究，都找不到简单的东西；随时随地都是这种相互制约的各项要素的复杂平衡。” 索绪尔在这里所说的“复杂项”，指的正是语言符号的非单元性。

早在 1936 年，美国语言学家雅可布逊（R. Jakobson）在比利时的根特城举行的第三届国际语音学会议上，提出了能否以对分法为基础来分解元音、辅音等音位的问题。1951 年，他在与范特（M. Fant）、哈勒（M. Halle）等语音学家合写的论文《语音分析初探》中，提出了对分法理论以及区别特征学说。他们认为，一切的音（无论元音或是辅音）都是可分的，可以根据它们的生理的或声学的特性，用对分法分成一对一对的“最小对立体”（minimum pairs）。例如，元音的舌位有“高-低”的对立，辅音的发音方法有“清-浊”的对立。他们

把这些最小对立体归结为 12 对区别特征（ distinctive features ），并且指出，世界上各种语言都可以用这 12 对区别特征加以描述。这样，过去一直被认为是不可分的单元性的元音、辅音就变成由若干区别特征组合而成的、非单元性的结构体了。这种区别特征理论已成为现代语音学进行音位分析的基础。任何一个音位都可以用区别特征的集合来加以描述。如某一个音位具有二项对立中的前项特征，记以正号“+”，具有二项对立中的后项特征，记以负号“-”，就可以做成一个矩阵表，作为对每一个音位的区别特征集合的描述。这种音位理论，已经在语音自动识别和合成的研究中得到应用，证明是行之有效的。这是语言符号非单元性的有力证明。

雅可布逊曾提到，他之所以提出音位对分理论，是受到了现代物理学的影响所致。他在《语音实体的辨识》一文中写道：“语音学分析及其得出的、不能再行分解的音位特征的概念，同现代物理学的研究成果有惊人的相似之处，物理学也正表明，物质具有粒子状结构，因为它们是由基本粒子构成的。”（Jakobson, 1949）[7]

物理学中关于物质具有粒子结构的观点，音位学中关于音位由 12 对基本的区别特征组合而成的观点，自然语言处理中关于语言符号由多个标记组合而成的观点，它们之间是何等的相似！客观世界中存在着的这种相似现象，说明了这些现象之间是有内在联系的，认识事物之间的这种相似性，可以增进我们进行科学研究的才干，提高研究工作的自觉性和目的性。英国物理学家法拉第（M. Faraday）受到他的老师戴维（H. Davy）把化学能转化为电能，又把电能转化为化学能的可逆过程的启发，立志要把已经发现的由电生磁现象转化为由磁生电，经过 9 年努力，终于完成了由磁生电的实验，建立了电磁感应学说的完整理论。(Nersessian, 1984)[8]正是这种对于事物之间相似性的信念，使我们更加坚信，非单元性确实是语言符号的又一个重要特性。

自然语言处理的理论和实践，加深了我们对语言符号的非单元性的认识。为了改进乔姆斯基的短语结构语法，在自然语言处理的许多理论中，都自觉地采用“复杂特征”的概念，使用“特征/值”系统来描述句子的结构。

自然语言处理还提出了非单元性的这种“复杂特征”进行运算的数学方法——“合一”(unification)运算，从而使我们有关语言符号非单元性的认识可以在计算机上进行实际的操作和演算。这种合一运算，并不完全服从于传统的集合论的运算。集合运算一般并不考虑运算对象的相容性，而合一运算则必须考虑运算对象的相容性。合一运算具有两种作用：

（1）合并原有的特征信息，构造新的特征结构，这与集合论中的“求并”运算类似。

（2） 检查特征的相容性和规则执行的前提条件，如果参与合一的特征相冲突，就立即宣布合一失败。

可见，合一运算提供了一种在合并各方面来的特征信息的同时，检验限制条件的机制。这正是非单元性的语言符号在计算机上运算时所需要的。所以，自然语言处理不仅在理论上证明了语言符号确实具有非单元性，而且还在实践上使这种非单元性获得了在计算机上进行运算的可能性。

3. 语言符号的离散性

我们平时说话时的语流似乎是连续不断的，但在实际上，这些连续不断的语流却是由许多离散的单元所组成的。在水平方向上，语流可以被分解为若干段落，一个段落又可以被分解为若干句子，一个句子又可以被分解为若干短语，一个短语又可被分解为若干单词，一个单词又可被分解为若干语素，一个语素又可被分解为若干音节，一个音节又是由若干个元音

[7] R. Jakobson, On the identification of phonemic entities, TCLP, Vol.V, 1949, p213.
[8] N. J. Nersessian, Faraday to Einstein, Dordrecht, 1984, p40.

和辅音音位组合而成的。在竖直方向上，语流中的各个成分又可引起联想，引出与之属于同一聚类的若干个离散单元来。所以，在连续语流的水平方向和竖直方向上，语言符号实际上都是与若干个不同的离散单元联系着的。

语言符号的这种离散性，在语流的停延时表现得特别明显，人们往往可以利用语流停延的这种离散性质，来区别语流的不同含义。

汉语的书面语在词与词之间是连写的，不像印欧语的书面语那样留有空白，因此，在汉语书面语中，词与词之间的离散特点体现不出来。这种情况，给汉语的自动句法语义分析造成了极大的困难。在中文信息处理中，汉语自动句法语义分析的第一步便是自动切词，根据词与词之间的离散特征，把相互连在一起的词切开。可以说，语言符号的离散性，是汉语自动切词在语言学上的理论根据。

美国语言学家朱斯（M. Joos）早就指出了语言符号的这种离散性。他说：“数学研究工具一般具有两种类型：连续分析（例如，无限小量的计算）或离散分析（例如，有限群理论），而可以称为语言学的那个部门则属于后者，这时，它不容许与连续性有半点儿妥协，因此，凡是与连续性有关的一切，都得排除于语言学之外。”(Joos, 1957)[9] “语言学的范畴是绝对的，是不容许任何妥协的。”(Joos, 1957)[10]他还说，“ 现在，语言学家把任何语言，也就是任何一个言语行为，看成是由叫做音位的不大数量的基本单位组成的，这些音位在重复出现时被认为是等同的。从物理学的角度来看，hotel 这个词对于不同的人或同一人发音，不可能完全相同地发两次，但从语言学的角度看，这里却有一个平均数 (t)，它始终是同样的，可以不管它们的细微差别，而把它们看作一个不可分解的语言学原子或范畴，这种原子或范畴，或者是完全等同的，或者是完全不同的。”(Joos, 1957)[11]这里，朱斯十分明确地把语言看成是“不可分解的语言学原子或范畴”离散地结合起来的，据此，他提出用离散数学来研究语言。他说：“物理学家利用连续数学来解释言语，如傅利叶分解、自相关函数等，而语言学家则与此相反，他们利用离散数学来研究语言。”(Joos, 1957)[12]

朱斯关于语言符号离散性的论述似乎有点儿矫枉过正。语言符号当然具有离散性的一面，但是，语言符号也有连续性的一面，特别是在语言的使用中，在语言的交际过程中，我们也可以利用一些连续数学的方法来研究它，而且实际上在这方面我们已经取得了不小的成绩。朱斯要把“凡是与连续性有关的一切”，“都得排除于语言学之外”，确实是太过分了。事实上，“离散性”和“连续性”都是语言符号本身所具有的性质，不过，在语言使用的交际过程中，我们强调语言符号的连续性，用连续数学的方法来研究它；在语言结构的分析中，我们强调语言符号的离散性，用离散数学的方法来研究它，而语言本身则是离散性和连续性的统一体。

根据语言符号的离散性，自然语言处理中采用集合论的方法建立了自然语言的集合论模型，并把这样的模型应用于机器翻译中，获得了很好的效果。这意味着，语言符号的离散性这一特性，在自然语言计算机处理的实践中已经得到了证实。

4. 语言符号的递归性

语言的句子是无穷无尽的，而语法规则却是有限的，人们之所以能够借助于有限的语法规则，造出无穷无尽的句子来，其原因就在于语言符号具有递归性。

语言符号的这种递归性，在不同的语言里表现不尽相同。汉语的句法构造的递归性突出

---

[9] 朱斯的这些论述，转引自 F. Harary, H. Paper, Toward a general calculus of phonemic distribution, <*Language*>, Vol.33, No.2, p143-169, 1957.

[10] 同上

[11] 同上

[12] 同上

地表现为句法成分所特有的套叠现象。在汉语里，由实词和实词性词语组合而成的任何一种类型的句法结构，其组成成分本身，又可以由该类型的句法成分充任，而无须任何的形态标志。这种套叠现象在主谓结构、偏正结构、述宾结构、述补结构、联合结构、复谓结构中都是存在的。这是由语言符号的递归性导致的汉语语法的一个重要特点。

例如，在句子“他嗓子疼”，中，“嗓子/疼”是主谓结构，这个主谓结构套叠在“他嗓子疼”中做谓语，与“他”又构成一个更大的主谓结构“他/嗓子疼”，这是主谓结构的套叠现象。又如，在短语“北大数学老师”中，“数学/老师”是偏正结构，这个偏正结构套叠在“北大数学教师”中，与它前面的名词“北大”又构成一个更大的偏正结构“北大/数学老师”，这是偏正结构的套叠现象。这些套叠现象都反映出汉语语法的递归性特点。

在自然语言处理的研究中，语言符号的递归性起着很大的作用。机器翻译的实质，就是把源语言中无限数目的句子，通过有限的规则，自动地转换为目标语言中无限数目的句子。如果机器翻译的规则系统不充分利用语言符号的递归性，要实现这样的转换是非常困难的，甚至是不可能的。

乔姆斯基在《乔姆斯基理论介绍》(中文版)(乔姆斯基,1982)[13]一书的序言中指出，早在 19 世纪初，德国杰出的语言学家和人文学者洪堡德（ W. V. Humboldt ）就观察到“语言是有限手段的无限运用”，但是，由于当时尚未找到能揭示这种理解所含的本质内容的技术工具和方法，洪堡德的论断还是不成熟的。那么，究竟应该如何来理解“语言是有限手段的无限”运用呢？ 乔姆斯基指出：“一个人的语言知识是以某种方式体现在人脑这个有限的机体之中的，因此语言知识就是一个由某种规则和原则构成的有限系统。但是一个会说话的人却能讲出并理解他从来未听到过的句子及和我们所听到的不十分相似的句子。而且，这种能力是无限的。如果不受时间和注意力的限制，那么由一个人所获得的知识系统规定了特定形式、结构和意义的句子的数目也将会是无限的。不难看到这种能力在正常的人类生活中得到自由的运用。我们在日常生活中所使用和理解的句子范围是极大的，无论就其实际情况而言还是为了理论上描写的需要，我们有理由认为人们使用和理解的句子的范围都是无限的。”

那么，怎样来刻画语言这个无限集的成分组成情况呢？

我们可以把语言中所有的元素列成一个表，进行简单枚举。例如，

L={Φ, a, b, aa, ab, ...}

这样的刻画办法，把后面一大部分东西省略掉了，后面未列出的部分，只好由我们根据给出的少量的元素去想象，这样的刻画办法显然是不好的。它不能体现“有限手段的无限运用”这一原则。

我们应该采用递归的方法来刻画语言，为此提出如下的公理系统的定义。

一个公理系统是一个有序三元组 (A, S, P)，其中，A 是符号的有限集，叫做字母表；S 是 A 上的符号串的集合，叫做公理；P 是在由 A 中的符号组成的符号串上的 n 位关系的集合， n≥2 （即 P 中的 n 元组至少必须是有序对），P 的元叫做生成式或推理规则。根据这样的公理系统，我们便可以从公理 S 出发，多次使用推理规则 P, 在符号集 A 上递归地生成语言中的句子，实现“有限手段的无限运用”。因而这个关于公理系统的定义是体现了递归的原则的。

如果我们把公理系统中的 A 想象成前面所述的短语结构语法中的非终极符号 V_N 和终极符号 V_T 的集合，把 S 想象成短语结构语法中的初始符号 S, 把 P 想象成短语结构语法中的重写规则 P，那么，我们马上就可以发现，短语结构语法与公理系统是十分相似的。所以我们可以说，短语结构语法是采用体现了递归原理的公理化方法来描述自然语言的语法。

现在，自然语言处理的理论业已严格证明，乔姆斯基的形式语法实际上等价于数学上的

[13] 乔姆斯基，《乔姆斯基理论介绍》（中文版），黑龙江大学出版社，1982 年。

一种公理系统 —— 半图厄系统（semi-Thue system），这种形式语法不过是数学中的公理系统理论在自然语言分析中的应用而已，语言的生成过程完全可以通过公理系统这一形式化的手段得到严格的描述(冯志伟，1985)[14]。正因为如此，乔姆斯基的形式语言理论，才会既在自然语言的信息处理中，又在计算机程序语言的设计中，得到如此广泛的应用(Chomsky, 1963)[15]。

所以，我们认为，语言符号的递归性，是反映了语言符号本质的又一个特点。自然语言处理深化了我们对语言符号的递归性的认识，普通语言学的理论对此应该给以足够的重视。

5. 语言符号的随机性

索绪尔在《普通语言学教程》中，把语言现象分为言语活动（langage）、言语（parole）和语言（langue）三样东西[16]，它们之间是彼此联系而又相互区别的。

他指出，“言语活动是多方面的、性质复杂的，同时跨着物理、生理和心理几个领域，它还属于个人的领域和社会的领域。我们没法把它归入任何一个人文事实的范畴，因为不知道怎样去理出它的统一体。”因此，“言语活动的研究就包含两部分：一部分是主要的，它以实质上是社会的、不依赖于个人的语言为研究对象，这种研究纯粹是心理的；另一部分是次要的，它以言语活动的个人部分，即言语，其中包括发音，为研究对象，它是心理•物理的。”

“把语言和言语分开，我们一下子就把（1）什么是社会的，什么是个人的；（2）什么是主要的，什么是从属的和多少是偶然的分开来了。”

他指出，“语言是一种表达观念的符号系统，因此，可以比之于文字、聋哑人的字母、象征仪式、礼节形式、军用信号等等，等等。它只是这些系统中最重要的。”而言语则“是人们说话的总合”，它包括言语行为的过程（也就是过程）和言语行为的结果（也就是口头的或书面的言语作品）。

索绪尔把语言比作乐章，把言语比作演奏，把语言和言语的关系比喻为乐章和演奏的关系。他说，“在这一方面，我们可以把语言比之于交响乐，它的现实性是跟演奏方法无关的；演奏交响乐的乐师可能会犯的错误绝不会损害这种现实性。” 这是一个非常贴切的比喻。

在索绪尔关于语言和言语区分的理论的影响下，乔姆斯基提出，必须把说具体语言的人对这种语言的内在知识和他具体使用语言的行为区别开来，并把前者叫做语言能力（competence），后者叫做语言运用（performance）。我们认为，乔姆斯基的语言能力，大体上相当于索绪尔的语言，乔姆斯基的语言运用，大体上相当于索绪尔的言语。

在言语（或语言运用）中，当我们用语言来进行交际活动的时侯，有的语言成分使用得多一些，有的语言成分使用得少一些，各个语言成分的使用并不是完全确定的，这种不确定性，就是语言符号的随机性。我们在学习语言时常常感到语言规则中总是有许多的例外，这些例外，就是由于语言符号的随机性造成的。所以，语言符号的随机性，也应该是语言的本质属性之一。

正因为语言符号具有随机性，所以我们很难用确定性的规则来描述它。语言使用中大量的例外现象使研究语法的语法学家们伤透脑筋，有的语法学家甚至因此而误入迷津，以偏概全，得出了错误的结论。为了避免以偏概全的错误，我国前辈语言学家曾提出“例不过十不立，反例不过十不破”的原则来制定语法规则，这个原则常常作为判断语言学家治学态度是否严谨的准绳。其实，对于言语活动这样的随机现象来说，找出十个例子来立某条语法规则并不难，而找出十个反例来破某条语法规则也很容易，以十个例子或十个反例来作为某条语

[14] 冯志伟，数理语言学，上海知识出版社，1985 年。
[15] N. Chomsky, G. Miller, Introduction to the formal analysis of natural language, In R. D. Luce, R. Bush, & E. Galanter, (Eds.), Handbook of Mathematical Psychology, Vol. 2, 323-418, Wiley, New York, 1963.
[16] langage, parole, langue 都是法语。

法规则破或立的标准，看来未必恰当。 最好的办法还是采用统计数学的方法来对交际活动中所出现的各种语言现象进行描述。如果我们从语言学理论的高度，把随机性看成是语言符号本身的一种自然特性，并采用恰当的数学工具来描述这种随机性，使用计算机来进行一般手工操作所难于胜任的大量的统计计算和分析，那么，我们对于语法规则中的各种各样的例外情况，也就不会再感到迷惑不解和束手无策了，因为这些例外的情况正是由于语言符号本身的随机性这一个特点而形成的。(冯志伟，2006)[17]

从自然语言处理的角度看来，在语言成分的出现这一个随机事件中，随机事件 A 与条件组 S 之间虽然没有完全确定的联系，但是，它们之间却有着统计上的联系。尽管当条件组 S 实现一次时，事件 A 可能发生，也可能不发生。但是，如果条件组 S 实现多次，事件 A 的发生就有着某种规律性，这种规律性就是统计规律性。自然语言处理认为，那些无一例外的必然的规律性，只不过是这种统计规律性的补充和表现形式罢了。

近年来，不少的语言学家开始认识到语言符号的这种随机性，自觉地使用统计方法来描述自然语言现象，这是令人欣喜的。在计算语言学中，根据语言符号的随机性，已经在计算机上作了很多统计工作，成果累累。我国学者进行的汉字字频统计、汉字部件统计、汉字笔画统计、书面语词频统计、汉字熵值计算、汉字冗余度计算、汉语语音统计、汉语方言亲疏关系的分析和统计，为汉语的自然语言处理研究提供了可靠的统计结果，推进了我国自然语言处理研究的发展。这些事实说明，一旦我们在理论上自觉地认识到语言符号的随机性，就会产生出巨大的物质力量。语言学的理论对于语言研究的实践确实有着重要的指导意义。

语料库语言学的研究，可以帮助我们从大量的经过标注的语言素材中，发现语言的统计规律，并把这些规律提炼为自然语言处理的规则。这种研究生动地体现了索绪尔所指出的语言和言语的相互关系。大量的语言素材相当于索绪尔定义的言语，语言学规则相当于索绪尔定义的语言，通过对言语的统计研究，就可以发现语言的规律。这是语言符号随机性的又一佐证。

6. 语言符号的冗余性

语言成分在交际活动中的出现是一个随机事件，语言成分之间彼此有着相互的影响和制约，也就是说，前后的语言符号具有相关性，我们根据前面出现的符号，常常可以预测后面的符号出现的可能性。当说话不清楚或文字有错落时，我们往往可以根据前后文来理解话语或文章的含义。就是当某个汉字或拉丁字母不清楚时，我们根据它们的残存部分常常就可以推断文字的全形。在有噪声或干扰时，我们仍然有能力根据已经听清楚的部分来识别那些不清晰的语音。这些事实说明，并不是语言中的一切成分对于传达语言符号整体所包含的信息都是绝对不可缺少的，就是缺少了某些部分，语言本身有能力把这些缺少的部分补充和恢复出来。这意味着，语言符号具有冗余性。这种冗余性是必要的和有益的，它保证了不理想的环境下（如书面文章中有遗漏，谈话是有嘈杂声，书写的字母不清楚，发音不清晰），仍能发挥其交际功能。因此，我们不能认为冗余度就真的是语言中“冗余”的或不必要的东西。恰恰相反，这种冗余度是语言传递信息时必不可少的。没有冗余度的语言在实际上是无法理解的，因为日常语言总有很大的灵活性，要想理解句子的意思就必须考虑到字母在单词中的位置和单词在句子中的上下文关系。我国著名语言学家李荣教授建议把“冗余度”改为“羡余度”，这是很有道理的。事实上，只要语言有结构性就会有冗余度，语言符号的冗余度就是语言的结构性在语言使用过程中的体现。这样看来，语言符号的冗余性也应该是语言符号的一个重要特性，它与语言符号的随机性一样，无时无刻不在语言的使用中表现出来。

自然语言处理已经根据各种言语统计的结果，计算出世界上许多种语言的冗余度。现在

[17] 冯志伟，当前自然语言处理发展的四个特点，《暨南大学华文学院学报》，2006 年，第 1 期（总 21 期）。

世界上各种语言的冗余度中，计算得比较精确的是英语。柏登（N. Burton）和里克里德（J. Licklider）两人通过大量的计算求出，英语书面语的冗余度在 67% 到 80% 之间。汉字是一个大字符集，要直接计算汉语书面语的冗余度，其工作量是非常大的，所以至今为止，我们还不能直接来计算汉语书面语的冗余度，只有通过间接的方法来估算。我国计算语言学研究者现已估算出汉语书面语的冗余度在 56% 与 74% 之间，其平均值约为 65%. 可以看出，汉语书面语的冗余度，其上下限都略低于英语书面语的冗余度。（冯志伟，1985）[18]

汉语的冗余度比英语低一些，说明汉语比英语“简练”一些，而“难懂”一些。所谓“简练”一些，就是对同一篇文章，中文将比英文短一些；而所谓“难懂”一些，就是指从平均的角度看，文章中对于同样长的字母序列，在语义方面给人们的预示能力差一些，或者说，它的语义更难捉摸一些，语义的不肯定性程度更大一些。自然语言处理的这些研究成果，与我们对于汉语和英语的实际体会是一致的。这说明，自然语言处理对于语言符号的冗余性的认识是正确的。

7. 语言符号的模糊性

索绪尔完全没有认识到语言符号具有模糊性。他在《普通语言学教程》中写道：“从心理方面看，思想离开了词的表达，只是一团没有定形的、模糊不清的浑然之物。哲学家和语言学家常一致承认，没有符号的帮助，我们就没法清楚地、坚实地区分两个观念。思想本身好像一团星云，其中没有必然划定的界限。预先确定的观念是没有的。在语言出现之前，一切都是模糊不清的。”他又说，“语言对思想所起的独特作用不是为表达观念而创造一种物质的声音手段，而是作为思想和声音的媒介，使它们的结合必然导致各单位之间彼此划清界限。” 显而易见，索绪尔认为，正是由于语言的作用，才使模糊的思想和声音的各个单位之间清晰起来。在索绪尔看来，语言本身是谈不上模糊性的。

关于语言的模糊性问题，在自然语言的计算机处理出现之前，就有不少学者进行过探索和研究。英国著名哲学家罗素（B. Russell）于 1923 年写过一篇《论模糊性》的论文（罗素，1990）[19]。他指出：“整个语言都或多或少是模糊的”，并且举例论证了这个问题：“ 由于颜色构成一个连续统，因此颜色有深有浅，对于这些深浅不同的颜色，我们就拿不准是否把它称为红色。这不是因为我们不知道“红色”这个词的意义，而是因为这个词的使用范围在本质上是不确定的。这自然也是对人变成秃子这个古老之谜的回答。假定一开始他不是秃子，他的头发一根根地脱落，最后才变成秃子。于是有人争辩说，一定有一根头发，由于这根头发的脱落，便使他变成秃子。这种说法自然是荒唐的。秃头是一个模糊概念；有一些人肯定是秃子，有一些人肯定不是秃子，而处于这两者之间的一些人，说他们必定要么是秃子，要么不是，这是不对的。排中律用于精确符号时是正确的；但是当符号模糊的时侯，排中律就不适用了。事实上，所有的符号都是模糊的。所有描述感觉特性的词，都具有‘红色’这个词所具有的同样的模糊性。”罗素这篇论文对传统逻辑学中的排中律提出挑战，从哲学和逻辑学上为模糊理论奠定了基础。

1933 年，美国语言学家布龙菲尔德（L. Bloomfield）在《语言论》一书中（布龙菲尔德，1980）[20]，也指出了自然语言中存在着模糊现象。

他说，“我们可以根据化学或矿物学来给矿物的名称下定义，正如我们说’盐’这个词的一般的意义是‘氯化钠’(NaCl)，我们也可以用植物学或者动物学的术语来给植物或者动物的名称下定义，可是我们没有一种准确的方法来给象‘爱’或者‘恨’这样一些词下定义，

[18] 冯志伟，数理语言学，上海知识出版社，1985 年。

[19] 罗素，论模糊性，中译文见《模糊系统与数学》，1990 年，第 9 卷，第 10 期。

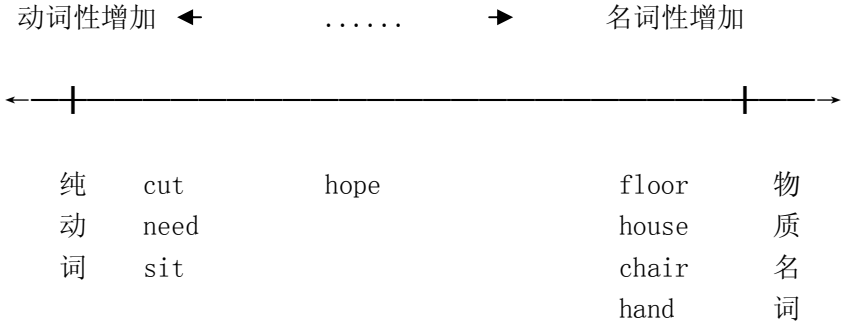
[20] 布龙菲尔德，《语言论》，中译本，商务印书馆，1980 年。

这样一些词涉及到好些还没有准确地加以分类的环境——而这些难以确定意义的词在词汇里占了绝大多数。”他进一步指出：“此外，即使我们有一些科学的（也就是普遍被承认的而又不准确的）分类，我们也还往往发现语言里的意义跟这种分类并不一致。”

这些研究都指出了自然语言里存在的模糊现象。直到 1965 年，著名数学家查德（L. A. Zadeh）发表了《模糊集》的著名论文后，模糊性的概念才第一次找到了完善的表示方法(查德, 1981)[21]。查德的研究是首先从观察语言符号的模糊性开始的。例如，“老年”这个概念就具有模糊性。七十岁算不算“老年”？如果算，那么，60 岁算不算“老年”？50 岁算不算“老年”？这是很难精确地回答的。查德把“老年”看成是建立在“年龄”这个论域上的集合，而把 70 岁、60 岁、50 岁都看成这个集合中的元素，这样，就可以研究这些元素相对于“老年”这个集合的隶属关系。这种隶属关系，很难用经典集合论中的“属于”或“不属于”某个集合的办法来描述，而可以用在多大程度上属于某个集合的办法来描述。也就是说，一个模糊集合 S 的特征，存在着一个隶属函数 μ ，对于论域中的每一个元素 x ，都有一个确定的值 μ (x)，这个值刻画着元素 x 隶属于模糊集合 S 的程度。查德把普通集拓广为模糊集，为模糊数学奠定了基础，这一开创性的工作不仅拓广了普通数学的研究领域，而且开辟了在软、硬科学（包括语言学）中提高数学适用性的广阔途径。

应该强调指出的是，模糊数学的产生和发展，首先是从观察和研究自然语言中的各种模糊现象开始的。查德本人在《模糊集》(查德, 1981)[22]一文中曾明确地说明：“模糊集合论的这个分支的起源是从语言学方法的引入开始的，它转而又推动了模糊逻辑的发展.....在即将到来的时代，我相信近似推理和模糊逻辑将发展成为一个重要领域，从而变成研究哲学、语言学、心理学、社会学、管理科学、医学诊断、判别分析以及其它领域的新方法的基础。”模糊语言的研究已引起了语言学家们的浓厚兴趣。1972 年在美国纽约举行的词典学国际讨论会上，美国语言学家雷柯夫（G. Lakoff）作了一个在词汇研究方面应用模糊数学的报告。雷柯夫高兴地说：“我们现在有了一个‘可爱的术语’——模糊集合”。 他在讨论会结束时又指出，模糊性将成为语言学研究的一个主要领域。

语言符号的模糊性不仅存在于单词的含义方面，语法方面也存在着模糊性。例如，许多语言中动词和名词的划界并不十分清楚，存在着“亦此亦彼”的现象，也就是说，动词和名词的划界是模糊的。美国语言学家洛斯（Ross）提出了“动/名连续统模型”来描述英语中动词和名词的划界问题。在连续统的两端分别是纯动词和物质名词，它们的界线是截然分明的。但是在这个连续统两端的中间，则存在着一系列界线模糊的过渡类，可图示如下：



可以看出，处于连续统中间的 hope （希望）这个词，兼具动词和名词的特点，表现了在词类归属上的模糊性。英语中的很多词，都可以根据它们在性质上的差异来确定它们在连续统上的位置。最近有学者采用这个“动/名连续统模型”来解决汉语的动词和名词的分界

[21] 查德，模糊集，中译文见《自然科学哲学问题》，1981 年，第 1 期。
[22] 查德，模糊集，中译文见《自然科学哲学问题》，1981 年，第 1 期。

问题，取得了较满意的结果。

在自然语言处理中，自然语言的表达和理解技术是一个十分困难的问题。学者们已经认识到，这个问题比他们原来预料的更加艰难，美国国会技术评价办公室最近指出，要使计算机具备一个五岁小孩的自然语言理解能力说不定是二十年以后的事。自然语言的表达和理解的主要困难在于自然语言本身的模糊性。这种困难的内在原因是我们对于人类如何贮存和处理模糊信息的机制还不十分清楚，外在原因是我们还没有一种适合于处理自然语言的模糊信息的工具。由模糊数学创始人查德亲自开拓的可能性理论、模糊语言方法以及由此而产生的模糊语言逻辑、自然语言语义表达和近似推理，已经构成一个知识分支，正在把克服上述自然语言理解和表达技术中的困难当作自己的研究目标，目前已取得了令人鼓舞的成果。可见，自然语言处理的研究将会推动我们更加深入地探讨语言符号的模糊性问题。

语言符号的模糊性与语言符号的随机性是两个不同的概念。

前面说过的语言符号的随机性是就事件的发生与否而言，但事件本身的含义是确定的，由于条件不充分，事件的发生与否有多种可能性，在［0,1］上取值的概率分布函数就是描述这种随机性的，它经常表现为字符或单词出现概率的大小。

语言符号的模糊性则是指元素对集合的隶属关系而言，事件本身的含义是不确定的，但事件发生与否是可以确定的，因而元素（事件）对集合的隶属关系是不确定的，在［0,1］上取值的隶属函数就是对于这种不确定性（即模糊性）的数学描述，它经常表现为单词含义对某一集合的隶属函数的数值的大小，也就是单词含义对某一集合的隶属程度的高低。

语言符号的随机性放弃了“一因一果”的决定论，反映了“一因多果”的规律性，因此，它是由于因果律破缺而造成的一种不确定性，在用统计方法来描述自然语言时，是满足排中律的。

语言符号的模糊性摆脱了“非此即彼”的确定性，反映了“亦此亦彼”的规律性，因此，它是由于排中律破缺而造成的一种不确定性。

研究语言符号的随机性，可以把语言学的领域从必然现象扩大到偶然现象，研究语言的模糊性，可以把语言学的研究领域从清晰现象扩大到模糊现象。因此，语言符号随机性和模糊性的发现，都加深了我们对于语言符号本质的认识，拓广了语言学的研究领域。

三、结语

由此可见，层次性、非单元性、离散性、递归性、随机性、冗余性、模糊性等七个特性也是语言符号十分重要的特性。索绪尔提出的语言符号的线条性可以用更为深刻的层次性来代替，而他提出的语言符号的任意性确实是“头等重要的”、“支配着整个语言学”的原则。因此，我们认为，语言符号的特性除了上述的七特性之外，还应该加上任意性，这样，语言符号就具有任意性、层次性、非单元性、离散性、递归性、随机性、冗余性、模糊性等共八个特性。自然语言处理的发展，使我们对于语言符号的这些特性的认识和理解更为丰富、更为深刻了。在这种情况下，我们不得不修正索绪尔的旧理论，而代之以反映当前人类对自然语言符号认识水平的新理论。这是自然语言处理在普通语言学的基本理论方面对理论语言学提出的挑战。

语言符号的任意性，也就是语言符号的社会约定性，它反映了语言符号的社会——人文的本质，这使我们有可能用社会科学的方法来研究语言。语言符号的层次性、非单元性、离散性、递归性、随机性、冗余性反映了语言符号的物质——自然的本质，这使我们有可能用自然科学的方法来研究语言。而语言符号的模糊性，则表现了人类心智活动和思维活动的特点，反映了语言符号的智能——心理的本质，这使我们有可能用思维科学的方法来研究语言。这样，原来作为纯粹人文科学的语言学，在计算机时代便大大地拓广了它的研究领域，

使它同时跨着人文科学、自然科学和思维科学三个领域。

法国著名数学家阿达玛（J. Hadamard）曾经说过：“语言学是数学和人文科学之间的桥梁”（冯志伟，1991）[23]，今天，我们可以进一步说：“语言学是自然科学、思维科学和人文科学之间的桥梁”。一向被人们看成是“冷门儿”的语言学，现在已经改变了它在整个现代科学体系中的地位，正在成长为一门带头的科学，成为现代科学技术研究的一个热点，以至于连许多自然科学家和计算机专家也认为电子计算机软件工作也可以看成是一种语言文字工作，这是每一个语言文字工作者应该引以为荣的(钱学森, 1994)[24]。

参考文献

[1] 冯志伟（2007），《机器翻译今昔谈》，语文出版社，2007 年。

[2] 冯志伟（1992），计算语言学对理论语言学的挑战，《语言文字应用》，1992 年，第 1 期。

[3] 索绪尔（1980），《普通语言学教程》，中译本，商务印书馆。1980 年。

[4] J. R. Firth（1948）, Sounds and Prosodies, 1948.

[5] A. M. Turing（1950）, Can a machine think? *Mind*, 50, 1950.

[6] R. Jakobson（1949）, On the identification of phonemic entities, *TCLP*, Vol.V, 1949, p213.

[7] N. J. Nersessian（1984）, Faraday to Einstein, Dordrecht, 1984.

[8] F. Harady, H. Paper（1957）, Towards a general calculus of phonemic distribution, *Language*, Vol. 33, No.2.

[9] 乔姆斯基（1982），《乔姆斯基理论介绍》（中文版），黑龙江大学出版社，1982 年。

[10] 冯志伟（1985），《数理语言学》，上海知识出版社，1985 年。

[11] N. Chomsky, G. Miller（1963）, Introduction to the formal analysis of natural language, In R.

[12] D. Luce, R. Bush, & E. Galanter, (Eds.), *Handbook of Mathematical Psychology*, Vol. 2, 323-418, Wiley, New York, 1963.

[13]冯志伟（2006），当前自然语言处理发展的四个特点，《暨南大学华文学院学报》，2006 年，第 1 期（总 21 期）。

[14] 查德（1981），模糊集，中译文见《自然科学哲学问题》，1981 年，第 1 期。

[15] 冯志伟（1991），《语言与数学》，湖南教育出版社，1991 年，长沙。

[16] 钱学森（1994），电子计算机与新时期的语言文字工作，《中文信息》，1994 年，第 2 期。

（冯志伟　教育部语言文字应用研究所）

[23] 转引自：冯志伟，《语言与数学》，第 1 页，湖南教育出版社，1991 年，长沙。

[24] 钱学森，电子计算机与新时期的语言文字工作，《中文信息》，1994 年，第 2 期。