汉语句间成分共享类型及分布研究*)

张 全[1] 吴 晨[1,2] 韦向峰[1]

(中国科学院声学研究所 北京 100080)[1] (中国科学院研究生院 北京 100039)[2]

摘 要 自然语言中语句之间经常出现句子成分共享的情况。本文以计算语言学理论为指导,首先明确了便于计算机自动处理的句子和句群的定义。以此为基础,获得了真实语料中句群单位内相关数据的统计结果。进而依据语句的定义分析了语句之间语义块共享的类型,给出了语句间语义块共享的具体分类,统计了真实语料中各共享类型的分布数据。同时本文还对统计数据进行了分析,统计结果符合常人对语料的直觉定性判断。本文的结果有助于语句之间成分共享和句群的计算机自动分析。

关键词 中文信息处理,自然语言理解,语句分析,句群分析,语义块共享

The Types and their Statistical Distribution on Sentence Elements Share between Chinese Sentences

ZHANG Quan[1] WU Chen[1,2] WEI Xiang-Feng[1]

(Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080)[1] (Graduate School of the Chinese Academy of Sciences, Beijing 100039)[2]

Abstract The sentence elements are usually shared among Chinese sentences. In order to analyze this problem, the definitions of sentence and sentence group are introduced based on a computing linguistic theory. The definitions are suitable to computer processing. The statistical data from corpus in one sentence group are presented according the definitions. Similarly, the problem of semantic chunks share between sentences is analyzed, and the types of share are put forward. Further more, the statistical data of the types are obtained from the corpus. And the statistical data are construed, the statistical data accord with people's intuition about the corpus.

Keywords Chinese information processing, Natural language understanding, Sentence analysis, Sentence group analysis, Semantic chunk share

1 引言

人们在运用自然语言进行交际时,往往需要用多个语句围绕一个主题,构成句群进行表达。让我们先来看一个例子。

例 1 1856 年 ～‖～,当特斯拉出生时 ～‖,那种天才的征兆‖便怪异地显露了。＋＋〈他母亲| 去做工〉时～‖,一场雷电交加的暴风雨‖向克罗地亚这座名叫斯米利安的小村庄‖袭来。＋＋7 月 10 日午夜 ～‖,母亲‖在隆隆雷声中‖生下了‖特斯拉。接生婆‖由此断言,[＃ 尼古拉‖肯定是‖"风暴之子" ＃]。……

(～‖|[＃ ＃]:是语句的语言空间标注符号,本文中语料标注符号参见文[7]。下同。)

在例 1 中,"接生婆由此断言"之前出现了 3 个句号,构成一个句群,围绕特斯拉出时特殊的天气情况进行表达,并以此作为这个句群的主题。下面的句子由"接生婆由此断言"开始,表述接生婆的语言内容,已不是这个主题,句群也由此结束了。

同时,如果让计算机进行自然语言理解处理,那么句群也是准确把握语义,获取整体表达内容的重要阶梯。

中科院声学所黄曾阳教授创立的概念层次网络(Hierarchical Network of Concepts,简称 HNC)理论立足于建立模拟人脑语言智能的自然语言交互引擎,设计了凸现概念联想脉络的语言概念空间,为自然语言语言重建了数字化的符号系统。语言概念空间包括层次鲜明、内在联系紧密的概念基元、句类、语境单元和语境知识表示等符号体系[1],它们大体对应于语言空间的词语、语句、句群和篇章。句群在自然语言理解中起关键的中枢作用,是进行篇章级处理的必由之路。另一方面,从语言交际研究的角度出发,一般[2]将自然语言分成:语素、词、短语、句子和句群等层次结构单位。其中句群是在语义上有逻辑联系,在语法上有结构关系,在语流中衔接连贯的一群句子的组合,是介于句子和段落之间的语言表达单位。

HNC 中句群就是围绕着一个特定概念展开的话语。HNC 概念基元符号体系为发现确定这个特定概念提供了丰富的线索。同时在 HNC 的语言概念空间中对于句群也设计了相应的表述模式——即语境单元空间[1]:

表示式 1 SGUN=(DOM;SIT;BACE;BACA)

SGUD=(8y:|DOM;SIT;BACE;BACA)

其中:

SIT=SCD(A,B,C) (SCD:领域句类)

DOM——领域

SIT ——情景

BAC——背景

BAC[E//A]——事件//述者背景

其中领域句类 SCD 是指为领域概念配置的特定句类,是

*)本文承国家 973 项目"自然语言理解的交互引擎研究"(2004CB318104)、中科院声学所知识创新工程项目"HNC 语言知识处理理论及技术"的资助。张 全 博士,研究员,博士生导师,主要研究方向为 HNC 自然语言理解处理技术、计算语言学;吴 晨 博士研究生,主要研究方向为自然语言理解处理技术、软件工程;韦向峰 博士,主要研究方向为自然语言处理技术。

领域世界知识的句类表示。领域句类的获取需要通过分析句群中的各个句子来获取。这包括两个层面的意义：一方面需要根据句类分析的结果判断句群的领域和对应领域代码，另一方面需要将各个句子句类分析的结果转换成领域句类的表示，从句类分析上升到句群分析，获取句群的整体信息。

在语言表达中，特别是汉语的表达中，大量采用语义块共享语句。从句类分析进入句群分析时，需要判定语义块的共享情况，这样一方面服务于句类检验，另一方面为确定语境单元中各项内容奠定基础。例 1 中句子之间的没有出现主语义块共享的情况，但却出现了辅语义块共享的情况："1856 年"，"当特斯拉出生时"，"7 月 10 日午夜"等几个时间条件辅语义块，在整个句群中是共享的；但第二句的"他母亲去做工时"，不是整个句群共享的辅语义块。所以在给出事件背景(BACE)描述时，就要区别处理这两类不同的辅语义块。主语义块共享也会产生同样的问题，因此需要判明句群中所共享的语义块，语义块共享的分析处理是句群分析应当也必须面对的一个重要问题。而完成分析处理的前提是对语义块共享的情况建立系统的分类，并对实际语料中语义块共享类型的分布有准确的了解。本文就是针对这一问题展开的。本文第 2 部分对句群的句子组成情况进行了一个分析和统计。第 3 部分则在此统计分析的基础上，对构成句群的句子间的要素共享情况按特征进行了分类。第 4 部分给出照本文分类标准对各类句群进行统计后获得的一个统计结果。最后对全文进行总结。

2 句群及相关的统计分布

语言学从语言交际的角度对句群已经积累了大量的认识[2,3]：在连贯话语中句群是相对独立的语言单位，它以一定的方式为组合标志，可以从语流中切分出来。一般认为句群由多个语句构成，单个句子不构成句群。文[3]中更明确指出，句群是两个或两个以上的句子的组合体，书面上用两个或两个以上表示句子的标点符号；句群再简单也必须是两个句子。同时对句群结构的分类和分析一般借鉴复句的分类和分析方法，如将句群分成单重和多重，对单重句群进一步采用单重复句的分类，分成并列、递进、选择、假设、总分、因果等等类型。这样作的原因在于[2]：(1)句群的内部结构和复句的内部结构相似，(2)借鉴复句的研究成果，便于人们掌握和了解句群。

上述从语言交际的角度对句群的研究有助于人学习和了解句群，也为从语言交互的角度，即从计算机理解自然语言的角度，研究句群提供了有益的参考[4]。然而，为了更好地服务于计算机自然语言理解处理，还需要进一步界定一些核心概念，使其形式化，以便计算机操作。首先关于句子的定义，一般认为句子是表示相对完整意义的语言单位。HNC 则进一步深化这一定义，提出了句类，用句类知识来具体表述语句的概念联想知识。在句类的指导下，主语义块在语句中的配置则明确给出了语义完整性的要求。其次在界定出句子的基础上，HNC 对句群也给出了明确的形式化描述——表示式 1。特定概念确定了句群所属的领域，句群表述的核心进一步表示为领域句类。

本文对语料分析时，所涉及的句群和句子，均采用 HNC 的定义。另外需要说明的是，根据 HNC 对于句群的定义，一个句子可以是一个句群。例 2 是语料分析的又一个句群实例。本例中包括 4 个句子，对应 4 个全局的句类代码，出现了两个句号，没有语义块共享现象。

例 2 他的发现‖不亚于‖法拉第的电磁感应，＋＋〈而他|对当今世界|的影响〉‖甚至超过了‖爱迪生。＋＋他‖才是‖真正的现代"普罗米修斯"，＋＋他的发现‖依然激励着‖当代研究者。

jD00J ＋＋ jD00J＃DBC1＝〈！ 111XY0J〉＋＋jDJ ＋＋ X291J

本文分析了 3 篇语料：一篇从英文翻译的科学家小传(C1)，一篇政治家生平(C2)，一篇党史讲话(C3)。它们的共同之处是都是关于历史的叙述，但文体和内容有较大的区别。这里分别对 3 篇语料中句子、句群、标示句子的标点符号(简称 句标号)、段落的分布情况进行了统计。表 1 中的字数，以汉字个数计，包括标点符号。

表 1 (句子、句群)－(句号、段落)统计

	句子	句群	字数	逗号	句标号	段落	平均字数/句	平均句数/句群	平均句数/句标号
C1	173	69	6066	134	94	25	35.06	2.51	1.84
C2	665	204	13495	621	314	51	20.29	3.26	2.12
C3	969	247	21399	888	423	115	22.08	3.92	2.29
总计	1807	520	40960	1643	831	191	22.67	3.48	2.17

从表 1 可以看出，HNC 定义的句子，比按照句标号划分的句子短。一个句标号中往往包含了多个 HNC 句子，从统计数据来看平均是 2 个左右。这些句子都是围绕着一个主题进行表述的，属于一个 HNC 定义下的句群。而且在实际的语料分析中，经常是一个句标号切分一个句群：对于 C1，一个句群大约包括 1.3 个句标号，C2 是 1.5 个，C3 是 1.7，总体平均为 1.6。这些数字都没有达到或超过 2，而一般从交际角度确定的句群至少是两个句子，这表明 HNC 对句群的界定更小更细。同时立足于 HNC 的句子和句群，可以解释"一般认为句群和复句结构相近"的现象，因为从 HNC 的观点看，实际上二者面对的都是句群。

统计数据中也出现了一些有趣的现象。C1 句子的字数较 C2 和 C3 多，C2 和 C3 的数值比较接近，这可能与 C1 是翻译文本，而 C2 和 C3 是汉语直接形成的文本有关。这一点从 HNC 的句子观察的更显著一些，对应的数值分别是 35.06、20.29、22.29；而句标号对应的数值分别为 54.53、42.97、50.58，尽管也反映了这一趋势，但是不如按 HNC 句子考察来得显著。另一个有趣的现象是，在 3 篇语料中，C1 每一句的字数最多，但句群中包括的句子数却最少。还有，C1 中句标号内包括的句数显著少于 C2 和 C3。结合句群中包括的句子数，表明 C1 更倾向于一个句标号构成一个句群。最后再来看一组数据——每一个句标号中包括逗号的平均值，C1 为 1.43 个逗号，C2 为 1.98，C3 为 2.10。C1 显著少于 C2 和 C3。如果将上述几点综合起来，可以看到：C1 的句子长，较少的句子构成句群或者说更多的情况在 C1 中是一个句标号构成句群，句标号之间少用逗号。这表明什么？表明 C1 中多用长句子，且一个句子表达意义相对完整。

综上，句群是一个重要的语言单位，HNC 提出了不同于语言交际角度的句群概念。这里结合 HNC 的句子定义，运用句群这一概念对实际语料进行分析调查。统计结果呈现出一些有趣的现象。这也从一个侧面反映出 HNC 句群这一概念进行语言理解研究的价值。

3 语义块共享类型分析

语言在表意时，在明确的前提下经常省略句子的成分。就具体的句子而言，这是省略，但是省略的内容往往会出现在上下文中，语言交际研究将这种情况归入回指(anaphora)中，并定义为零形回指，与其他的代词回指、名词回指构成回指研究的主要内容。对回指的研究是语言学研究的一个热点[5]。如果从 HNC 的句群观点出发，这种现象可以理解为句子语义块的共享。HNC 已经对句子中主语义块共享的情况进行了深入的研究，已经形成了关于共享句的迭句、链句、环句和塔句等概念[6]。

本文将在已有研究的基础上结合 HNC 句子语义块描述，对句群中语义块的共享情况做进一步的分析。

文[1]中给出了 HNC 句子的语义块表示式如下：

表示式 2　SC=GBK1+EK+GBKm(m=2−3)

SCR=SC+fKm

其中：SC 是句类空间的数学描述，SCR 是增加辅语义块的表述。在 HNC 句类中，句子主语义块的类型和数量由句子的概念联想脉络确定，一般不超过 4 个；辅语义块根据句子表述的需要添加，弱依赖于句子概念联想脉络。

根据表示式 2，可以演绎出句群中两个句子语义块的共享情况，而多于两个句子的情况可以作为两个句子语义块共享的叠加处理。

表 2　语义块共享基本类型

	整体共享	部分共享	
		非句蜕	句蜕
主语义块	Ⅰ型	Ⅱ-1 型	Ⅱ-2 型
辅语义块	Ⅲ型	Ⅳ-1 型	Ⅳ-2 型

表示式 2 中有两类语义块：主语义块(GBK 和 EK)和辅语义块(fK)。句群中这两种语义块都可能出现共享的情况，因此可以首先分成主语义块共享、辅语义块共享两种语义块共享类型。主语义块又包括两类：广义对象语义块 GBK 和特征语义块(EK)。特征语义块集中体现一个句子的语句级概念联想类别，如果特征语义块共享，则句子概念联想脉络的激活点缺失，句子将很难判读和理解。当然，语言现象非常复杂，不排除在特定语境下出现特征语义块共享的句群，但至少这种情况比较罕见，因此本文对特征语义块共享的情况不讨论，这里对主语义块共享情况的分析研究仅限于广义对象语义块。语义块具有自身的内部构成，因此，语义块共享还需要从另一个角度考察——整体共享和部分共享。这样，就形成四种语义块共享的类型。对于部分共享的情况，由于语义块构成存在句蜕和多元逻辑组合两种情况，因此需要进一步区分这两种情况。具体情况见表 2。

对于表 2 中的各种语义块基本共享类型还需要进一步分析说明。在 HNC 理论中，主语义块和辅语义块是两类不同性质的语义块，在句群中不存在主语义块和辅语义块整体交叉共享的情况，即在句群中一个句子将另一个句子的辅语义块整体作为自己的主语义块共享，或反之。因此，语义块的整体共享只可能是主语义块和主语义块的共享，辅语义块和辅语义块的共享。部分共享的情况比较复杂，应当出现在语义块具有复杂构成的情况下，复杂构成包括句蜕。因为语义块构成复杂，故而可以有一部分内容被句群中另外的句子共享，或者作为语义块的全部，或者作为语义块的一部分，因此，辅语义块和辅语义块之间、主语义块和主语义块之间可以存在部分共享；辅语义块和主语义块之间也可以存在部分共享。这里只是可能性的分析，语言的表达不可能是这些可能的排列组合，本文根据实际语料分析的情况，对部分共享的情况做了简化，其中Ⅱ型共享特指主语义块间，不考虑主辅交叉的情况；Ⅳ型特指辅语义块的部分被主语义块共享的单向情况，而且此时只考虑辅语义块的部分充当整体主语义块一种情况。简言之，对于辅语义块共享，在本文中只考虑了一种整体共享的情况。

主语义块共享中，除了辅语义块部分被作为主语义块共享的情况外，无论整体共享还是局部共享，都存在着主语义块共享位置的变化，即表示式 2 的 SC 等式中一个句子的 GBKm 被共享后充当另一个句子的 GBKn，这里 m 可以等于 n，也可以不等于 n。表 3 对此进行了具体的分析。这里以句子最多具有 3 个广义对象语义块的情况进行说明。

表 3　广义对象语义块共享基本类型

		共享句子		
		GBK1	GBK2	GBK3
源句子	GBK1	11	12	13
	GBK2	21	22	23
	GBK3	31	32	33

表 3 中，源句子指提供共享语义块的句子，即共享的语义块在这个句子中；共享句子是共享其他句子语义块的句子，表中给出了各种主语义块共享情况的编码。表 3 中 11 就是迭句；如果源句子有 GBK3 则 31 构成链句，如果只有 GBK2 则 21 是链句；对于源句子和共享句都是 3 个广义对象语义块的句子，33 是塔句，否则根据不同情况 22，32，23 都可能是塔句；可能是环句的代码包括 12，13。如果考虑源句子和共享句子都是两个广义对象语义块的句类，那么迭句、链句、环句和塔句正好是对主语义块共享情况的封闭描述。但是考虑 3 个广义对象语义块的句类，则情况就复杂起来了。另外还需要说明的是，在句群语义块共享中，并不要求源句子和共享句子具有相同的广义对象语义块个数。

对于主语义块部分共享的情况，还可以进一步考虑源句子主语义块的部分充当共享句子语义块的部分、整体，以及源句子语义块整体充当共享句主语义块部分的 3 中情况。

综上，这里通过对 HNC 句子的语义块表示式的演绎分析，对句群中语义块的共享情况进行比较全面的分类，并作了简化。下文对语料语义块共享情况的调查就是以此为基础的。

4 语义块共享类型的分布

表 4 是本文对所选取的语料进行语义块共享类型分布情况的统计数据，语料的情况和 2 中相同。例 3～例 7 是语义块共享各类型的实例。

例 3 迭句　他‖拥有了‖自己的公司——特斯拉电气公司，＋并争取获得‖\〈以交流电为基础的|新电气技术〉的专利/。

R611J ＋ XYa0 * 21J # YC=〈! 111XY0J〉

说明：句群中第一句的“他”被第二句共享作为 GBK1 共享。

例 4 链句　特区的高速发展‖带动了‖全国的对外开放，＋～形成了‖全面开放的新格局，＋～有力地促进了‖

改革和现代化建设事业。

P21J ＋～ XY0＊21J ＋～ XY60＊21J

说明：这个句群有两个链句，第一句的“全国的对外开放”被第二句共享作为 GBK1 共享，第二句的“全面开放的新格局”被第三句共享作为 GBK1 共享。

例 5 环句 无论什么样的生产关系和上层建筑‖～，[-都要-]随着生产力的发展‖而发展。＋＋ 如果它们‖不能适应‖生产力发展的要求，＋而成为‖\{生产力发展}和{社会进步}的障碍/，＋＊ 那就必然要发生调整和变革。

Y4J ＋＋ R711Y01＊211J ＋ Y02J＃YB2＝\{Y4J}{Y4J}/ ＋＊ ！31XY60＊21J

说明：句群的最后一句 GBK1 是社会，省略了；GBK2 就是“它们”，即不适应生产力发展要求的生产关系和上层建筑，将前一句的 GBK1 作为自己的 GBK2，形成环句。

例 6 辅语义块部分共享句（Ⅳ-2 型） {刘邓大军|进入|大别山地区}后～‖～，对国民党在长江以南的广大统治区‖形成了‖直接威胁，

Cn&ReB！31XY10＊21J＃Cn＝{T2bJ}

说明：全局 GBK1 省略，实际上是共享了原型句蜕形式条件辅语义块的 GBK1；全局 GBK1 共享辅语义块的部分。

例 7 主语义块部分共享句（Ⅱ-1 型） 他‖化名‖邓斌，＋任‖中共广西前敌委员会书记，＋％ 同张云逸等‖～ 于 12 月～‖发动‖百色起义

D01Y02＊20J ＋ X10J ＋％ XY10＊21J

说明：第三个句子的 GBK1 在句子中只给出了一部分，“同”前面的内容应当是“他”“邓斌”。非句蜕形式的主语义块部分共享。

表 4 汉语句群中语义块分布数据

	Ⅰ型(句群)		Ⅲ型(句群)		Ⅳ型(句)		Ⅰ型(句)				Ⅱ型(句)	
	群数	%	群数	%	Ⅳ-1	Ⅳ-2	迭句	链句	环句	塔句	Ⅱ-1	Ⅱ-2
C1	27	39.13	8	11.60	0	0	30	2	0	0	1	3
C2	110	53.92	19	9.31	0	6	231	23	1	1	9	2
C3	144	58.30	30	12.15	1	0	488	15	1	0	7	0
总计	281	54.04	57	10.96	1	6	749	40	2	1	17	5

从表 4 中的统计数据可以看到：

(1)大量的句群中存在主语义块共享现象。不同的文体，出现主语义块共享句群的情况差别显著，C1 中出现主语义块共享的百分比小于 C2、C3，C2 和 C3 出现主语义块共享的情况相近。这与 C1 的句子长，句标号之间少用逗号且一个句子表达意义相对完整的表述风格有关。

(2)从数据来看，辅语义块共享的情况没有主语义块共享情况普遍。引起这一现象的一个原因是句群中辅语义块的数量远远少于主语义块，有些句群甚至不出现辅语义块。在语料标注的过程中发现一般涉及时间地点范围等的条件辅语义块往往在句群中共享，为句群描述的主题提供叙述背景。这里三篇语料中辅语义块共享在句群中所占的比例相近。

(3)在主语义块整体共享中，迭句的情况占据了绝大多数的情况；链句在标注的语料中偶尔可以见到，而其他类型的主语义块共享情况，比较罕见，而且经常伴随其他共享或省略出现，如例 5。

(4)主语义块存在部分共享的情况，这种情况不常见。在本文研究的语料中这种共享方式的句子不多，因此没有具体再细分成 3 类情况分别考虑。这里，遇到的部分共享情况都属于部分-整体共享，即在一个句子中的一部分共享为另一个句子的主语义块整体，或反之；没有遇到有部分－部分情况的共享。在统计数据中非句蜕复杂构成的主语义块共享情况，多于句蜕的复杂构成的主语义块共享情况。

(5)在辅语义块部分共享中只考虑辅语义块的部分共享为主语义块整体的情况。这类情况集中出现在 C2 这篇语料中，同时 C2 中这类共享全部来自辅语义块是句蜕构成的情况。

综上，这里对上一节的分类进行了具体的语料验证，给出了部分典型例句和共享类型的分布数据及其分析。从所得结果来看，本文的分类情况比较全面地反映了汉语句群语义块共享的情况，适宜于进行汉语句群分析使用。

结束语 句群是自然语言中一级重要的语言单位，本文根据 HNC 理论关于句群和句子的定义，首先明确了 HNC 句群和语言交际研究中句群概念的差异，其次运用 HNC 句子和句群概念具体统计了语料中句子、句群、标示句子的标点符号等的分布数据，并对这些数据进行了解读。这些数据也说明了 HNC 的定义能够比较好地反映语句的特征。同时，在迭句、链句、环句、塔句概念的基础上，对语义块共享问题进行了进一步的分析：以演绎为基础分析了共享的类型，结合实际语料验证分类的合理性，统计了各种类型的分布数据，并对数据进行了分析。本文的结果将服务于语句之间成分共享和句群的计算机自动分析。

本文的研究工作还需要进一步深入。首先，应借鉴语言交际的研究成果，将共享研究和指代研究结合起来联合攻关。其次，需要进一步在本文分类的基础上研究共享和指代的内容恢复策略。就共享而言，根据本文的分析，应当重点解决好迭句。第三，需要研究句群之间语义块的共享情况，本文在语料分析中观察到这一现象，但没有进行具体的调查。

致谢 本文的诸多认识得益于黄曾阳教授的句群语料标注及注释，在此谨致谢意。

参 考 文 献

1 黄曾阳. 语言概念空间的基本定理和数学物理表示式[M]. 北京：海洋出版社，2004
2 吴为章，田小琳. 汉语句群[M]. 北京：商务印书馆，2002
3 庄文中. 句群[M]. 北京：人民教育出版社，1990
4 吴晨，张全. 自然语言处理中句群划分及其判定规则研究[J]. 计算机工程，2006
5 徐赳赳. 现代汉语篇章回指研究[M]. 北京：中国社会科学出版社，2003
6 韦向峰. 基于 HNC 理论的扩展句类分析平台研究：[博士学位论文]. 北京：中科院声学所，2005
7 基于 HNC 的语句概念结构语料库标注规范(草案稿)[S]. 2004.9