齐夫定律和语言的"熵"

—— 普通语言学论集

曹聪孙 著

天津人民出版社

齐夫定律和语言的“熵”

——普通语言学论集

曹聪孙　著

天津人民出版社
1994年

（津）新登字001号

齐夫定律和语言的“熵”

——普通语言学论集

曹聪孙 著

天津人民出版社出版发行
（天津市张自忠路189号）
天津师范大学印刷厂印刷
850×1168毫米 32开本 9印张20.3千字
1994年2月第1版 1994年2月第1次印刷
ISBN 7—201—01699—7/H.26

定价：6.80元

弁言

这是一本近40年有关普通语言学和汉语研究的个人部分论文结集。文章全部是发表过的。发表的报刊有《中国语文》、《语文学习》、《辞书研究》、《中国翻译》、《知识工程》、《天津师大学报》、《文汇报》等，收入一些专门论文集的如《语法图解法的比较研究》等，就没有阑入。最早的一篇1955年的《是不是倒装》是全国语法界讨论主语宾语问题时发动讨论的三篇论文之一；最近的是发表于1993年的《语言的T型结构刍议》。附录是作者几本著作的例言或结语。

这些文字虽然接触到了语言研究中的某些理论问题和实际应用领域，但恐怕都是浅尝辄止，未遑深入。山积而高，泽积而长。这里既无山积之高，也无泽积之长。既没有做到研幽阐微，又未能考本论末。因此，它只是一个探索语言问题过程的纪录，而不是什么有实质性结论的研究成果。好在今天语言学的研究工作正方兴未艾，如日中天。这也算是在雄伟的大合唱中加入一个弱音符罢了。是为序。

曹聪孙

1993年12月于天津师范大学

齐夫定律和语言的“熵”

——普通语言学论集

目　　录

词汇学

语法学

附录：

齐夫定律和语言的"熵"

一 语言的繁复与简缩

虽然谁都不能直接观察到语言的系统演变过程，但谁也都相信语言是在发展变化着的。历史记载了语言的演变。语言学史、语音学史、词汇学史、语法学史乃至文字学史，都记述和论证了这些变化。

历史比较语言学告诉了我们语音变化与对应的规律；词源学向我们揭示了词汇的历史来源；地理语言学探索了方言间的差异及其发展。历时语言学既有前瞻的探索，又有回溯的研究。具体的语言系统要素各自遵循着某些规律在历史的长河中不断地向未来流去。

我们在当前还不能概括全部语言发展的内部规律，这是由于科学发展的历史阶段性所决定的。不过，我们在这里可以提出一个易于观察而又未被详细论述过的规律性现象。这就是统摄语言变化的缩简趋势与拉长趋势的统一。它也可以表述为"齐夫定律"和语言的"熵"。

齐夫定律（zipf's law）是语言学中的一个著名的定律。齐夫是美国的语言学家。他在对莎士比亚等作家的作品进行数理统计后建立了这个定律。它可以概括地解释为：语言中使用

频率最高的词就是那些最短的词。由于表达的需要，表达者尽量把语言说（写）得简短些，这就是语言的节约原则或经济原则。

法国著名语言学家若热·穆南在他的《语言学词典》①中解释过："为齐夫所完成的统计语言学研究表明，一个词的级词，符合它与其他词的关系（处于同一素材之中）。词是以使用频率来分类的。排列在第一的是使用频率最高的词。其余类推。这就叫齐夫定律。"英国的哈特曼与斯托克的《语言与语言学词典》②中的释义是，"（齐夫定律）是指关于谈话者或写作者使用的词的分布和频率(Frequency)的总描述。F×R=C。这个方程式中，F=frequency（频率）；R=rank(级)，即频率表上的位置；C=constant（常数）。这个方程式表示，词使用的总次数和词在频率表上的位置之间有一个固定比率。……语言中使用频率最高的词也就是最短的词。"

语言学规律和物理学的定律不同，它们都不是预测性质的。它们只能对已有的语言现象作出合理的描述与解释。齐夫定律也不例外。它观察了千百种语言现象并从历时语言学的研究中得到启发。齐夫的结论是对语言事实的一种科学的归纳和理论概括。

在具体的言语活动中，人们感知到的语言结构和他所能（可能）理解到的信息之间，有时是一致的；但有时并不完全一致。表达者可以使用几种不同的结构组合方式把他要表达的信息传达给接受者。这种结构可简可繁。必要时，还可以把它简略到最低限度，达到最省力、最经济的接近"安全阀"的位置。

昂得勒·马丁内指出，言语活动中有一种内部发展的力量。这种力量就是表达需要与人的自然惰性的矛盾。惰性要求

减少消耗而表达则需要日趋繁复。它们之间的冲突的结果就形成了语言始终处于不断发展变化之中。

什么叫语言的日趋繁复呢？这就是一种新兴的理论，可以概括为“语言的‘熵’”。

信息论借用了统计物理学中的“熵”（entropy）的概念来建立自己的理论基础。“熵”，是热力学函数的概念，指的是一个系统中自发的变化总会导致混乱。宇宙间的事物发展有时是从无序到有序；有时又是从有序到无序（混乱）。“熵”在物理系统的自发变化中，总是向着增大的方向发展，即从有序朝无序的方向变化。

语言的“熵”和这一物理量的物质变化是可以比拟的。在语言长期发展变化的过程中，语音的数量总是有所增加，书写的长度有所伸延。明显的趋势是：表达相同概念的语音单位在历史的进程中被拉长了。语言趋向繁复。

语言活动酷似分子的热运动。在千百万人的不断使用当中，不断增加它的无序和混乱程度。多数语言的平均长度有拉长的趋势。即，沿着“熵”增大的方向发展。$ds(L) \geqslant 0$ ③。

齐夫定律把语言单位缩短，而语言的“熵”又把语言单位拉长。一缩一拉，这就造成语言不断变动、经常变化的总趋势。也恰恰是这两种力量的不断平衡才形成语言的相对稳定性。这正是语言发展史上的辩证的统一。

二 “可懂度是语言损耗的极限”

L·R·帕默尔在他的《语言学概论》里讲过，“可懂度是语言损耗的极限”。这话的意思是说，表达所需要的最后限度的语言总量（包括文字长度）是以人们的可以理解为最后依据的。

语音的简化是我们可以从多种语言中历史地观察到的一种

现象。比如，汉语的全浊声母清化现象就指示出了这种趋向。中古汉语存在过36个声母。其中一套浊声母并〔b〕、定〔d〕、群〔g〕、澄〔ɖ〕、从〔dz〕、床$_1$〔dʒ〕，床$_2$〔dʑ〕、邪〔z〕、禅$_1$〔ʒ〕、禅$_2$〔ʑ〕匣〔ɣ〕等，到了近古时期已经变化为相同发音部位和发音方法的清声母。由于受声调的影响，浊塞音、塞擦音的平声变为送气音；仄声变为不送气音、入声的消失也同样说明了这一点。这是遵循语音变化中的一条：处于词尾的音素往往受到压力而变弱、损耗、脱落的规律而形成的。粤方言仍有入声，还有9个声调。福州话保留了一个〔-k〕，吴方言残留了一个喉塞音〔ʔ〕。这历史耗损过程当然是渐进的。学者们的预测是：先是三个入声合并成一个，而后又变成喉塞音，最后全部消失。也有人认为，历史上北方的少数民族众多，他们学汉语时发音缓慢，易拖长音，结果就把字尾脱落了。上述现象表明，声母精简了，声调类型减少了，语音的物质性表达类型变得简单了。

浊声母
入声字 ＞部分保留＞全部消失

不妨列一个表看一看汉语语音的发展历史。古汉语语音构拟采取时贤成说。（见下页）

总的来看，现代汉语比起古代汉语来，明显的变化是：声母、韵母、声调的数量全都减少了。表面看来声调似乎没有少，但它的有区分的调值类型是具体地减少了，象入声的三个类型收〔-p〕〔-t〕〔-k〕的，全部消失了。这些都符合从齐夫定律引申出来的规律。

“弱化”也是一种省力和经济原则的体现。它以较轻微的发音动作来代替较费力的发音动作。〔ə〕是一个省力的央元音，用它来代替较费力的其他元音就是一种弱化作用。汉语：

时　　期	声母	韵　　部	声　　调	注
先　　秦 （前207）	33	29—30 （战国时代30）	4（平、上、长入、短入）	有人主张古有五声说
两　　汉 （前206—公元220）	33	29	4	王力认为去声尚未产生
魏晋南北朝 （220—589）	33	42	4（平上去入）	
隋　　唐 （581—907）	33	50	4	上去调值相近，可以通押
晚唐五代 （907—960）	36	40	4	浊上变去
宋 （960—1279）	21	32	4	三类入声并为一类
元 （1271—1368）	25	19	4	平分阴阳，入派三声
明 （1368—1644）	21	15	4	
清 （1644—1911）	20—23	15	4	
现代汉语普通话 （1911—）	23	16	4	方言声调多。厦门、苏州：7个；广州：9个。

“桌子”的“子” 〔tsl〕 ⟶ 〔tsə〕

“西瓜”的“瓜” 〔kua〕 ⟶ 〔kuə〕

英语的冠词a〔ei〕人们习惯把它发成〔ə〕，an〔æn〕也常常发成〔ən〕，the〔ði：〕也发成〔ðə〕。央元音〔ə〕岂不神通广大！

丹尼尔·笛福所写《鲁滨逊飘流记》的原型亚力山大·塞尔格克独自一人在荒岛上生活了4年。他在1901年1月被救离荒岛后，英语遗忘得很厉害。他几乎把所有词的词尾都“吞掉”了。这就是语音学上的“脱落”。古拉丁语宾格的鼻音词尾在今天的任何一种罗曼语当中都已消失不存了，如terra m(土地）中的m。现代法语这个词是terre。这正好象古汉语的韵尾〔-m〕在现代汉语普通话中已经不存在了一样。真是无独有偶。

在印欧语系的语言中，辅音的减少与辅音组合的简化、某些元音音位的消失等，无一不是语音简化与语言省力、经济原则的体现。古拉丁语有19个辅音，发展到民间拉丁语就减少了5个而成为14个。法语的塞音体系比起拉丁语来是明显地简化了。12—16世纪，法语的复合元音变成了单元音。

日常口语现象是描写语言学的描述对象。它的省（略）音（素）现象在各种语言中屡见不鲜。汉语的“三个”被说成三〔ə〕，省去了声母〔k〕。“教育局”竟然被说成“教局”，干脆省掉了一整个音节。一次，笔者明明听到有人把托赛里的小夜曲中的歌词“yon and only you”（你，也只有你）〔yɔ：and ounli yɔ：〕唱成了〔yɔ：ən li yɔ：〕，一下子把〔d〕〔oun〕全都省去了。

汉语的同音省略可以找到极好的实例：

这是谁的摊儿？*是那个卖菜的的。

是那个卖菜的。

这馅饼是什么馅的？"是羊肉的的。

是羊肉的。

应该说两个"的"，第一个"的"是"的字结构"，第二个"的"是"是……的"的一部分。可是人们在口语、书面语当中只用一个"的"，谁也没觉得自己或别人在"省略"了什么。习惯使然。

上边提到的是词尾的脱落，也还有词头的脱落。象法语的les boches原应为Alboches（德国的，有贬义），le troquet原为mastroquet（小酒馆），词首的Al、mas都被省略了。

齐夫定律在语音中这种表现为什么不妨碍表达上的需要呢？有没有什么科学上的而不是经验上的根据来说明这个问题呢？有的。心理学做过一些实验。实验证明，在语境当中，词的"平均识别"时间又早又快。具体地说，在语流里，听话人在听到一个词的多一半的时候，已经可以清楚地辨认出它的含义了。科学家测定出，人们只需要200毫秒就可以听懂一个词，而一般词的平均时长为375——420毫秒④。看来，人的语言分辨和反应能力可以实现最佳效率的信息利用。这样的实验可以印证一个传统的理论：表达的可以印入，如果隐去的也可以印入，那就无需再表达了。

至于词汇方面，从印欧语系的语言来看，不少语言的词长，词形都有缩短的趋势。布龙菲尔德在《语言论》里说，"(现代英语)词法简单，词形大大地缩短了。""语言中的变化倾向于词的构造更规则和更短小。语音变化把词'缩短'了。法语的voiture automobile(机动轿车)可以单独说成voiture或auto mobile，两者均可以当"小汽车"解。词形缩短了近一半。

人们都会注意到，近二三十年来，各种语言中的"缩略

词”大大风行起来了。缩略法进入了构词法。缩短的类型有前截、后删、中略、前后节略、缩写等。汉语译作“笔会”的“笔”（pen）就是一个词头缩写词，即poet（诗人）、Essayist（散文家）、Novelist（小说家）三个词的词首字母。缩写成的pen又恰好是“笔”的意思。“托福考试”的“托福”（TOEFL）也是一个缩写词，即 Test of English as Foreign language（作为一门外语的英语测试），由每个实词的词首字母组成。英语的“夏季时间”称D.S.T，就是daylight saving time（日光节约时间）。令人谈虎色变的爱滋病（AIDS）就是 Acquied Immune Deficience syndrome（后天性免疫力缺损综合症）。全是词首字母的合写。

汉语的新词“三胞”（台湾同胞、港澳同胞、海外侨胞）“四眷”（台湾、港澳、华侨、外籍华人的家属）也是缩略词的一种类型。

英语的缩略词语据《英语缩略语词典》（史群编、商务1979年版）所收，约有4万条。法语的《法汉缩略词典》收词3万条。俄语的《俄汉缩略语词典》收词4万3千条。可以想见，各语种的缩略词都是数以万计的。

简化表达是浓缩信息、节约交流时间的有效方法。在汉语，这方法是古已有之的。“不可”缩略为“叵”，“之于、之乎”缩略为“诸”，见诸古籍。“不用”合成“甭”，“勿曾”合为“艚”（fēn），是方言。

表达上的简化是这一缩略现象的依据。当然，也还有某些其他原因促成这一现象的形成。例如，缩略语和委婉说法之间存在着一定的联系。为了使某些谈法表达得不太粗俗和低级，选用比较文雅的缩略写法也许是可取的。象英语的son of a

bitch（骂人话，婊子养的）就写成SO B，bullshit（狗屁，骂人的粗话）写成BS⑤。

语法方面的简化现象也是多种语言在历史的发展过程中经常出现的。古汉语有“格”的变化。方言客家话至今仍有主格和所有格，成为古汉语的遗迹。俄语“数”的范畴也省略了一部分，双数消失了。古俄语的名词变格较多，有五种。发展到现代俄语，减少到了三种。一些阳性名词的词尾受阴性名词的影响，由-N变成-Ы，如复数主格СТОЛИ（桌子）变为现代俄语的СТОЛЫ，这也是语法上词形变化简化的结果。英语名词的格现在只剩下三个了（主、宾、所有格），德语还有4个。芬兰语格变复杂，现在仍有15（或16）个。语言学家以为，它比较接近拉丁语。从“格”这一语法范畴来说，印欧语系各语言都是类型越来越少，词形变化越来越小。英语的主、宾格也是名存实亡，词形变化已经不见了。

三　“熵”把语言单位拉长

在一个热力体系中，自发的变化随着运动的不规则性会增加其无序程度。统计物理学使用“熵”这个概念是指热的本质就是物质分子的不规则运动。我国物理学家李仙洲在30年代创制了“熵”这个字来翻译entropy这个词。

“熵”在近代信息论中被引入，它跨过了物理学的边缘，进入某些社会科学领域⑥。在语言学方面，它是这样被描述的：在长期的言语活动当中，语言会逐渐增加其混乱程度。为了降低这种混乱程度，要制造（或者已经自发地形成）一个负熵过程。这个过程就是避免苟简，清楚地表达概念和事实。

把语言单位拉长的趋势“熵”可以从语音与词汇合一的角度来加以观察。因为，它们的内容和形式是完全一致的。

古汉语是以单音词为主的，而现代汉语则是以双音词为

主。有人统计，在百万字的资料范围内，双音词的个数约占总词数的80%以上⑦。当然，单音节词的出现频率仍然是较高的。古汉语的单音节词发展到现代汉语的双音节词，如日——太阳、月——月亮等等，例子不胜枚举。

英语语音也不全都是由长变短，由多化少的。有些元音的演变也是“拉长”的佐证。例如house（房屋），中古音读作〔hu：s〕，到了现代英语则读成〔hous〕，当代读为〔haus〕。wine（酒），中古念〔wi：n〕，现代念〔wein〕，当代则读〔wain〕。这说明。英语元音有一部分是从中古的长元音变为现代的复元音，音素比过去增加了。法语的塞音体系的简化从擦音体系的增加上得到了补充。

可以指出一种构词法上拉长语言单位的趋势来说明“熵”的明显现象。这就是“嵌入音”的构词方式。象汉语形容词的AliA B式。

胡涂—胡里胡涂　　马虎—马里马虎

罗嗦—罗里罗嗦　　古怪—古里古怪

这是一种“中缀”（infix）。这一形式印欧语言里很少，东方语言里却较多地存在着。匈牙利语、芬兰语里也有。把词形拉长有着加强原有形容词词义的作用。

语言的“熵”的另一方面——歧义的产生也是必须予以注意的。歧义往往产生于句法的简化。由于表达者过于简略地组织句子，于是就产生了模糊言语。“我们三个一组”有两个语义：“我们三个人在一个组”和“我们每三个人组成一组”。因为省略了句子的动词，句义就变得不清晰了。我国古代有一个著名的例子“夔一足”。这句话到底是“夔（乐正）有一个就够了”还是“夔，只有一只脚。”要不是长沙楚墓出土的晚周帛画上画有一只脚、有角的似蛇的动物和安阳殷墟有出土的一足

雕龙，这争论恐怕还要继续下去。古汉语诠释、训诂、注解之所以聚讼纷纭，时常是由于句法过于简化的结果。现代汉语也不是没有这类问题。“一中学生”是“一个中学生”呢还是“第一中学的学生”？“一竖，一边一点儿”到底是“卜”字呢还是“小”字？看来，苟简的确是一种毛病。简略超过了界限，语言的损耗极限被打破了，可懂度降为不明晰的言语，符号产生了破损。

要想解决这个问题，只有重新“拉长”这些言语，提高它们的明晰性和可懂度。

我们上边提到过的缩略词，它们的流行过程是：繁复的形式——缩略为简单的形式——发生了言语的模糊性——需依靠上下文和注释——？（改变词形等）。这可以用实例来说明。比如ABC这样的缩略语，已经有了多达100个以上的义项。只要看一看E. T. Crowley的《首字母缩略词与缩略词词典》⑧就可以知道，几十个解释的首字母缩略词和一般缩略词是不少的。ABC是“美国广播公司”的略语，可同时，它又是“澳大利亚广播公司”的略语。汉语简称“新三论”（耗散结构论、协同论、突变论）与“老三论”（系统论、控制论、信息论）的称谓极不确切，引起科学分类上的混乱（见1987年1月4日《人民日报》）也是一个明证。

从语法方面来观察，古今汉语的词序虽然没有大的变化，但现代汉语语句的长度是大大增加了的。这一方面与词形拉长有关；另一方面也是添加了新的因素。现代汉语发展出不少量词来，表示时态的助词也增加了。重音、词缀、重迭形式也都有所发展。从句法去说，现代汉语的补语比古汉语扩展了很多，使成式、处置式、被动式都有了新的类型。总的趋势是向复杂化与精密化方面发展。这一则是表达的需要；二则是人的