汉语和英语逗号的对比分析及其翻译处理*

张　全

（中国科学院 声学研究所 语言语音部 100080，E-mail: zhq@mail.ioa.ac.cn ）

摘要：逗号在书面语表达中起着非常重要的作用。汉语和英语的逗号用法有差异，翻译中逗号处理不当会严重影响译文质量。因此，在机器翻译中需要对逗号进行相应的处理。本文运用 HNC 的视点对汉英两种语言中逗号的用法进行了详细分类，结合真实语料调查了各种用法的分布情况，对比研究了两语种中逗号用法的异同。同时，本文调查了汉英对照语料中逗号主要用法的对译情况，并给出了汉英逗号的翻译准则，统计数据与准则符合。

关键词：机器翻译，HNC 理论，逗号分类，汉英翻译中的逗号处理

The Comparison of Comma between Chinese and English and the Translation Processing of Comma

ZHANG　Quan

(The Institute of Acoustics 100080，CAS, 100080, E-mail: zhq@mail.ioa.ac.cn)

Abstract: Comma plays an important role in written language. There are many differences between Chinese and English, and the errors of comma will make the translation unreadable. Therefore, the Chinese-English machine translation system needs a step to process the comma. The main part of this paper discussed the difference of comma between Chinese and English and the principles of translation. It includes a new category of comma usage, the statistical data on real corpus about the category, the comparison of comma between Chinese and English and some principles of comma on translation from Chinese to English. The HNC (Hierarchical Network of Concepts) is the foundation of this paper.

Keyword: Machine Translation, the theory of HNC, the category of comma, the processing of comma on translation from Chinese to English

1 引言

书面语以文字的形式来表达和传递信息。为了正确完整地表达语义，需要使用标点符号。

*本文承 973 项目(G1998030506)、国家语言文字应用"十五"重点项目"语句语义标注规范研究"（ZDI105-43C）和中国科学院声学研究所创新项目的资助。

当然在特殊语境中也有不使用标点符号的，如某些为了特殊表达需要的文学作品、古汉语等，这时对于文章的理解主要借助于上下文的语义关系。但现代社会中大量流通的文字材料，都具有标点符号。因此，在计算机理解自然语言的过程中，处理好标点符号对于正确获取语义信息具有重要意义。在机器翻译中，标点符号的处理对于译文的质量，同样具有重要意义。

例 1

邓小平在中国人民政治协商会议第一届全体会议上，被选为中央人民政府委员。

Deng Xiaoping is at the first plenary session of the Chinese People's Political Consultative Conference, Elected as the committee member of People's Central Government.

Deng Xiaoping was elected a member of the Central People's Government at the First Plenum of the Chinese People's Political Consultative Conference.

例 1 中的斜体英文是用机器翻译软件翻译的结果，下面是作为对比的人工翻译结果。从中可以看出，机器翻译的结果由于对逗号处理不当，造成译文错误。

逗号在书面语中最常出现，同时它的用法也灵活多样，给计算机理解处理带来很大的困扰。本文拟采用 HNC 的视点对汉语和英语中逗号的用法及其翻译处理中的准则进行探讨。

这里说明一下本文统计使用的语料。本文使用的语料主要是新闻类语料，对汉语逗号的分析，使用的是原文为汉语的语料；英语的情况相同。其中汉语语料 23288 字（含标点符号），英语语料 7478 词（不含标点符号）。这样作的目的是为了更好反映逗号在该语种中的表现。翻译语料采用的是由汉语到英语的语料，以汉语计 7743 字（含标点符号）。另外，统计点号使用情况的语料多于上述的语料，汉语为 1136258 字，英语 1080424 词。

2 关于逗号

根据标点符号的作用，可以将它们分成两类：点号和标号。点号可以进一步分成句末点号和句中点号。句末点号包括：句号、问号、叹号；句中点号包括：逗号、顿号、分号、冒号。在点号中，逗号最常出现，用途最广[1][2]。在本文使用的语料中，各类点号出现的情况如表 1 所示。从表中的数据也可以看到无论在汉语还是在英语中，逗号的出现率都是第一位的。

表 1 中英文点号出现频率统计

	句号		问号		叹号		逗号	
	频数	频率	频数	频率	频数	频率	频数	频率
中文	20542	27.8%	109	0.1%	139	0.2%	37607	50.8%
英文	56232	47.7%	874	0.7%	86	0.1%	57758	49.0%

	顿号		分号		冒号		总计	
	频数	频率	频数	频率	频数	频率	频数	频率
中文	12150	16.4%	1467	2.0%	1972	2.7%	73986	100%
英文	/	/	792	0.7%	2105	1.8%	117847	100%

为了规范现代汉语中标点符号的使用，我国出台了《标点符号用法》[3]。其中说明了使

用逗号的四种情况，分别为：句子内部主语与谓语之间如需停顿；句子内部动词与宾语之间如需停顿；句子内部状语后边如需停顿；复句内各分句之间的停顿。当然，这里的说明只是框架性的，具体汉语中逗号的用法要更多一些。[1]在[3]的基础作了更详细的分类和说明。英文逗号的用法，[2]的归纳较为全面，共计 28 种。

综上，逗号很常用，具有较大的模糊。语言学在逗号用法方面已经作了大量的研究。这些研究主要是面向人使用逗号的。

3 逗号用法类型及其分布

HNC 要建立的是语言概念空间，在这里用突现概念之间关联性的概念化的语义基元网络表示语义，用句类和语义块表述句子的句义。主语义块的语义特征和配置与句类密切相关，是句类知识的重要内容[5]。简单地说，在计算机理解语句的过程中，可以根据句类对主语义块的预期知识及主语义块出现的情况判定句子句义的完整性，为语句级的理解提供了基础。同时，也将[3]中关于句子定义的“表示相对完整意义的语言单位”内容进一步深化，便于计算机的处理。

在考察逗号的用法时，有两类语义基元网络需要给予特殊的关注：一个是语言逻辑概念网络，一个是语法（习）类概念网络。前者主要表述句义结构的各种标志符和说明符，包括句间的语义结构符号（*lb*）,对应于传统的介词和句间连词等。后者主要表述语言中的多种习惯用法和句子的附属成分，包括插入语、独立语以及传统语言学意义上的句子语气类别（在 HNC 中用 f 表示）等等。

对于句群，HNC 使用了“叠句”和“合句”两个概念：叠句，具有两个或两个以上独立全局特征语义块、共用至少一个广义对象语义块的句子集合；合句，具有两个或两个以上独立全局特征语义块、不共用广义对象语义块的句子集合。在下面的例句中“‖”表示语义块的切分位置，“+”表示叠句的分隔，++表示合句的分隔。例句中有下划线的部分为对应情况的示例。

汉语逗号的用法及其分布。

C-a) 用于叠句之间(S1，!3nS2)。

泰国的年人均国民生产总值增长率 ‖ 达 ‖ 百分之八点二，+ 居 ‖ 世界各国和地区的首位，++ 韩国 ‖ 以百分之七点八 ‖ 居 ‖ 第二位，++ 中国和新加坡 ‖ 以百分之六点九‖ 并列 ‖ 第三位。

C-b) 用于合句之间(S1，S2)。

C-a 例中的另外两个逗号就是这种情况。

C-c) 用于辅块与主句之间(fK,S)。

据世行统计 ‖，１９９４年 ‖ 中国的国民生产总值‖ 为 ‖ ６３０２．０２亿美元。

C-d) 语义块的复杂构成(J(f)K=K1+K2，而 K1，K2)。

包括两种情况：1）主语义块的复杂构成；2）辅语义块的复杂构成。

“吃水要分用，大人一铁勺，小孩一调羹”‖ 是 ‖ 全州水源短缺的真实写照。(主语义块)

C-e) 独立语(f2，S)。

更令人可喜的是，人均收入快速增长的发展中国家 ‖ 增多。

C-f) 主块加辅块(JK1+fk，S)。

一大批企业 ‖ 通过技改调整产品结构 ‖ ，_提高了 ‖ 市场占有率，+焕发出 ‖ 新的生机。

用法中还有：

C-g) 在块扩部分多个扩展子句之间(Sr1，Sr2)，C-h) 插入语(f1，S)，C-i) 补充及列举(S，f41)，C-j) 短语的并列成分(fFK1，fFK2)，C-k) 非基本格式时语义块之间(JK1+^JK2，S)，C-l) 句间连词(/b，S)，C-m) 判断逻辑(j/uv1，S)，C-n) 重复指代(f84，S)，C-o) 呼语(f3，S)，C-p) 一般感叹(f51，S)，C-q) 含主块的一般感叹(JK1+f51，S)

下面是汉语逗号用法的分布情况。由于 C-m～C-q 主要用于口语，这里没有出现。

表 2 汉语逗号用法分布频率统计（共计 1008 个逗号）

C-a		C-b		C-c		C-d	
频数	频率	频数	频率	频数	频率	频数	频率
321	31.8%	252	25%	194	19.2%	78	7.7%
C-e		C-f		C-g		C-h	
频数	频率	频数	频率	频数	频率	频数	频率
67	6.6%	25	2.5%	23	2.3%	20	2.0%
C-i		C-j		C-k		C-l	
频数	频率	频数	频率	频数	频率	频数	频率
11	1.2%	7	0.7%	7	0.7%	3	0.3%

从表中数据可以看到，C-a 和 C-b 实际上是逗号作为句子之间分隔符使用的情况，他们所占比例超过一半。在句子内部，逗号主要用于分隔辅块、语义块的复杂构成中和独立成分；另外，这里主块加上辅块（JK1+fK，编号 C-f）后使用逗号的情况也比较多见。最后，上面的统计结果也说明了 HNC 在拼音处理时期考虑了汉语逗号使用的最常见情况[4]（C-a～C-e 超过 90%）。

英语逗号用法采用的分类和编码与汉语一致，遇到英语中没有的情况，则跳过相应的编码；英语中特有的逗号用法，则延续上面的编号。

E-a) 用于叠句之间(S1，!3nS2)。

Kongdan Oh and Ralph C. Hassig ‖ review ‖ the policy options,__+ underscore ‖ their flaws.

E-b) 用于合句之间(S1，S2)。

The answer ‖ is probably ‖ yes,_++ if a price ‖ can be agreed upon.

E-c) 用于辅块与主句之间(fK,S)。

But in Baltimore ‖ _for example, 14.2 percent of babies born in 1998 ‖ were ‖ low-birthweight.

E-d) 语义块的复杂构成(J(f)K=K1+K2，而 K1，K2)。

More importantly, the Tesla Coil ‖ opened ‖ the way to radio and TV transmission,_and to Tesla's most amazing discovery: the transmission of electrical power through thin air. (主语义块)

E-e) 独立语(f2，S)。

见前面 E-c 例句中的第二个逗号。

其他还有：

E-g) 在块扩中子句之间(Sr1, Sr2)，E-h) 插入语(f1, S)，E-i) 补充及列举(S, f41)，E-j) 短语的并列成分(fFK1, fFK2)，E-k) 非基本格式时语义块之间(JK2, S)，E-l) 句间连词(/b, S)，E-o) 呼语(f3, S)，E-p) 一般感叹(f51, S)，E-r) 对仗型省略(S1, S2；S2 为省略句)，E-s) 英语

特有的、固定用法，计 9 项。

表 3 英语逗号用法分布频率统计（共计 505 个逗号,表中未包括“E-s”，3.0%。）

E-a		E-b		E-c		E-d	
频数	频率	频数	频率	频数	频率	频数	频率
57	11.3%	106	21.0%	88	17.4%	21	4.1%
E-e		E-g		E-h		E-i	
频数	频率	频数	频率	频数	频率	频数	频率
48	9.5%	10	2.0%	16	3.2%	60	11.8%
E-j		E-k		E-l		E-r	
频数	频率	频数	频率	频数	频率	频数	频率
62	12.3%	4	0.8%	13	2.6%	5	1.0%

上面分别考察了逗号在汉语和英语中用法的分布，这里结合两种语言作一点说明。

首先，根据本文的分类，逗号在汉语和英语中的用法相近。表现如下：

1） 它们的主要用法相同，从它们的用法分布表中可以看出。

2） 它们的用法分布相近。

3） 汉语和英语都出现了一定数量的独立语。因为新闻报道注重客观，常常要给出信息的来源以及有关观点的出处。

其次，统计数据也表明汉语和英语在逗号运用上存在差异。

1） 汉语表达中有“主块加辅块(JK1+fk，S)”（C-f）后用逗号的情况，而英语没有，这个现象是汉语特有的。

2） 英语中的叠句少于汉语，说明汉语更多地使用句子之间的语义关联，而英语更注重形式的完整。

3） 英语中“短语中的并列”（E-j）出现频度也大大高于汉语，这可能是由于汉语中有顿号，而英语中没有顿号的原因造成的。

4 汉英机器翻译中的逗号处理

表 4 是本文选取的汉英对照语料汉语源语言中逗号用法的分布情况。

表 4 汉语源语言逗号用法分布频率统计（共计 352 个逗号）

C-a		C-b		C-c		C-d	
频数	频率	频数	频率	频数	频率	频数	频率
146	41.5%	39	11.1%	105	29.8%	28	8.0%
C-f		其他		总计			
频数	频率	频数	频率	频数		频率	
11	3.1%	23	6.5%	352		100%	

这里只列出了出现频度较高的主要用法，这也是下面要具体分析的用法，除此之外的其他用法归入“其他”中。这里共计有 352 个逗号，在英语中有 103 个逗号在对应的位置上消失，占 29.3%；有 27 个变成了句号，占 7.7%。

对于上述五种情况，根据逗号断开的是句还是非句，将他们分成两组：叠句（C-a）合句（C-b）为一组，辅块（C-c）语义块复杂构成（C-d）以及主块加辅块（C-f）为一组。由于C-f是汉语种特有的现象，应当独立出来考虑，成为第三组。

在第三组中，汉语中与辅块连在一起的主块构成简单，而辅块构成比较复杂。这种情况类似于第二组中C-c逗号的用法，实际是为了给出辅块的下边界。辅块的上边界，由于主块构成简单，辅块自身又具有标记，不需要用标点符号给出。这里的辅块还应该分成两类：句蜕构成和非句蜕构成。在翻译时，这两种情况的主块都要放到主句中去，而辅块的位置有区别。句蜕构成的辅块将放到主句之前，用逗号与主句隔开。而非句蜕构成的辅块，则直接接到主句后面，不用逗号。这种方式有利于平衡句子的结构，也使句义清晰。例1中的句子，就属于后一种情况。

在第二组中，部分情况与第三组相同，即可以按照语义块构成是否有句蜕来确定逗号的出现与否。这在C-c中符合的比较好，而C-d中有所不同。对于C-d，还要看句蜕结构的核心，如果核心是蜕前句的特征语义块，则会将整个块放到主句中，同时对应特征语义块中的v（动态）概念也以g（静态）概念的形式出现。如果核心是蜕前句的广义对象语义块，或者是原型句蜕，则逗号仍然存在，同时v（动态）概念以非谓语动词的形式出现。考察的语料中辅块的句蜕类型都是原型句蜕。最后，英语中的and很常用。汉语中语义块内部构成的成分之间用逗号隔开，而译成英语时，最后一个逗号会被and替代。如果只有两部分，则英语中只用and，而不用逗号。

在第一组中，如果逗号断开的汉语叠句共用的是第一个句子的JK1（第一个广义对象语义块），则英语中逗号也出现在叠句之间。如果共用第一个句子的JK2（第二个广义对象语义块），则英语要用代词补齐汉语中省略的部分，把句子从汉语的叠句变成英语的合句，合句之间有逗号。如果叠句之间存在句间连词，则也要转换成英语的合句。这是由于英语的形式约束严格，叠句用途受限。另外，英语中也有用and替代逗号连接叠句和合句的情况。

考察汉英翻译语料的逗号情况，也较好地符合了上述原则。逗号转换的具体情况如下：

1） 叠句(C-a)经常转为：叠句(E-a，43.2%)，去掉逗号(32.2%)，合句(E-b，8.9%)，句号(7.5%)，辅块(E-c，4.1%)，短语并列(E-j，1.4%)。

2） 合句(C-b)：合句(E-b，30.8%)，去掉逗号(28.2%)，句号(23.1%)，叠句(E-a，7.7%)，辅块(E-c，7.7%)，补充列举(E-i，2.5%)。

3） 辅块(C-c)：辅块(E-c，71.4%)，去掉逗号(20.9%)，叠句(E-a，2.9%)，其他（合句、补充列举等，4.8%）。

4） 语义块复杂构成(C-d)：去掉逗号(42.9%)，语义块复杂构成(E-d，39.3%)，叠句(E-a，7.1%)，句号(7.1%)，辅块(E-c，3.6%)。

5） 主块加辅块(C-f)：辅块(E-c，54.5%)，去掉逗号(27.3%)，合句(E-b，18.2%)。

语料中不符合的情况，主要原因有两个，一是翻译中发生了句类与格式的转换，另一个是译者的个人倾向。对于后者，按照上述原则处理逗号，也是大局不错，不致产生类似例1的错误。

此外，在“其他”情况中，汉语的逗号可以直接对应为英语的逗号，包括：插入语、独立语、句间连词和补充列举等情况，主要原因在于这些句子成分无论在汉语还是英语中都是

相对独立的。汉语中的顿号也直接翻译成英语的逗号，不过最后一个顿号要使用 and。

5 结语

本文的有关内容还需要进一步深化。首先，需要结合句类与格式转换来进一步研究逗号的翻译处理。其次，本文目前只考察了从汉语到英语的情况，还需要进一步考察英语到汉语的逗号处理。第三，在汉英翻译中除了源语言中逗号对应到译文中外，还需要根据实际情况添加逗号。对于这个问题，除了前文提到了英语特有的、固定的逗号用法(E-s)可以利用外，还要结合句类进行研究。

对比两种英语（由汉语翻成的英语和原文本身就是英语）逗号的用法，可以看到向源文“趋同”的现象。为什么这样说？因为从前面的数据可以看出，英语为源语言的语料中，叠句的比例并不高，低于合句的比例。而在这里汉语语料中逗号用于叠句的比例高于合句，同时(C-a)转换成(E-a)的比例也高于合句的情况，所以对应的英语中，叠句也多于合句。进一步考察语料，印证了这一推断，在汉英对照双语语料中，英语中逗号用于叠句的频率为 20.2%，用于合句的频率为 8.2%。这里有一个问题，“趋同”可能导致译文不够地道、没有原汁原味的感觉。但有一点可以肯定，就是这种翻译可以被接受。如果按“信达雅”来区分翻译的境界，那么最基础的应是“信”，机器翻译也应当先作到这一点。

另外，如果将本文的研究应用于机器翻译中，首先要弄清楚逗号在源语言中的用途。如前所述，逗号在现代汉语中被赋予了太多的功能，因此消解逗号用途的模糊就成了首要问题。在这个问题中，确定逗号分隔的是句子还是非句子又是首当其冲的。当然，也可以另辟蹊径，限制逗号的用途。黄曾阳教授就曾呼吁将逗号仅用于断句。其出发点正是考虑到逗号的复杂现象和计算机处理自然语言能力之间的鸿沟而要帮计算机个忙。当然要让人接受这个呼吁并非易事。不过笔者以为，在各种语料库标注中，应当区分逗号用途中句与非句的不同，不然真是“句读（dou4）之不知”，计算机之难解了。

参考文献

[1] 吴直雄，《实用标点符号手册》，国际文化出版公司　北京 1996

[2] 语恒，《英语的标点符号 [海外中文图书]：正确使用标点符号，提升英语阅读与写作技巧》，笛藤出版图书公司，台北 2001

[3] “现代汉语规范词典”编写组编，“GB/T 158334-1995 标点符号用法”，《语言文字规范使用指南》，上海辞书出版社，上海 2001

[4] 黄曾阳，《HNC（概念层次网络）理论》，北京：清华大学出版社，1998。

[5] 黄曾阳，“HNC 的发展和未来”，《HNC 与语言学研究（张全 萧国政等编）》：52~68. 武汉：武汉理工大学出版社，2001。

注：本文的诸多认识得益于黄曾阳教授的指点，在此谨致谢意。