Canadian TCSL Association
加 拿 大 中 文 教 学 学 会

The 6th Canada-China TCSL Conference
Collection of Papers

1.

基于附码语料库的对外汉语教学知识挖掘研究
Study on Knowledge Mining for TCFL Based on Annotated Corpus

盛玉麒**教授**，山东大学中文信息研究所

提要：本文运用语料库语言学的理论和方法，论述了汉语的特点和对外汉语教学面临的知识短缺以及基于附码语料库进行汉语知识挖掘的可行性，重点介绍了本文所采用的语料库加工后的主要属性数据库的类型和特点，从词语动态频度、兼类词分布、词语搭配及句法模型等方面，介绍了对外汉语教学知识挖掘的类型和具体方法。
关键词：语料库 对外汉语教学 知识挖掘

Abstract: This paper will inquire into the characteristics of Chinese language, the knowledge shortage encountered in teaching Chinese as a foreign language and the feasibility for knowledge mining based on annotated corpus. It lays emphases on introduction to styles and characteristics of attributive database adopted after knowledge processing. It also offers an introduction to styles and methods for knowledge mining in teaching Chinese as a foreign language in terms of word dynamic frequency, homonym distribution, word collocation as well as syntactic models.

Key Words: Corpus; Teaching Chinese as Foreign Language; Knowledge Mining.

一、导言

对外汉语教学面临诸多知识短缺的问题，一方面是因为汉语本身的复杂性和汉字记录汉语过程中的噪音干扰与信息缺失所致，另一方面是因为长期以来汉语的研究、工具书和教材的编写几乎都是面向母语学习者的，即使是名之为对外汉语教学的，也多以低年级母语学习者的读本为参照，所以，教和学都陷入的境地。

权威的汉语教科书讲解词类划分标准的时候，都以会说汉语为前提，例如：“形容词一般能受程度副词修饰”、“动词后面可以带动态助词”、“及物动词能带宾语”、“不能重叠”……等等。试想，对于外国学习者来说，根本不懂汉语，自然

不知道词与词之间能否搭配。所以，这种讲解对他们来说毫无用处。

正如维特格斯特说的那样，“用法即意义”。汉语词语究竟有多少种用法谁也说不清楚，但是，可以肯定地说远远超出静态的词典工具书所描述的范围，从发展的观点看，更是如此。因此，基于语料库的汉语知识挖掘就显得十分必要。特别是对外汉语教学领域，要解决所遇到的知识短缺的问题，唯有从现代汉语流通语料库中挖掘，舍此没有其他捷径可走。

这里说的知识挖掘并不是增加对外汉语教学的内容和知识量，而是提高教学效果，把最简明、最实用、最有规律的知识找出来教给学生。

二、汉语知识短缺与对策

1．词汇语义知识短缺

汉字数量繁多、结构复杂、理据多样、信息冗余、读写繁难；同音、谐音、连读音变影响词汇语义的理解和运用；有声语言中的语调、节奏、轻重音等都因为汉字无法记录而被忽略。常常会遇到“听—说”的时候没有歧义，用汉字写出来再念的时候，就有了歧义，如“喝口水”绝不会听成“喝/口水”，却很可能被留学生们读错。

字义笼统，随词而变；词义模糊，随句再变；二语习得者茫然难得要领。同一个字用在不同的语句中，意义改变很大。有些表面上看起来词形类似，却不能望文生义、随意类推，如：

“妈妈═妈”、“爸爸═爸”、“哥哥═哥”；

“爷爷≠爷”、“老爷爷≠爷爷”、“老公公≠老公”。

“有没有关系——有关系——没关系——搞关系——搞好关系”等短语中的“关系”的意义各有细微的差异：

◆“没关系”起码有三个用法：

①惯用语，礼貌用语。如：“打搅你们吃饭了，对不起。”“没关系，别介意。”

②惯用语，不要紧，没有问题：没关系，我自己会做。

③短语，无关、没有关系。如：我可不愿意搬弄是非去管那些跟我没关系的闲事。

◆“有关系”：

①短语，存在某种关联。如：这消息与你的家乡有关系。

②短语，在疑问句中表示有问题。如：这有关系吗？

③短语，表示特殊关系。如：他与领导有关系

◆“搞关系”

短语，贬义，指为某种利益驱动进行的疏通、贿赂行为。如：他很会搞关系。

类似的例子比比皆是，举不胜举。

2. 句法知识短缺

虚词和语序是汉语表示语法功能的重要手段。虚词和语序自然是对外汉语教学的重点。

虚词是封闭的类，数量有限，但是由于“古今参杂”，必须加以筛选。筛选的依据就是实用性，这就离不开通过语料库的统计分析。

语序知识表现在彼此的相关性上，或者说组合搭配关系上。汉语词与词之间的组合搭配关系并不像有形态的语言那样，掌握了词类和形态标志就条理清楚了。汉语曾经有过“词无定类”、“依句辨品、离句无品（“品”即词性。麒按)”状态。长期以来，词都是“自由自在”地“兼类”、“活用”着，最权威的《汉代汉语词典》标注词性也是从第五版开始的。离合词问题、动宾结构带宾语问题等，成了教学中“剪不断、理还乱”的难题，由于对这类现象缺乏定量定性分析，常常是就事论事，缺乏规律性解释。

例如“在桌子上写字”、“在火车上写字”、“在飞机上写字”中“上”的意义不一定是指“上面”。“火车上”有“在火车里面”和“在火车外面”两种解：而“飞机上”则只能理解为“飞机里面”。

至于“刚才——刚刚”、“突然——忽然”、“帮助——帮忙”的区别，用“解词”的方法不如用不同的搭配举例说明来得简洁明了。

静态系统的“兼类”并没有指出所兼的“类”之间的主次。语言学习不是照着词典学、只要背会词典就能说话。实际使用的句法知识才是语言学习的主要对象。因此，采用定量定性统计分析数据描述的句法知识，不但是中文信息处理智能化的需要，也是提高对外汉语教学科学性和实践效果的需要。

3. 语用知识短缺

汉语共时系统中口语、书面语、网语多元杂糅，非标准普通话随处可闻，文言词语偶尔参杂期间，甚至有些冷僻词语一夜之间成为“流行语”。词语在实际使用中的意义和用法与词典中的解释往往不尽相同。如果拘泥于词典的释义，在实际交际中就会遇到费解的情况。

例如接电话时说“好，好，我就来。”其中的“来”所表示的是“去”的意思。这就不是词典中解释的“从远处到近处”的意思。

日常交际中常使用省略、简称，有大量不完全句；表达含蓄、不直白，这与文化习俗方面讲究“礼节”和形式的传统习惯有关。许多常用的模糊词语如“还行”、“还可以”、“不错”、“很有意思”、“很有特点”、“好吧”、“看看再说”、“研究研究”之类，需要在具体的语境中仔细玩味才能体会其中所表达的意义。

长期靠天吃饭的农业社会形态和粗放式生活方式也会影响到话语表达方式，特别在日常词语中，如称谓语、礼貌语、情感表达等表现明显。崇尚自由、顺应自然的审美取向，注重心领神会、淡化言语表述，喜欢隐喻的方式，对微言大义津津乐道，加上历史悠久、文献浩瀚，典故丰富、崇古尚文，致使文言词语、诗词名句，超常搭配时有所见。这些都无形中增加了对外汉语教学的难度。

由于多元信仰和多源禁忌，传统文化对生老病死、婚丧嫁娶、吃喝拉撒、交易旅行等形成了多种多样的委婉语。这些文化层面的知识在对外汉语教学中几乎是一个禁区，成了“说不好、不好说、不说好”的教学“瓶颈”。实际上，如果处理得好，

文化词语不但可以丰富教学内容，还可以提高学习汉语的积极性，因此，挖掘常见习用的文化色彩词语知识也应是对外汉语教学领域的一个新课题。

三、基于语料库知识挖掘的可行性

语言教学的目标是“听说读写”，必须立足于活的、实用的语言，因此，从大规模真实文本语料库中挖掘静态系统所缺乏的知识，应该是对外汉语教学领域一项重要的基础工程。

1. 从举例证明到定量分析

语言学是一门最具人文社会性的实证科学，“约定俗成”的规则实际上就是统计学的“大数定律”。因此，不能满足于“例不十、法不立”的传统标准，而应注重定量定性分析的方法。

2. 从静态系统到动态系统

维特格斯特曾说“用法即意义”，只有在动态系统中才能真正发现意义。由自然语言的真实文本所组成的语料库是语言动态系统“子集”，是内部语言的外化。各种用法、各种意义表达上的细微差别都是静态系统所无法比拟的。

3. 从充分描写到充分预测

当代语言学对语言现象和语言规律的研究讲求的是“充分描写、充分解释和充分预测”。这里所说的“充分”实际上是一个理想的目标，任何时候只能是相对的“充分”。采用基于语料库的知识挖掘方法，借助计算机大容量、高速度的优势，可以在所建立的语料库范围内，实现充分和穷尽式的描写和分析，这就已经远远超越了以往任何个人研究能力和时间周期的局限性。

“预测”是从已知推测未知的复杂的探索过程。许多语言现象所蕴含的规则或规律，需要满足统计学中的“大数定律”，也就是从“量变”到“质变”的规律。那种“一叶知秋”的能力，“见瓶水之冰而知天下之寒”的预测力，需要大量经验知识的积累和复杂的逻辑推理。现在我们所做的只是一种探索和尝试，尽我们所能挖掘

所缺乏的语言知识。虽然我们不知道距离理想的目标究竟有多远，但是我们相信经过不懈的努力总会逐步逼近这个目标。

4. 克服语料库的局限性

语料库所收入的语料总有一个局限，无法收入所有已经说出的言语作品，更无法收入那些“能说”但没有说出来的句子。从这个意义上可以说，语料库永远都是“不完备”的。

实际上任何研究都会受到研究者认知能力和范围的限制，即使内省式的研究可能从内部词库中搜索出语料库之外的例子，但是也还是有局限性。在语言知识挖掘方面，特别需要把“内省式”的研究和语料库语言学方法结合起来，采用科学合理的抽样方法，尽量保证语料库的规范性和代表性，再充分发挥研究者内部词语知识库和“见微知着”的能力，实现基于语料库的汉语知识挖掘的根本突破。

四、本文所用语料库

1．本研究所用语料库概况

目前语料库语言学受到学界的普遍重视，不同规模、不同用途的语料库纷纷建设，并投入使用。加之网络的普及和数字化文本的与日俱增，给语料库的研究和建设带来了极大的便利。

本文研究所采用的自建语料库主要以现代汉语文学作品抽样语料库。总字符数将近600万字，其中汉字符号近580万字。（详见表1）

表 4-1：文学语料库的基本数据

词长	词次	词频	字数
单音词	8575	2985054	2985054
双音词	49614	1214121	2428242
三音词	11278	72846	218538
四音词	7702	34222	136888
五音词	631	1718	8590

六音词	178	287	1722
七音词	81	136	952
合计	78059	4308384	5779986

本语料库采用国内流行的中科院计算所研制的自动分词和标注词性软件进行加工处理。因为文学文本的特殊性，分词正确率达到 90%以上，词性标注正确率超过 80%。因此需要人工校对。

加工后的语料库分别建立不同的数据库，用于知识挖掘和数据分析。这些数据库主要有：

①词频数据库（音序、降频）

②属性数据库（词长、词性、频数、频率）

③兼类词数据库；

④二词搭配数据库；

⑤三词搭配数据库；

⑥四词搭配数据库；

⑦五词搭配数据库；

⑧句法模式数据库。

2. 搭配关系数据库

1）二词搭配数据库

由两个词相互搭配所组合成的一种关系。通过二词搭配数据库，不仅可以得到词与词之间的组合搭配关系，还可以得到有关结构模式的量化信息。从降频表中可以看到高频组合的出现频度和前后两个词在搭配选择上的分布情况。

表 4-2：二词搭配降频例表

序号	词 1	词 2	使用频数	使用频度%
01	不/d	是/v	6424	0.14910
02	他/r	的/u	6391	0.14833
03	我/r	的/u	5193	0.12053
04	她/r	的/u	4441	0.10307
05	也/d	不/d	4317	0.10019
06	自己/r	的/u	2817	0.06538
07	这/r	是/v	2792	0.06480
08	就/d	是/v	2705	0.06278
09	不/d	能/v	2650	0.06150
10	我/r	不/d	2491	0.05781

2）三词搭配数据库

三词搭配“是不是”的“v+d+v”模式就概括了所有的动词“正反重叠”式，如“会不会”、“去不去”、“能不能”等等；

表 4-3：三词搭配降频例表

序号	词 1	词 2	词 3	使用频数	使用频度%
01	是/v	不/d	是/v	861	0.03951
02	两/m	个/q	人/n	563	0.02583
03	也/d	不/d	是/v	388	0.01780
04	对/p	我/r	说/v	382	0.01753
05	也/d	不/d	会/v	377	0.01730
06	是/v	一/m	种 /Question	352	0.01615
07	笑/v	着/u	说/v	352	0.01615
08	一/m	句/q	话/n	348	0.01597
09	这/r	件/q	事/n	339	0.01556
10	有/v	一/m	天/q	311	0.01427

3）四词搭配数据库

四词搭配的“躺在床上”的“v+p+n+f”模式可以生成“走在路上”、“写在信里”、“记在心里”、“体现在行动上”等等；

表 4-4：四词搭配降频例表

序号	词 1	词 2	词 3	词 4	使用频数	使用频度%
01	躺/v	在/p	床/n	上/f	119	0.0081
02	叫/v	了/u	一/m	声/q	116	0.0079
03	你/r	是/v	不/d	是/v	105	0.0071
04	看/v	了/u	一/m	眼/q	103	0.0070
05	说/v	不/d	出/v	话/n	72	0.0049
06	抬/v	起/v	头/n	来/f	68	0.0046
07	是/v	怎么/r	回/q	事/n	68	0.0046
08	摇/v	了/u	摇/v	头/n	57	0.0039
09	走/v	来/f	走/v	去/v	56	0.0038
10	出/v	了/u	什么/r	事/n	46	0.0031

4）五词搭配数据库

五词搭配的“叹了一口气”的“v+u+m+q+n”模式可以生成“说过一句话”、“读了三年书”、“听了一节课”等等。

表 4-5：五词搭配降频例表

序号	词 1	词 2	词 3	词 4	词 5	使用频数	使用频度%
01	叹/v	了/u	一/m	口/q	气/n	80	0.00560
02	说/v	不/d	出/v	话/n	来/f	50	0.00350
03	看/v	了/u	我/r	一/m	眼/q	48	0.00336
04	一/m	天/q	比/p	一/m	天/q	31	0.00217
05	一/m	动/v	也/d	不/d	动/v	24	0.00168
06	这/r	是/v	怎么/r	回/q	事/n	21	0.00147
07	喘/v	不/d	过/u	气/n	来/f	19	0.00133
08	你/r	说/v	是/v	不/d	是/v	17	0.00119
09	话/n	也/d	说/v	不/d	出来/v	17	0.00119
10	我/r	吓/v	了/u	一/m	跳/q	17	0.00119

五、可挖掘知识类型

1．动态频度知识

词语动态系统的使用频度是评价该词“活性”的重要功能指标，在研制教学用词表、设计教学大纲等方面都有重要的参考价值。降频表就是一个最常用的动态频度表。（见表 1）

表 5-1：降频词表前 20 条例样

序号	词条	词长	频次	频度
1	的/u	1	178944	4.1469
2	我/r	1	68510	1.5877
3	是/v	1	61417	1.4233
4	不/d	1	57292	1.3277
5	了/u	1	56362	1.3062
6	他/r	1	55585	1.2882
7	一/m	1	52310	1.2123
8	了/y	1	42064	0.9748
9	在/p	1	38961	0.9029
10	说/v	1	36861	0.8542
11	她/r	1	36442	0.8445
12	你/r	1	35647	0.8261
13	着/u	1	35357	0.8194
14	就/d	1	29013	0.6724
15	也/d	1	25040	0.5803
16	这/r	1	24793	0.5746
17	地/u	1	24068	0.5578
18	有/v	1	23718	0.5497
19	人/n	1	21857	0.5065
20	到/v	1	19699	0.4565

2．兼类分布知识

对于兼类词来说，从动态频度表中可以看出兼类词所兼的各类词之间在使用频

度上的差别，从而判断轻重主次，以便把最主要的功能类型作为教学的重点。可避免平均用力或主次不分，从而提高教学效果。（见表 5-2）

关于统计数据的评价，知识挖掘的角度与以往词频统计的信度判断指标不同。在词频统计中，一般只注重高频区的词语，认为低频区缺乏参考价值可以忽略。但是从知识挖掘和知识发现的角度，就要特别注意低频区的情况，因为，低频区除了属于误切分造成的“特殊”情况之外，很可能就是具有潜在价值的“关键少数”，因此，应给予特别的关注。

例如在“兼类词动态频度”统计中发现，“好”的连词用法只有 1 次。（表 5-2）如何分析评价这种只见到一次的用例，就是知识挖掘过程中常见的问题。不能轻易判为“误切分”、“误标注”，而要慎重对待。

通过在原库中认真查找，实际的用例是：

“要王七一将来再写书，**好**编给他听！”

这是一个合法的用例，这就启发我们扩大语料面，搜索更多的用例。根据这个思路，笔者随意从网上搜索一下，很容易发现了一些例子：

“我给槟榔拍一砖，**好**让他清醒……”

“你不爱他了,请放手,**好**让他知道;”

“可以和他提一下离婚的问题，**好**让他把事情说清楚。”

“我明儿也再生一个，**好**让他当老师并参加新课改”

“版主帮我看看我小孩的八字，要补什么 ，我**好**帮他其名字...”

（上述各例都保留了超链接指示出处，读者若直接检索字符串也可以从网上找到原文。玉麒注）

这就是前文所谈的“内省法”与语料库语言学方法相结合实现知识挖掘突破的一个有力证明。

表 5-2：兼类词动态降频表例样

兼类词	词类序	词类	频次	频度
把	1	介词	14188	0.3288
	2	量词	934	0.0216
	3	数词	54	0.0013
	4	动词	18	0.0004
和	1	连词	14154	0.3280
	2	介词	1788	0.0414
	3	人名	46	0.0011
	4	动词	26	0.0006
	5	副词	2	0.0000
来	1	动词	9071	0.2102
	2	趋向动词	7482	0.1734
	3	数词	185	0.0043
在	1	介词	38961	0.9029
	2	副词	3072	0.0712
	3	动词	758	0.0176
过	1	助词	9649	0.2236
	2	动词	1104	0.0256
	3	副词	328	0.0076
好	1	形容词	10167	0.2356
	2	副词	147	0.0034
	3	动词	11	0.0003
	4	连词	1	0.0000

3．句法搭配知识

组合与聚合是语言的基本关系。词与词之间的搭配就是组合关系的具体表现。不同的搭配关系往往反映不同的词汇语义功能，能够体现出维特根斯坦所说的“用法即意义”的思想。

表 5-3：名词“关系”的常见搭配表

词 1	词 2	结构关系	使用次数	使用频度
有	关系	动词+名词	41	0.0016
没有	关系	动词+名词	28	0.0011
是	关系	动词+名词	10	0.0004
没*	关系	动词+名词	9	0.0004
断绝	关系	动词+名词	9	0.0004
毫无	关系	动词+名词	7	0.0003
搞好	关系	动词+名词	4	0.0002
拉	关系	动词+名词	2	0.0001
托	关系	动词+名词	2	0.0001
找	关系	动词+名词	2	0.0001

（注：“没关系”作为相对稳定的惯用语在语料库中发现使用 134 次，频率为 0.0031%。）

表 5-4：动词“关系”的搭配：

词 1	词 2	结构关系	使用次数	使用频度
关系/v	到/v	动补	16	0.00037
关系/v	着/u	动助	7	0.00017
关系/v	大家/r	动宾	1	0.00002
关系/v	那/r	动宾	1	0.00002
关系/v	你/r	动宾	1	0.00002
关系/v	全局/n	动宾	1	0.00002
关系/v	人家/r	动宾	1	0.00002
关系/v	我国/n	动宾	1	0.00002

其中，“关系到”的搭配使用频数最高。从扩展的成分看，动补结构后接名词性成分。下面是从 4 级相关性数据库中检索出来的用例。

表 5-5：“关系+到”与后接成分表

动补关系		后接成分	
关系/v	到/v	爸/n	妈/n
关系/v	到/v	傅/nr	桂英/nr
关系/v	到/v	民生/nz	大计/n
关系/v	到/v	你/r	将/d

关系/v	到/v	全县/n	工业/n
关系/v	到/v	省会/n	城市/n
关系/v	到/v	世界/n	革命/vn
关系/v	到/v	他们/r	之间/f
关系/v	到/v	我国/n	革命/vn
关系/v	到/v	我们/r	党/n
关系/v	到/v	子孙后代/n	的/u

进一步找出语料库中“关系+到”后接成分扩展的用例，可以看出更多的搭配信息。下面是从 9 级相关数据库中检索出上表相关的后接成分扩展用例。

表 5-6 “关系+到”后接成分扩展例表

关系+到	后接成分
关系/v 到/v	爸/n 妈/n 如何/r 尽快/d 培养/v 你/r 成材/v
关系/v 到/v	傅/nr 桂英/nr 的/u 去留/vn
关系/v 到/v	民生/nz 大计/n 的/u 事/n 都/d 可以/v 马虎/an
关系/v 到/v	你/r 将/d 来/v 的/u 人生/n 前途/n
关系/v 到/v	全县/n 工业/n 生产/vn 上/f 一个/m 新/a 阶段/n
关系/v 到/v	省会/n 城市/n 的/u 面貌/n 问题/n
关系/v 到/v	世界/n 革命/vn 前途/n 的/u 大事/n
关系/v 到/v	他们/r 之间/f 感情/n 的/u 谈话/vn 中/f
关系/v 到/v	我国/n 革命/vn 前途/n 的/u 大事/n
关系/v 到/v	我们/r 党/n 和/c 国家/n 生死存亡/i 的/u 运动/vn
关系/v 到/v	子孙后代/n 的/u 前途/n 呢/y

六、余论

基于语料库的汉语知识挖掘是一个全新的课题，涉及到汉语本体、应用以及语料库建设与加工等诸多理论与实践问题。本论文只是一个初步的研究报告，作为引玉之砖以引起学界的关注。还有许多富有挑战性的问题，例如汉语句法标记知识、词缀与类词缀知识、歧义与消歧知识、语法化知识等等，期待更多的专家学者参与和参加到知识挖掘的系统工程中来。

欣闻第六届加拿大中国对外汉语教学学术研讨会在温哥华召开，谨以此文表示祝贺，并就教于大方之家。文内不当之处，敬希批评指正。

参考文献

1 盛玉麒（2006）.《语言文字信息处理》(济南:山东大学出版社)。

2 Ruth Kempson, Wilfried Meyer-Viol, Dov Gabbay (2001). *Dynamic Syntax: The Flow of Language Understanding*《动态句法学：语言理解的流程》(Oxford: Blackwell Publshers)。

3 （美）帕蒂（Partee,B H），（美）默伦（Meulen,A），（美）华尔（Wall,R E）(1993),*Mathematical Methods in Linguistics*《语言学中的数学方法》(Boston: Kluwer Academic),冯志伟导读, 世界图书出版公司，2009-3-1。

4 杨惠中 (2002).《语料库语言学导论》(上海:上海外语教育出版社。

5 周荐(2004-12).《汉语词汇结构论》(上海:上海辞书出版社)。

6 魏荣华,<论动词的句法搭配模式>,《中国俄语教学》,1997 年第 4 期。

7 范继淹等,<工智能和语言学>,《中国语文》1980 年第 4 期。

8 王素格等,<自动获取汉语词语搭配>,《中文信息学报》,2006 年第 20 卷第 6 期。

9 孙茂松等,<汉语中的兼类词、同形词类组及其处理策略>,《中文信息学报》,1989 年第 4 期。

10 罗振声等,<汉语句型自动分析和分布统计算法与策略研究>,《中文信息学报》,1994 年第 2 期。