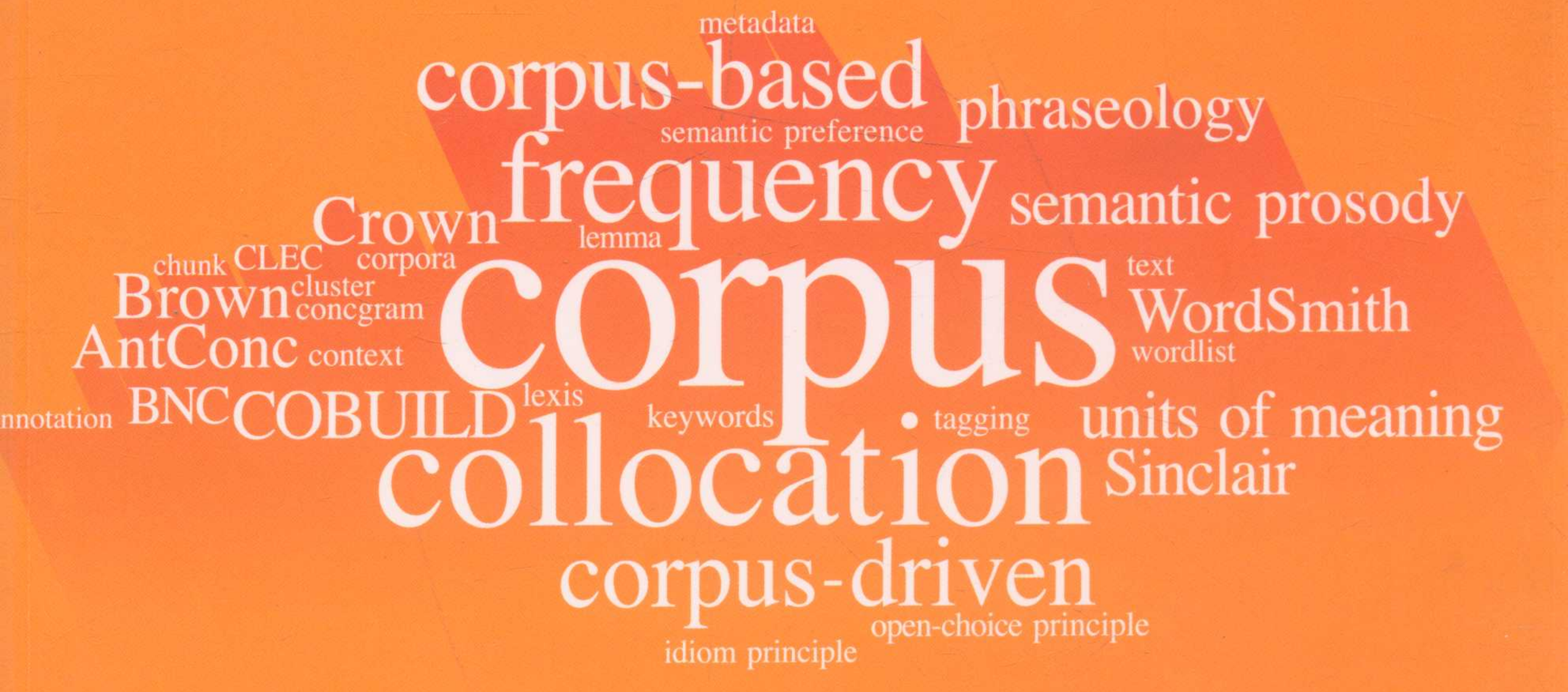
《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

语料库语言学

CORPUS LINGUISTICS

1 Vol. 3 No.1
第3卷 第1期
2016

北京外国语大学中国外语教育研究中心
梁茂成 许家金 主编



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

图书在版编目 (CIP) 数据

语料库语言学. 2016. 1 ：汉、英 / 梁茂成，许家金主编. -- 北京 ：外语教学与研究出版社，2016.6
ISBN 978-7-5135-7762-5

Ⅰ. ①语… Ⅱ. ①梁… ②许… Ⅲ. ①语料库-语言学-汉、英 Ⅳ. ①H0

中国版本图书馆 CIP 数据核字 (2016) 第 144599 号

出 版 人　蔡剑峰
责任编辑　毕　争
执行编辑　解碧琰
封面设计　外研社设计部
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路 19 号（100089）
网　　址　http://www.fltrp.com
印　　刷　中国农业出版社印刷厂
开　　本　787×1092　1/16
印　　张　7.5
版　　次　2016 年 6 月第 1 版　2016 年 6 月第 1 次印刷
书　　号　ISBN 978-7-5135-7762-5
定　　价　12.00 元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
法律顾问：立方律师事务所　刘旭东律师
　　中咨律师事务所　殷　斌律师
物料号：277620001

《语料库语言学》
2016年　第3卷　第1期

目　录

研究论文

研制开发

书刊评介

CORPUS LINGUISTICS

Volume 3, Number 1, 2016

Table of Contents

从推理类话语标记的演化看翻译与现代汉语的互动*

曲阜师范大学 秦洪武 刘丹丹 杜肖颖

提要：本文以汉语推理类话语标记为例，基于汉语历时类比语料，考察翻译在汉语话语标记历时变化中所起的作用及其作用方式。研究发现，较文言文和旧白话文，现代汉语使用话语标记的频率更高，组织手段也更多样，这一变化与翻译，尤其是20世纪初的白话文翻译部分相关。翻译影响汉语话语标记使用的方式多样，但总的说来汉语总是有选择地接受翻译语言的影响，汉语语言手段被翻译调用、发挥并最终影响汉语的发展。

关键词：推理标记、类比语料、现代汉语、翻译语言

1. 引言

我们组织言语表达时通常要明示言语的进程（开始、结束或者过渡）、话题、上下文关系、个人态度等等，这时需要调用各种手段，而最常使用的是语言手段，如词、短语、小句等，这类语言手段通常称作话语标记（discourse markers）。话语标记具有程序意义，有助于形成语篇的连贯性与条理性，并起到一定的指示与提示作用。

根据秦洪武、王玉（2014），现代汉语的话语标记在二十世纪发生了明显的变化，甚至有一些古代汉语和旧白话里没有使用过的话语标记形式频繁出现在现代汉语里，而且这种变化和同一时期汉语翻译语言的变化同时发生。这一变化对于现代汉语交际功能的完善有重要意义，但长期以来，学界对它关注很少，更缺少系统的分析。本文基于历时对比语料库，通过考察汉语推理类话语标记，分析话语标记语在汉语中使用的变化过程，并探讨翻译在其中所起的作用。

2. 推理类话语标记

Fraser（1999）将话语标记分为两类：关联语段信息的话语标记和关联话题的话语标记。有些关联语段的话语标记归纳总结上文信息，推测原因或者引出结论，

*本研究为国家社科基金重大项目“大规模英汉平行语料库的建立与加工”（10&ZD127）的阶段性成果。

如“因此”、“综上所述”、“显然”等，我们称之为推理类话语标记（inferential discourse markers，IDM）。在英语里，这类标记可以使用词（therefore、hence）、词组（in conclusion、in total）或者小句（to sum up、whatever happens）。

现代汉语频繁使用推理类话语标记，这种篇章组织手段的使用频率和组织方式与英语很相似。这一变化既是语言自身发展的结果，又有诸多外部力量推动，翻译可能是其中的一个重要因素，这也是本文探讨的问题。

3. 翻译语言与目标语言的互动：编码复制理论

现代汉语早期发展包含两个重要阶段：晚清白话文运动和五四时期的白话文学。在五四白话文学时期，汉语变化最为明显，大批支持新文化运动的学者提倡通过翻译“创造出许多新的字眼，新的句法”以改造汉语。傅斯年（1918）欧化主张最为明确、激进，他提出改造汉语的三条途径：（1）读西洋文学和从西洋文译过来的文本；（2）自己直译获得新的表达方式；（3）坚持欧化，做文章时，运用读、译西文所得仿造西文。

怎么仿造呢？在汉语里直接植入西洋语法（如屈折变化）会导致系统混乱，全部移植就更不可能。可行的做法是按汉语的语言规范局部移入。Johanson提出的“编码复制理论”就涉及局部引入问题。该理论认为，翻译作为一种间接语言接触，会引发复制行为（copying），即基本码（basic code，即目标语）使用者复制模型码（model code，即源语）中的成分（Johanson 2008：62）。汉语欧化就是模型码欧洲语言（主要是英语[1]）的表达形式通过翻译被复制进基本码汉语（秦洪武、王玉 2014）。但这种复制不是简单的拷贝，也不是更替原有代码，而是将源语中的代码成分嵌入目标语代码，是一个顺应过程，形成两种代码间的互动（Johanson 2008：62）。从这个角度说，模仿是有选择的，而模仿中语言性质的变化是语法复制，这一变化也就是Johanson（2008：62）所说的“选择性语法复制”（selective grammatical coping），包含词形复制（material copying）、语义复制（semantic copying）、搭配复制（combinational copying）和频率复制（frequential copying）（董元兴、赵秋荣 2012）。从理论上说，话语组织方式上的复制也应如此（秦洪武 2010）。

翻译一般只是复制源语中的一个或多个特征，而且，新引入的表达形式都要经过严格筛选以适应汉语自身的规律，这样的互动过程才会最终引发形态-句法变化，或者目标语话语组织方式上的变化（秦洪武、王克非 2009；秦洪武 2010）。本文以推理类话语标记为例，从编码复制角度探讨翻译和汉语话语标记演化之间的关系。

4. 研究方法

4.1　类比语料数据

五四运动以来，推理标记的使用伴随着该形式在翻译语言中的使用而发生变化，那么，我们就可以通过类比分析的方法，比较汉语原创文本和汉语翻译文本中的话语标记成分，找出那些受翻译语言影响的话语标记使用现象。为此，本研究使用了现代汉语历时类比语料库，其构成如下：

表1.　现代汉语历时类比语料库（使用AntConc统计）

类比语料库	子库名称	库容（词）	语料构成
汉语原创语料库	1911年前汉语原创子库	1,229,168	文学：1,136,056（92%）；新闻报刊：93112（8%）。晚清白话文运动：主体为白话小说，部分白话报刊，无科技文献。
	1919-1930年代汉语原创子库	1,236,273	文学：1,029,263（83%）；非文学：207,010（17%）。白话文用于非文学明显少于文学，文学文体丰富。
	1990年代-当代汉语原创（LCMC）	834,007	改革开放后：语料构成平衡。
汉语翻译语料库	1919-1930年代英汉翻译子库	2,713,469	文学：1,193,695（44%）；非文学：1,519,774（56%）。语料构成较为平衡。
	1980年代-英汉翻译子库	2,123,097	文学：1,006,694（47%）；非文学：1,116,403（53%）。文学、非文学语料构成较为平衡。
类比库库容合计		8,136,014	

表1所列的汉语翻译和汉语原创类比语料都是历时的，时间横跨一个世纪；每个阶段的间隔至少为20年，即一个语言代际。另外，为反映话语标记使用的实际特点，本项研究使用的语料包含非文学文本，文本类型多样。

4.2　数据提取

在这些语料库里，我们发现话语标记一般出现在句子起始位置，或位于句中某个小句的开头。这一使用特征为我们提取数据提供了方便，我们据此使用定位检

索，只提取由1-4词构成且独立使用的句首成分（检索时附加标点‘,’），即使用正则表达式提取语块[2]，然后进行人工识别和分类，最终筛选出推理类话语标记。

基于提取的数据，我们先对比白话文初期和当代汉语，找到当代汉语不同于早期白话的语言特征，接下来观察这些特征在各个时段的表现，以此考察翻译在这一历时变化中所起的作用。

5. 发现和讨论

5.1 汉语推理类话语标记使用频率的历时变化

为便于观察，我们先将数据标准化，按10万词为单位计算各标记的使用频率，标准化的数据见表2。

表2. 句首话语标记使用频率的历时变化（按句首独立的1–4词检索）

句首话语标记	汉语原创			汉语翻译	
	1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）	1919-1930年代英汉翻译	1980年后英汉翻译
推理标记标准化频率（十万词）	27.9	29.3	114	77.6	67.5
汉语句首话语标记总频率（十万词）	65.55	84.6	322.72	190.51	255.18

表2显示，在20世纪，汉语句首话语标记的使用频率逐渐增高，汉语翻译文本中相应的句首标记语也呈现相同的走势，只是1930年代以后的汉语翻译文本中的推理标记使用频率略有下降，这主要是受30年代以后“文白论战”中“白话文反思运动”[3]的影响。另外，1980年后的汉语原创中的推理类标记在频率上远高于同时期的汉语翻译文本，这与文本的性质有关。1980年后英汉翻译子库中多为文学文本，而汉语推理类话语标记多用于论证性的非文学文本。

由于各个子库之间库容差异很大，绝对频次不能直接表现话语标记使用频率在各个时期的差异，为此，我们使用似然率分析观察话语标记使用的历时差异的显著性。见表3、表4。

表3.　五四前后白话文中推理类话语标记使用差异的似然率分析

	1911年前汉语原创（频次）	1919-1930年代汉语原创（频次）	对数似然率	显著性
推理类标记	343	337	24.48	0.000

表4.　旧白话和当代汉语中推理类话语标记使用差异的似然率分析

	1911年前汉语原创（频次）	1990年代-当代汉语原创（LCMC）（频次）	对数似然率	显著性
推理类标记	343	950	44.54	0.000

与1911年前的白话文相比，现代汉语中推理标记的历时使用频率逐步增高，这一历时变化和汉语翻译语言中话语标记的变化趋向基本重合，这提示我们需要通过类比来分析汉语翻译语言对汉语发展的影响。

5.2　汉语原创推理类话语标记的历时变化

一般来说，汉语话语标记的变化有三种情况：一是话语标记的使用频率发生变化；二是出现功能相对稳定和独立的话语标记；三是出现全新的话语标记。这三种情形是否适用于推理标记尚不可知。我们就以1911前汉语原创子库为参照，观察这类标记的使用特征。

表5.　汉语原创推理类话语标记的历时变化

句首推理类话语标记	1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）
类符	34	32	56
形符	343	337	950

从表5和附录中看出，汉语原创推理类话语标记的使用总体看来变化明显，类符和形符均出现明显变化。这种变化在九十年代以后尤为显著。值得注意的是，1919-1930年代汉语原创文本中推理类话语标记的种类和频率有些微减少，这主要是受到了“白话文反思运动”的影响。另外，五四以来受欧化影响比较重的是汉语的书面语而非口语，欧化文体的使用有语体限制，一般都限于书面语，较少进

入口语。而本研究所用1930年代以后的汉语原创语料开始出现一定比例的非文学文本，口语体比例相对减小，受欧化影响的推理标记的使用就相应减少。

表6. 推理类话语标记的历时变化（详见附录）

1911年前汉语原创		1919-1930年代汉语原创		1990年代-当代汉语原创（LCMC）	
总（而言）之	27	总（而言）之	56	总的讲/总的说/总起来看/总前所述/综上所述/	46
因此	29	因此/因之	32	因此/因而/因为	295
所以	11	所以	12	所以	77
于是	1	于是	13	于是	106

表7. 推理类句首话语标记的历时变化（例示，详见附录）

1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）
就这样/要能这样/由这样看来/这样、这等……、这么说、这一来	既然这样/若不是这样/照这样看来/照这样做/这样……/这样一来、看这情形、从这里看来	就这样/再这样下去/这样（……）、这么说、这还不说、这是因为、这一来、如此这般
如此/若果如此/事已如此/虽然如此/原来如此、为此、因此、由此看来、据此	既然如此/如此（……）来、继此、如是、因此	如此这般/虽然如此/长期如此、故此、如是、为此、因此、由此看来
总（而言）之	总（而言）之、笼统言之、	综上所述、总的讲、总的说、总起来看、总前所述、总之

表6、表7和附录中的例证说明，推理类话语标记在1919年以后才大量出现，这类标记在1919年以前要么不用或者罕用，要么不能作为独立的句首成分使用。这些话语标记在1919年后经历了从出现到频繁使用的变化过程。另外，推理类话语标记在历时变化中也新出现了很多变式。因此，我们可以将推理类话语标记的历时变化特点总结为三方面：1. 形式更加丰富；2. 组合方式更加多样；3. 使用频率变高。这些剧烈的变化显然不能简单归结为语言自身演化的结果，应该与外部因素如翻译有关。

5.3　翻译与推理标记的语用化进程

表8显示，现代汉语（总体）中推理标记的使用特征和汉语翻译语言基本重合，汉语翻译语言中出现的推理类话语标记在源语英语中有对应的表达形式（如表9所示），且其句法位置和语用功能相同，说明这一变化包含代码复制过程，我们可以据此假定这和翻译有关。

表8.　推理标记在汉语原创和翻译子库中的使用

	汉语原创	1919-1930年代翻译	当代翻译
推理标记	这样（……）、总的……、一句话、显然（……）、简单地说、……归纳、很……	这样（……）、显然（……）、很……	这样（……）、总的……、显然、很……

表9.　推理标记的模型码英语对应词

汉语中的标记	主要英语对应词	其他英语对应词
显然	obvious(ly)、evident(ly)	clear(ly)、apparent(ly)
这样	so、thus、so that	so that

代码复制到复制品成为话语标记可能需要经历一个从概念意义到程序意义的语用化过程。但需要指出的是，代码复制本身无法复制源语的语用化过程，但能加快这一过程。比如，“显然”本来有实在意义，指显著、显扬、显赫，一般用于描述事物容易观察和理解的性质，如：“事迹显然，无可臻惑”；一般不会作用于一个命题信息，也就不会有置于句首充当话语标记的用法。语料检索显示，1911年前的汉语原创和CCL语料库古代部分里就只有词汇意义，没有语用标记这类用法。如：

（1）a. 省疏，并见周氏遗迹，真言显然，符验前诰，二三明白……（概念意义）

b. 就终不回，私与恭疏曰：“大人率厉炖煌，忠义显然，岂以就在困危之中而替之哉？”（概念意义）

但在1919-1930年代汉语原创中已有12例开始出现程序意义，作用于命题内容。如：

（2）a. 爱情很快被销蚀了——这显然不是使那产业车轮运转着的原动力。（概念意义+程序意义）

b. 两个肩头很有力，显然是做惯了苦工的缘故。（概念意义+程序意义）

可以看出，“显然”的概念意义已经减弱，开始用于标记说话者对所述信息的逻辑关系所作的推断，具备了独立用作语用标记的可能。

然而，同时期该话语标记语在翻译语言中也大量使用。在其对应的模型码（英语）中，表达式obvious(ly)、evident(ly)、clear(ly)、apparent(ly)等的典型位置也是句首且独立使用。翻译这些表达形式时，“显然”的语用潜力就得以充分挖掘，成为可独立使用的语用标记。如：

（3）a. *Obviously*, individuals who share this judgment will regard the legal mechanisms under discussion here as incomprehensible at best and perhaps perverse at worst.

显然，持有这种看法的人至多会认为这种法律机制不可思议，甚至会认为它是邪恶的。

b. *Clearly* the Secretary could not contract away his statutory authority.

显然，内政部长不能因为订立了合同就不再行使其法律上的职权。

上面的例证说明，启用目标语中现成的表达形式并不是简单的调用，而是一种改造。“显然”本有的词义改变了，由评价性表达（一般放在被评价成分后边）转变为话语标记（置于句首位置）。也正是从1919年后，“显然”的功能开始变化：它不仅能引出要陈述的内容，还可以独立使用，用于标记后面的推断性命题。鉴于我们很难找到其他可能的原因，现代汉语中句首推理标记的高频使用应是受翻译语言影响的结果，也就是间接受模型码英语影响的结果。汉语翻译语言挖掘了汉语语言中既有表达式的表达潜力，同时又加速了该表达式的语用化进程。表10的统计也可以说明这一点。

表10. “显然”标记的使用与翻译的互动

	1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）	1919-1930年代英汉翻译	1980年代-英汉翻译
句首	0	12	57	54	94
句首/独立	0	1	23	18	31

5.4 翻译与现代汉语的互动

从上面的描述和分析中可以看到，汉语句首话语标记的使用频率在短时间内明显增高，翻译在外部起到了重要的推动作用。但翻译语言以什么方式参与了这一历时变化过程呢？本文认为可以从以下两个方面观察。

5.4.1 翻译丰富了现代汉语话语标记的表达形式

比如，表示“总结”意义的话语标记在1919年前（CCL）白话文里只有“总（而言）之”；在1919-1930年代的白话文中，还有“笼统言之”；而在现代汉语

中它的组合方式丰富、灵活，可以检索到更多高频使用的变式，如“综上所述”、“总的讲/说”、“总起来看”、“总前所述”等，这些组合方式大都可以在汉语翻译文本中找到，其中的相互影响是显而易见的。

5.4.2　现代汉语有选择性地接受翻译语言的影响

任何语言的变化都是为了维系既有信息交流系统顺畅运行而进行局部修补或者优化，不会允许整个系统的改换。所以，汉语接受由翻译推动的语言变化时，会按自身的特点有选择地吸收或接受来自翻译语言的影响。就推理标记来说，在翻译时，基本码汉语复制模型码英语对应成分的组合方式及其语篇和句法位置。但这种复制不是将语法特征全部复制（4a中的关系代词that就无法复制），而是保留其语篇位置，但使用汉语的组合方式，如：

（4）a. It is *obvious* that this development is one which could not have taken place, had not circumstances favored the development of a caste of priests.

那是显然的，这个发展，如果情形是不适于一个僧侣阶级的发展，便不会发生出来。

b. They’re the first thing to disappear from bathrooms, *apparently*.

很显然，它们是最先从浴室里消失的东西。

这说明，代码复制受基本码本身形态句法性质的制约，并非单纯的形式拷贝。翻译语言本身也没有力量改变目标语，如果汉语中有翻译语言的某些特征，那只是汉语按照自己的需要接受了翻译语言的影响。因此，翻译语言和目标语之间是互动关系，这种互动既丰富了双语转换中对等成分的使用，又促进了目标语自身的发展。

6. 结语

翻译语言的某些特征以异乎寻常的方式进入汉语和时代有关，20世纪初的白话文运动有意识地推动翻译语言中多少带有“异味”的语言表达方式进入汉语，其中就包含话语标记语。汉语翻译语言引入话语标记语的主要方式是选择性语法复制，这种复制产生高度同构的结构，使得代码在双语间互译性更强，同时也丰富了汉语篇章组织的语言手段，总体上起到了积极的作用。

从现代汉语推理标记使用的历时变化上看，翻译中的语法复制实际上是充分挖掘汉语既有的言语表达资源，或是使用现成但不常用的语言表达形式，或是加速某些表达形式的语用化进程。这说明，汉语接受外来语言的影响有一个前提：一切改变均是为了维系汉语既有语言系统运行顺畅。这意味着，汉语翻译语言也受到汉语本身的约束，与之相关的语法复制只能是局部的，不可能涉及语言系统的改变。

注释

1. 根据Kubler（1985：25），在20世纪初大量译入中国的作品中，译自英国和美国的作品占到全部翻译的62%，来自英语的最多，也最具影响。

2. 如下面的正则表达式：‘<s>(\b[^\x00-\xff]+/[a-z]+\b\s){1,4},’（检索实例：‘<s> “/w 依/v 我/r 想/v ,’）；‘<s>(\b[^\x00-\xff]+/[a-z]+\b\s){1,4},’（检索实例：‘<s>很/d 明显/a ，/w’）

3. 白话文反思运动：1930年代以后，林纾、章士钊、陈寅恪、钱穆等学者对“白话文运动”提出了反对和质疑。继而出现了白话文反思阶段，反对过度欧化。

参考文献

Baumgarten, N. & D. Özçetin. 2008. Linguistic variation through language contact in translation [A]. In P. Siemund & N. Kintana (eds.). *Language Contact and Contact Languages* [C]. Amsterdam: John Benjamins. 293-316.

Fraser, B. 1998. Contrastive discourse markers in English [A]. In A. Jucker & Y. Ziv (eds.). *Discourse Markers: Descriptions and Theory* [C]. Amsterdam: John Benjamins. 301-326.

Fraser, B. 1999. What are discourse markers? [J]. *Journal of Pragmatics* 31(7): 931-952.

Johanson, L. 2008. Remodeling grammar: Copying, conventionalization, grammaticalization [A]. In P. Siemund & N. Kintana (eds.). *Language Contact and Contact Languages* [C]. Amsterdam: John Benjamins. 61-81.

Kubler, C. 1985. *The Development of Mandarin in Taiwan: A Case Study of Language Contact* [M]. Taipei: Student Publishing.

Urgelles-Coll, M. 2010. *The Syntax and Semantics of Discourse Markers: Continuum Studies in Theoretical Linguistics* [M]. London: Continuum.

董元兴、赵秋荣，2012，编码复制框架视角下翻译对现代汉语发展变化的影响——以被动语态为例 [J],《中国地质大学学报（社会科学版）》(3)：129-133。

傅斯年，1918，怎样做白话文 [J],《新潮》(2)：171-184。

贺　阳，2008,《现代汉语欧化语法现象研究》[M]。北京：商务印书馆。

李秀明，2011,《汉语元话语标记语研究》[M]。北京：中国社会科学出版社。

秦洪武、王克非，2009，基于对应语料库的英译汉语言特征分析 [J],《外语教学与研究》(2)：131-136。

秦洪武、王　玉，2014，从详述类话语标记看翻译与现代汉语话语组织的发展 [J],《外语教学与研究》(4)：521-530。

秦洪武,2010，英译汉翻译语言的结构容量：基于多译本语料库的研究 [J],《外国语》(4)：73-80。

冉永平，2003，话语标记语well的语用功能 [J],《外国语》(3)：58-64。

谢世坚，2009，话语标记语研究综述 [J],《山东外语教学》(5)：15-21。

赵秋荣、王克非，2013，英译汉翻译语言的阶段性特点——基于历时类比语料库的考察 [J],《中国翻译》(3)：15-19。

朱一凡,2011,《翻译与现代汉语的变迁（1905-1936）》[M]。北京：外语教学与研究出版社。

附录：

推理类话语标记使用的历时变化

1911年前汉语原创		1919-1930年代汉语原创		LCMC当代汉语原创	
既……	97	不管……	2	不管……	15
就这样	2	不论如何	2	而且	30
据此……	6	不然(……)	32	概而言之	1
那么说	3	从这里看来	2	故此	1
那么着	6	大概言之	2	归根到底	1
如此……	15	既然 如此/这样	5	归纳……	3
如果……	2	继此	1	果然	8
如今看来	2	继上说来	1	很……	9
如若……	7	简单的一句	1	既然……	4
若	9	简单地说	1	假如	4
若果如此	4	结果	4	结果	10
若是不然	17	看这情形	1	究其原因	1
事已如此	1	笼统言之	1	就这样	20
虽然如此	3	那并不	1	举例来说	1
所以	11	那末	44	看来/得出	17
倘/倘若……	13	那么	84	看样子	1
为此	2	如此（……）来	5	可见	15
无论如何	16	……不然	3	明显得很	1
要能这样	1	若不是这样	1	那么	66
要不然	5	如是	3	那末	5
因此	29	所以	12	然则	1
由现在/这样看起来	4	无论……	19	如是	7
由此看来	1	以上……	4	如此这般	2
由此可知	2	因之	1	如果(……)	12
于是	1	因此	31	……表明	11
原来如此	7	应当说	1	说到底	1
再怎么说	1	于是	13	虽/虽然（如此）	2
照……	12	照这样看来	1	所以	77

（待续）

（续表）

1911年前汉语原创		1919-1930年代汉语原创		LCMC当代汉语原创	
这等……	15	照这样做	1	推而广之	1
这么说	3	这样(……)	47	为此	45
这样	4	这样一来	5	无论怎样说	1
这样……	24	总(而言)之	52	显然	20
这一来	2			一句话	6
总(而言)之	27			因此	266
				因而	15
				因为	14
				由此	7
				由此看来	17
				于是	106
				再进一步	1
				再这样下去	1
				长期如此	1
				这还不说	1
				这是因为	6
				这么说	4
				这一来	1
				针对……	21
				正因为……	9
				正是我国春秋时代	1
				综上所述	4
				总的讲	1
				总的说	1
				总起来看	1
				总前所述	1
				总之	38
				这样(……)	61

通讯地址：273165　山东省曲阜市曲阜师范大学外国语学院

语料库驱动的机器词典构建关键问题探讨

解放军外国语学院 曹 蓉
浙江工商大学 濮建忠
解放军外国语学院 黄金柱

提要：语料库驱动的语言研究试图最大限度地摆脱已有理论束缚，力求发现新的反映语言使用本质的事实。在这一理念指引下，我们重新审视和探讨了机器词典构建的几大关键问题（如：语言描述的核心、基本单位、释义模式等），并在此基础上提出一种新的机器词典构建理念，即：以意义为描述核心、以语言使用（或文本）为获取意义的本源、以集词汇、语法、语义、语用于一体的“扩展意义单位”为基本描述单位，采用列举与归纳相结合的释义架构和正则表达式的表示方法。

关键词：机器词典、语料库、扩展意义单位

语料库自诞生之初即与词典构建结下不解之缘。早在1898年，德国学者Kaeding就通过统计单词在大量文本语料中的出现频率编写了《德语频率词典》，这被认为是最早的语料库及语料库在词典编纂上的应用。尽管Kaeding所使用的语料并非机器可读，然而这种从大量真实文本语料出发构建词典的理念可以说是开创性的。1959年，英国伦敦大学的Randolph Quirk发起了“英语用法调查”（SEU）语料库项目，并利用该语料库编写出版了《现代英语语法》。1980年，由英国伯明翰大学的John Sinclair负责的“柯林斯伯明翰大学国际语言资料库”（COBUILD）项目启动，并在此基础上先后出版了多部词典、语法书和用法指南等。2008年，欧洲辞书学会创始人Sue Atkins和国际著名的语料库词典学家Michael Rundell合作撰写了《牛津应用词典学指南》，对基于语料库的单语和双语词典编纂过程进行了详细介绍。在我国，诸如《常用汉字登记表》、《普通话三千常用词表》、《现代汉字综合使用频度表》、《现代汉语用字频度表》等在语料统计基础上构建的词典先后出现，为中文信息处理相关标准的建立提供了科学的基础数据。

可以说，语料库的出现给词典编纂注入了新思路，其在词典构建中的应用也日益受到重视，然而上述应用终归只是将语料库作为词典编纂的一种研究方法，甚至只是一种数据生成的手段或操作工具，而真正由语料库驱动所得的一些重要