现代蒙古语词语搭配分布特征初探

— 以单词 GAR、JALAGV 为例[1]

春岭　斯琴

内蒙古大学蒙古学学院　呼和浩特市　010021

E-mail: chunling830122@126.com，xisiqin@126.com

摘要：本文以单词 GAR、JALAGV为例，探讨现代蒙古语以高频形容词和名词为中心的词语搭配的分布情况，归纳出了一种根据共现频率和MI值获取词语间的搭配强度显著搭配词及最佳窗口。

关键词：语料库，　词语搭配，　统计法，，　MI 值

Discussion of Distributional Characteristics Of The Modern Mongolian Collocation

- Taking the word GAR and JALAGV as example

CHUN LING　　SI QIN

Inner Mongolian University Mongolian institute　Huhhot　010021

E-mail: chunling830122@126.com，xisiqin@126.com

Abstract: we take the word GAR and JALAGV as the example in this paper, discussing modern Mongolian which is based on the collocation distributed situation by taking highly frequency adjective and the noun as the central words. In order to gain intensity remarkable matching words and The best window, one kind of methods which is based on co-appearance frequency and the M I value is concluded.

Key words: derivation, collocation, Statistics, MI value

1　引言

词语搭配在自然语言处理的很多方面都起着重要作用,如语言教学，尤其语言生成和机器翻译等领域显得特别重要。对自然语言生成而言，从一定的逻辑模式生成文本需要大量词语组合知识,由于大多数的句子至少含有一个词语搭配,因此利用词语搭配能够显著提高文本生成的质量。对词义消歧而言，词义消歧就是在特定的语境下,为多义词选择正确的义项。从人理解自然语言的角度来看,通常只需要利用多义词上下文中的一两个词就能辨别出多义词的义项,因此词语搭配被认为是多义词消歧中的一个重要特征。自然语言信息处理工作给搭配研究带来了新的发展可能性①。用机器准确的切分，生成语言时需要把语言知识让机器理解，然而搭配研究就是机器翻

[1]作者简介：春岭（女）斯琴（女）内蒙古大学蒙古学学院硕士研究生 研究方向:蒙古文信息处理

译，语料库加工，语言生成系统等信息处理工作的基础。

蒙古文搭配研究起步较晚，从20世纪80年代起，义位搭配关系方面的一些研究成果陆续问世了。例如，田峰的《词义的组合问题》、斯琴，德乐格日玛的《语义学》(1996年)等。额尔敦朝鲁的《面向信息处理的蒙古语动词语义研究》，从信息处理的角度研究了动词和名词语义搭配。姜迎春的《面向信息处理的蒙古语形容词语义研究》，对蒙古语形容词语义类别、形容词语义类和名词语义类的搭配规律、形容词配价以及如何在“蒙古语语法信息词典”中设置形容词语义属性字段等问题进行了探讨。对蒙古文信息处理工作来说，建立词汇知识库是不可缺少的一环，而研究词语搭配正是建立词汇知识库的一个重要步骤。本文借鉴先人的研究成果，并以100万词级《现代蒙古语语料库》及其配套软件为基础，考察了以单词 “GAR”和“JALAGV” 为中心词的词语搭配分布情况，根据其共现频率和MI值，获取了它们的搭配强度显著搭配词及最佳窗口。

2 搭配词的确定

众所周知，从计算机处理的角度看，词语搭配通常被作满足一定规则的，重复出现的，具有正常的句法结构的词语组合。因此我们依据短语标注语料库确定了搭配词。

本文的研究借鉴传统蒙古语语法研究成果的基础上，对现代蒙古语兼类词与其他词类上及其在搭配上的变异进行了大规模的，系统的考察，分析，比较。因为蒙古语与其它语言一样，也有相当多的兼类词。100 万词级《现代蒙古语语料库》中兼类词有 115235 条，占总语料库的 11.4%，所以，处理兼类词问题占重要地位。由于兼类词的词类与该单词的具体环境有着密切的联系，完全实现自动标注难度较大。

本章使用《现代蒙古语语料库程序》来为关键词提供该词在 100 万词级《现代蒙古语语料库》中每次出现时的上下文环境。再此对给定的词考察左，右两边的搭配。蒙古语搭配数目极大，几乎所有的实词（名词，动词，形容词，时位词，数词等）都可以作为中心词，与其他词语组合在一起构成搭配。因此我们能力有限只选择若干有代表性的词来考察。本文从名词、形容词中各选一个频率最高，兼类的简单的词作为这两类词的代表进行考察。这两个词分别是“GAR”和“JALAGV”。我们对这些词的搭配对象的考察大体上如下步骤进行：（1）对于关键词在语料中的每个实例，抽取其左右各 4 或 5 个词（不记标点符号），构成子语料库。我们所用的语料是已经标注过的，但仍有一些兼类词的 100 万词级《现代蒙古语语料库》。（2）人工标注出子语料库中所有可以与关键词组成合法搭配的词。（3）抽出所有的搭配词，构成一个表（Excel），姑且称之为搭配词全集。（4）统计分析共现搭配词在每个位置（[-5, +5]或[-4, +4]）上的分布情况。本章运用统计搭配词与关键词的共现频率和测量共现词项间的 M I 值的方法来实现了词语间的搭配强度显著搭配词。该方法首先利用检索工具对关键词作带有语境的检索(KWIC)，然后提取一定跨距内与关键词共现的所有词项并统计这些共现词项的频数。我们，在语料库中与关键词共现频数达到 3 次（等于）以上的词项才可被认为是关键词的搭配词。以往研究结果表明，就英语而言，将跨距界定为[-4, +4]或[-5, +5]是较为合适的。计算 MI 值是通过比较搭配词的观察频数的差异来确定某一词语搭配在语料库中出现概率的显著程度(Hunston 2002: 70)② 。MI 值(Mutual Information Score，互信息值)表示的是互相共现的两个词中，一个词对另一个词的影响程度，或者说一个词在语料库中出现的频数所能提供的关于另一个词出现的概率信息。MI 值越大，说明关键词对其词汇环境影响越大，对其搭配词吸引力越强。因此，M I 值表示的是词语间的搭配

强度。基于语料库的词语搭配研究中通常把 MI 值等于或大于 3 的词作为显著搭配词。因此，我们也跨距界定为[-4, +4]或[-5, +5]，与关键词共现频数达到 3 次（等于）以上的词项，且 MI 值等于或大于 3 的词作为显著搭配词。MI 值的计算公式为：

$$MI(x,y) = \log_2 P(x,y)/P(x) \times P(y) = \log_2 F(x,y) \times N/F(x) \times F(y)$$

3 词语搭配的分布特征

搭配是一种具有任意性的，重复出现的词的组合。但它有三个要件：（1）必须是合法的词语序列，这是前提条件，我们用规则和人工校对来保证。（2）具有任意性（即强制性）。（3）是重复出现的组合③。我们抽取的搭配既包括连续组合，也包括非连续的和跨层次的组合。因为“GAR”、“JALAGV”作为兼类词，也是比较代表性的，较为常用的词。下面以名词“GAR”和形容词“JALAGV”来说明。我们首先利用统计的方法抽取“GAR”和“JALAGV”的 10 个共现数据，“JALAGV”在语料中共出现了 499 次，“GAR”在语料中共出现了 530 次。后用人工的方法统计测量，并标注兼类词“GAR”、“JALAGV”的词性其关系类型，在我们抽取的 100 万词级《现代蒙古语语料库》中人工识别出不同的搭配对“GAR”321 种，“JALAGV”365 种，这些搭配对共出现了 505 次和 1012 次。通过计算 MI 值得出 55 次和 68 次。下面是“GAR”“JALAGV”频率高的 10 个搭配词：

“GAR”（Nt）的分布特征及MI值统计测量数据：

共现搭配词	词性	语料中的总出现次数	共现频数	MI 值	搭配分布位置
HOL	Nt	328	22	7.00	L1（R1）
BARAGVN	Ne	408	14	6.03	L1（L2）
JEGUN	Ne	392	11	5.75	L1
PVV	Nt	110	10	7.44	R1
HOGOSON	Ac	171	11	7.05	L1
DURU/JU	Vt	22	8	9.44	R1
ALCIGVR	Nt	46	9	8.54	R1
HODEL/JU	Ve	133	6	6.43	R1
ERGIGUL/JU	Vt	52	6	4.46	R1
BARILCA/N	Ve	6	6	10.32	R1

“GAR”（Vt）的分布特征及 MI 值统计测量数据：

共现词搭配	词性	语料库中的总出现次数	共现频率	MI 值	搭配分布位置
TERGE/N-DU	Nt	60	3	6.58	L1
AB/V/GAD	Vt	635	3	3.18	L1
HVRDVN	Ac	214	2	4.16	L2

由此可以看出，通过考察[-5, +5]范围内的词，随着距离的增大，搭配词集元素增加的数目

越来越少。如果以 [-1, +1]作为观察窗口，在该窗口内，搭配词的种数和次数分别是55和295。

MI 值可清楚表现共现词间的相互吸引程度,它可以帮助我们确定把哪些词作为关键词的可能搭配词而重点加以研究。但是MI 值高的搭配词不一定和关键词共现的频数就高。以表中的BARILCA/N 一词为例。尽管BARILCA/N在语料库中频数较低(仅出现6 次) ,但从MI 值(10.32) 来看二者搭配显著,这是因BARILCA/N 在语料库中几乎都是与"GAR"结伴共现(5 次) 。这说明MI 值表示的词语连结信息并不总是可靠。如果一个词在语料库中的出现频率较低,而出现时又多与关键词共现,那么二者的MI 值肯定很高。这也说明对于语料库中的低频词(频率小于10) ,MI 值信度较低,因为我们不能确定这一结果是源于二者的真正关联还是源于语料库的特殊本质。

据考察词语搭配的内部关系, "GAR" 的各类型搭配分布表如下所示:

结构关系	定体关系	联合关系	宾述关系	辅助关系
出现次数	14	5	22	3

进一步分析可以看出，在L1、R1 位置上的搭配词基本上都是支配关键词的动词（构成宾述关系）和修饰关键词的词语（构成定体关系）。

"JALAGV" (Ac)的分布特征及MI值统计测量数据:

共现搭配词	词性	语料中的总出现次数	共现频数	MI 值	搭配分布位置
NASV/N-ACA-BAN	Nt	10	4	9.67	R2, R1
NASV	Nt	290	34	7.90	L1, R2
NASV-BAN	Nt	56	6	7.77	R1
NASV/N-DAGAN	Nt	62	5	7.36	R1
CIBAGANCA	Nt	38	18	9.91	R1
VR1	Ac	18	9	9.99	L1
BOSEGUI-YI	Nt	11	3	9.11	R1
GEYICIN	Nt	19	4	8.74	L1
IDER	Ac	32	5	8.31	L1, R2
EMEGTEI	Nt	213	14	7.06	R3, R1, R2

"JALAGV" (Nt) 与有关搭配词的最高 MI 值统计测量数据

共现搭配词	词性	语料中的总出现次数	共现频数	MI 值	搭配分布位置
SVYVL-TAI	Ne	11	3	9.11	L1, L2
GOYO	Ac	114	14	7.96	R2, R1, R4
GEYICIN	Nt	19	4	8.74	L1
ADVGVCIN	Nt	46	7	8.27	L1
YABV/G_A	Ve	221	7	6.01	L2, L1

SILIDEG	Ac	91	3	6.07	L1, R3
NASVTAN	Nt	57	3	6.74	R3, L1
HOGSIN	Nt	456	8	5.16	L1, L3, L4, R1, R2
SIR_A	Ac	329	6	5.21	L1, L4
ONO	Rj	235	5	5.43	L1, L3

根据“JALAGV”的使用特点，在跨距为[-4, +4]间提取，检索，计算排序，位置按出现频率顺序排列，统计结果中将其剔除后，得到与“JALAGV”共现频数大于（等于）3 的 MI 值大于 3（等于）的最高 10 个词项，如下表所示. 从表我们可以看出，搭配词在 R1, R2 和 L1, L2 四个位置上的分布明显比较集中，取[-2, +2]为最佳窗口合适。由于基于语料库的词语搭配研究中通常把与关键词共现频数达到 3 次（等于）以上的, MI 值等于或大于 3 的词作为显著搭配词，所以以上所示的共现搭配词都能与“JALAGV”构成显著搭配, 二者之间有较强的连结关系。搭配词的种数和出现次数分别为 67 和 573。

据考察词语搭配的内部关系，“JALAGV”的各类型搭配分布表如下所示：

结构关系	定体关系	联合关系	体述关系
出现次数	25	7	25

进一步分析可以看出，在 L1、R1 位置上的搭配词基本上都是被关键词支配的名词（构成体述关系）和修饰关键词的词语（构成定体关系）。

4 结论

本文从名词、形容词中各选择了一个代表性的词语，在 100 万词级《现代蒙古语语料库》中统计出了其搭配词语的分布情况，得出这两类词的搭配词语的最佳观察窗口，即：“JALAGV”是[-2, +2]，“GAR”是[-1, +1]。通过这项研究，我们可以得出这样的结论：从本文中我们可以看到在蒙古语中名词、形容词的搭配词语在分布上的差异。首先，(1)搭配词语的分布特征是语言和结构相关的。(2)搭配词语的分布和关键词及其搭配对象的词类有关。再次是，共现频率等于或大于 3，MI 值越大的话两个词的相关性越强, 集中在同一个位置上。

注解

① 全昌勤、刘辉、何婷婷 《基于统计模型的词语搭配自动获取方法的分析与比较》 计算机应用 2005 年 第 19 期

② 邓耀臣、王同顺《词语搭配抽取的统计方法及计算机实现》 外语电化教学 2005 年 105 期

③ 王素格、杨军玲、张武《自动获取汉语词语搭配》 中文信息学报 2006 年 第 6 期

参考文献

[1] 黄昌宁《语料库语言学》商务印书馆 2002年 北京

[2] 格尔泰. 现代蒙古语语法（蒙古文版）[M]. 呼和浩特 内蒙古人民出版社，1999.

[3] 卫乃兴《基于语料库和语料库驱动的词语搭配研究》语言文字学 2002年 第10期

[4] 华沙宝 《现代蒙古语文数据库》程序设计 内蒙古大学学报（人文·哲学版） 蒙文版，1992年第2期，P68-86

[5] 华沙宝 《蒙古语语料库建设现状分析和完善策略》 全国第七届计算语言学联合学术会，2003年，

[6] 邓耀臣 《词语搭配研究中的统计方法》 大连海事大学学报 2003年 第4期

[7] 华沙宝 《对蒙古文语料库的词类标注系统——AYIMAG》内蒙古大学学报（人文科学版）1999年5期