搭配强度计算公式的特点与应用

——以国际中文教育为例 *

张永伟 [1]　梁敬芝 [2]
（1. 中国社会科学院语言研究所语料库暨计算语言学研究中心，北京　100732；
2. 中国社会科学院大学国际教育学院，北京　100102）

摘　要：[**目的 / 意义**] 分析搭配强度计算公式在汉语窗口搭配和依存搭配自动提取中的特点和性能差异，旨在为汉语搭配研究和国际中文教育提供参考。[**方法 / 过程**] 选取 7 种典型的搭配强度计算公式，从真实语料库中为 60 个典型的词语提取窗口搭配和依存搭配，邀请专家进行评分验证后，分析不同公式的性能表现。[**结果 / 结论**] 面向国际中文教育时，公式 Dice 系数、MI^3 和对数似然比在搭配提取中表现较好，而互信息和搭配词频次表现较差，依存搭配提取的精确率普遍高于窗口搭配，并用 MI^3 和 Dice 系数可以取得最高召回率，但仍难以达到 100%。研究结果为搭配强度计算公式的选择和搭配提取工具的研制提供了依据。

关键词：窗口搭配　依存搭配　搭配强度计算公式　语料库
分类号：TP391.1　H195
DOI：10.31193/SSAP.J.ISSN.2096-6695.2025.02.06

0　引　言

搭配（collocation）是一种重复出现的词语组合，具有任意性、结构性和领域相关性等特点，体现了语言的习惯表达方式[1-2]。人们可以从与一个词结伴使用的其他词来了解这个词的含义[3]。语言是文化的载体，搭配是语言组织的规律体现，蕴含着丰富的文化信息。在国际中文教育中，系统化的搭配教学有助于提升汉语词汇教学质量。

搭配由节点词（node）和搭配词（collocate）构成，其中节点词是教学与研究的对象，搭配

* 本文系国家社会科学基金一般项目“融合句法信息的大规模汉语语料库分析工具研制研究”（项目编号：22BYY086）研究成果之一。
[作者简介] 张永伟，男，研究员，研究方向为语料库语言学、计算语言学、计算词典学，Email：zhangyw@cass.org.cn；梁敬芝，女，硕士生，研究方向为语料库语言学、汉语国际教育，Email：2393509280@qq.com。

词是与节点词共现的、辅助节点词学习与理解的其他词语。提取搭配时，搭配强度被用来衡量节点词和搭配词之间关系的紧密程度，搭配强度越大，二者关系越紧密。搭配强度计算公式（又称关联度量）是用于计算搭配强度大小的数学公式。许多语料库分析工具都提供了自动搭配的检索功能，搭配检索是语料库分析工具的核心功能之一[4]。

搭配分为专家搭配和自动搭配，前者由人工提取，后者由计算机自动提取。专家搭配的提取依赖专家的知识和经验，质量高但主观性强且耗时耗力。自动搭配的提取具有客观性和高效性，不受专家主观因素的影响，但存在覆盖面不足、质量不高等问题。如何客观高效地提取高质量搭配，使自动搭配的质量逐渐接近专家搭配，是语料库语言学长期关注的重要研究课题。其中，搭配强度计算公式是最核心的问题之一，旨在利用统计方法模拟专家对搭配的认知。

在国际中文教育领域，非母语学习者缺乏丰富的语言背景知识，词汇学习和运用更具挑战性，有效的词语搭配有助于学习者更全面地理解和掌握汉语词汇的含义及用法。然而，现有搭配提取工具通常提供多种搭配强度计算公式供使用者选择，但并未提供公式选择的指导，导致使用者难以做出合理选择。为此，本研究拟对自动搭配提取的三个关键问题进行探讨：（1）常用搭配强度计算公式有何特点；（2）不同搭配强度计算公式的关系如何，哪些相似度高，哪些差异大；（3）在国际中文教育领域，如何选择搭配强度计算公式。研究结果可以帮助国际中文教育者和汉语学习者等根据实际需求选择合适的搭配强度计算公式，提高搭配提取效率和准确性，提升词语教学和学习质量。

文章各部分安排如下：第 1 部分梳理了搭配自动提取的相关研究，第 2 部分介绍了搭配自动提取和专家验证实验的细节，第 3 部分参照专家评分结果对自动搭配结果进行分析，最后进行总结。

1 搭配自动提取研究概述

1.1 搭配自动提取方法概述

二元词语搭配的自动提取方法有四种：基于窗口的方法通过设定窗口大小，提取与目标词在一定距离内的共现词作为搭配词；基于语法的方法利用句法分析，提取与目标词存在特定语法关系的词语作为搭配词；基于语义的方法借助语义信息、同义词替换、翻译一致性等，判断词语间的语义关联，进而筛选候选搭配词；而基于分类的方法则综合利用上述三种方法的特征，结合机器学习算法对候选搭配词进行分类，判断其是否与结点词构成搭配[5]。当提取的搭配数量较多时，需要由专家进一步识别典型搭配。为了提高专家识别效率，可以引入搭配强度的计算和搭配排序筛选机制，为专家识别提供依据。

基于窗口的方法和基于依存语法的方法提取的搭配分别简称为窗口搭配和依存搭配，相关研究较多。许多语料库分析工具都支持窗口搭配和依存搭配的自动提取，如 CQPWeb、Sketch Engine、English Corpora、WordSmith、AntConc、依存搭配检索系统（DCS）[6]、汉语助研等，支持窗口搭配的提取，CCA 中文搭配助手[7]、DCS 系统等支持依存搭配的提取。

1.2 搭配强度计算公式概述

搭配强度计算公式直接影响自动搭配的提取效果。在直接提取的搭配数量较多，但仅需选取几个典型搭配进行教学与学习时，搭配强度的计算尤为重要。自动搭配在多大程度上可以替代专家搭配，是衡量搭配强度计算公式有效性的重要指标。

Wermter 和 Hahn[8] 将搭配强度计算公式分为基于频次的方法、基于信息熵的方法和基于统计学的方法。许多搭配提取工具都提供了多种搭配强度计算公式，如 CQPWeb 支持互信息（mutual information, MI）[9]、对数似然比（log-likelihood ratio, LLR）[10]、MI³[11]、T 值（t-score）[12]、Z 值（z-score）[13]、Dice 比率（Dice ratio）[14]、对数比率（log ratio）、保守似然比（conservative LR）、排序频率（rank frequency）9 种公式。DCS 系统支持点间互信息、平方互信息（square mutual information, SMI）[15]、T 值、对数比率、对数似然比、Dice 系数（Dice's coefficient，即 Dice 比率）、相对频次、共现频次、搭配词频次 9 种公式[6]。

前人关于搭配强度计算公式的研究通常先从语料库中提取所有二元组，随后运用一种或多种搭配强度计算公式对这些二元组进行量化评估，并将高分搭配作为研究对象。这些研究可大致分为两类：一类以汉字为基本单位，从语料库中提取汉字二元组，通过观察其成词情况来分析搭配强度计算公式的特性[16–17]；另一类则以词语为基本单位，从语料库中提取词语二元组，利用搭配强度计算公式评估这些词语二元组的关联程度，从而识别词语搭配[15, 18–21]。此外，还有少量基于固定节点词的搭配强度计算方法比较研究。如梁敬芝[22] 对 10 种搭配强度计算公式进行了对比分析，但每个节点词仅选取 20 个搭配词进行分析，搭配数量较少，也未探讨并用多种计算公式的效果。然而，值得注意的是，搭配的用途多样，其评判标准也不相同。因此，针对搭配强度计算公式在不同用途时的特性，有必要开展更有针对性的研究。

2 搭配自动提取与专家验证

2.1 搭配自动提取

2.1.1 语料来源与处理

实验选择北京大学 CCL 现代汉语语料库①作为语料来源。下载所有路径（path）包含“txt”的检索行，使用哈尔滨工业大学语言技术平台（LTP）4.3 版（Base1 模型）对检索行的正文进行切分标注，包括分句、分词、词性标注和依存句法分析等。以句子“各国媒体之间保持紧密合作关系，具有重要意义。”为例，经过切分标注后的可视化结果见图 1。

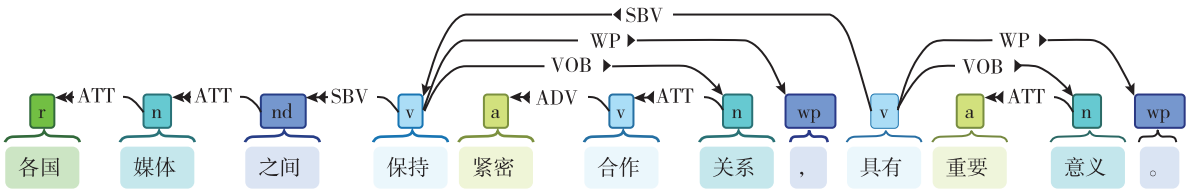

图 1 语料切分标注样例

注：图 1 借助 brat 工具②绘制。图 1、表 3 和表 4 中的词性和依存关系标签含义参见 LTP 在线文档附录③。

以“紧密合作”为例，图 1 显示其被切分为紧密和合作两个词，词性分别为 a（形容词）和 v（动词），“紧密”依存于“合作”（箭头指向的词依存于箭尾指向的词），依存关系为 ADV（状中结构）。

2.1.2　计算公式

梁敬芝[22]对 9 个常见搭配提取工具进行了分析，发现这些工具共支持 30 种搭配强度计算公式。其中，互信息、对数似然比、MI^3、T 值、Z 值以及 Dice 系数（包括其改进形式）至少被 6 个工具支持，是应用最为广泛的计算公式。本研究在计算公式的选择上遵循代表性与多样性相结合的原则，既考虑计算公式支持的普遍性，又兼顾不同的计算公式类型。因此，基于信息熵的方法选择了支持最为广泛的互信息和 MI^3，基于统计学的方法选择了 T 值、对数似然比和 Dice 系数。虽然基于频次的方法未被普遍支持，但搭配频次和搭配词频次常用于衡量搭配词的典型性，因此本研究仍选择了这两种基于频次的方法。最终选择的七种代表性计算公式见表 1。

表 1　搭配强度计算公式详情

名称	简短表示	公式
互信息	MI	$\log_2 \frac{f_{AB}N}{f_A f_B}$
MI^3	MI^3	$\log_2 \frac{f_{AB}^3 N}{f_A f_B}$
T 值	TS	$\frac{f_{AB}-\frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$
对数似然比	LLR	$2\cdot(xlx(f_{AB})+xlx(f_A-f_{AB})+xlx(f_B-f_{AB})+xlx(N)$ $+xlx(N+f_{AB}-f_A-f_B)-xlx(f_A)$ $-xlx(f_B)-xlx(N-f_A)-xlx(N-f_B))$
Dice 系数	Dice	$\frac{2\cdot f_{AB}}{f_A+f_B}$
搭配频次	FreqAB	f_{AB}
搭配词频次	FreqB	f_B

注：公式及符号定义参考 Sketch Engine 在线文档④。其中，f_A 为节点词频次、f_B 为搭配词频次，f_{AB} 为搭配频次，N 为语料库词数，$xlx(N)$ 为 $fln(f)$。

2.1.3　节点词选取

为确保节点词的代表性，实验利用以下标准来选取节点词：（1）为确保研究成果直接服务国际中文教育实践，所有候选词均取自《国际中文教育中文水平等级标准》词汇表，为避免词长对后续专家评分的影响，统一选择双字词；（2）为确保所选词语具有充分的使用价值，仅选取高频词；（3）考虑到多义词在搭配判别过程中可能引发的歧义问题，仅选取单义词；（4）基于现代

汉语词类体系的特点，从名词、动词、形容词这三个主要词类中各选取 20 个代表性词语作为节点词。

具体实验时，词频数据通过 CCL 现代汉语语料库切分标注后统计获得，词义数则依据《现代汉语词典》（第 7 版）确定。最终确定的 60 个节点词详见表 2。

表 2　节点词列表

词类	节点词
名词	政府、部门、产品、总统、政策、银行、专家、方式、价格、事件、原因、农村、方法、机会、作品、医院、选手、行业、眼睛、母亲
动词	认为、知道、举行、提高、继续、包括、实现、增加、获得、达到、造成、产生、宣布、实行、扩大、相信、减少、看见、考虑、离开
形容词	重要、明显、著名、广泛、优秀、复杂、显然、彻底、显著、热烈、独特、准确、详细、出色、愉快、艰难、可爱、精心、有趣、珍贵

注：60 个节点词的平均频次为 61 578.78，其中“珍贵”频次最小（7148），“政府”频次最大（253 572）。

2.1.4　搭配的自动提取

实验使用 DCS 系统对切分标注后的 CCL 现代汉语语料库进行了窗口搭配和依存搭配的提取。实验用 7 个公式为 60 个节点词分别提取搭配强度最高的 50 个依存搭配和 50 个窗口搭配，共获得了 42 000 条搭配信息。为确保搭配具有一定的使用度，凸显公式自身的特点，实验设置搭配的原始频次不低于 2。

在提取窗口搭配时，实验区分搭配词词性，窗口大小为 5。为便于专家验证的同时又简化窗口搭配信息，实验记录搭配词出现在节点词左侧还是右侧，但不记录搭配词在窗口中的具体位置。在提取依存搭配时，实验区分不同搭配词词性和依存关系。以 Dice 系数为例，与“机会”搭配强度最大的 10 个窗口搭配信息和 10 个依存搭配信息分别见表 3 和表 4。

表 3　“机会”的高强度窗口搭配信息（Dice 系数）

序号	搭配词	搭配词频次	左侧频次	右侧频次	搭配频次	搭配强度
1	就业 /tv	28 411	2413	94	2507	0.063 9
2	抓住 /tv	20 423	1357	149	1506	0.042 8
3	创造 /tv	44 875	1633	81	1714	0.036 1
4	提供 /tv	107 429	2327	96	2423	0.030 8
5	利用 /tv	70 505	1529	161	1690	0.028 0
6	把握 /tv	10 445	557	237	794	0.026 3
7	难得 /ta	6160	515	174	689	0.024 5
8	获得 /tv	87 714	1359	188	1547	0.022 5
9	偶然 /ta	7121	573	13	586	0.020 5
10	多 /ta	274 142	1958	1306	3264	0.020 1

注：搭配词的词形和词性用“/”分隔。左侧频次指搭配词位于节点词左侧的频次，右侧频次指搭配词位于节点词右侧的频次，左侧频次 + 右侧频次 = 搭配频次。

表 4　“机会”的高强度依存搭配信息（Dice 系数）

序号	搭配词	搭配词频次	搭配频次	搭配强度
1	就业 /v/att	17 514	2315	0.038 6
2	提供 /v/vob	92 471	3111	0.031 9
3	创造 /v/vob	33 066	1769	0.026 1
4	给 /v/vob	46 456	1839	0.024 7
5	抓住 /v/vob	18 629	1481	0.024 5
6	利用 /v/vob	55 685	1912	0.024 2
7	次 /q/att	140 392	2906	0.023 9
8	好 /a/att	90 204	2120	0.022 0
9	获得 /v/vob	81 539	1562	0.017 0
10	多 /a/att	119 971	1700	0.015 3

注：搭配词的词形、词性和依存关系用“/”分隔。

2.2　搭配的专家验证

实验邀请 6 名国际中文教育硕士生作为专家，对自动提取的搭配进行评分。评分采用 5 分制：确定是有效搭配且需要教学的评 5 分，倾向于是搭配但不是必须教学的评 4 分，不好确定是否为搭配且可以不教学的评 3 分，倾向于不是搭配且对教学没有价值的评 2 分，确定不是搭配且会成为教学负担的评 1 分。其中，“需要教学”指的是该搭配符合国际中文教育的需求，常见且有助于学习节点词。“不是必须教学”指的是虽然是搭配，但对于国际中文学习者来说，掌握该搭配并非必要，或者可以通过其他方式习得。“可以不教学”指的是该搭配的有效性存疑，即使是搭配，其教学价值也不高，不纳入教学内容也不会对学习者造成影响。

评分时，搭配之间彼此独立，不受其他搭配分值和高分搭配数量的影响。搭配需符合国际中文教育需求，搭配词常见并且有助于节点词的学习，比如，将“矿管　部门”作为节点词“部门”的搭配、“减少 3804 万”作为节点词“减少”的搭配时，二者均无助于节点词的学习，此时这两个搭配均应评低分。

提取的 42 000 条搭配信息中，去重后剩余 10 901 个，平均每个节点词有 181.68 个搭配，专家对它们的评分结果统计见表 5。

表 5　专家评分详情

平均分	搭配数量	平均搭配数	比例（%）	累积比例（%）
$0 \leqslant score < 1$	4718	78.63	43.280	43.280
$1 \leqslant score < 2$	3153	52.55	28.924	72.204
$2 \leqslant score < 3$	1407	23.45	12.907	85.111
$3 \leqslant score < 4$	950	15.83	8.715	93.826

续表

平均分	搭配数量	平均搭配数	比例（%）	累积比例（%）
4 ⩽ score ＜ 5	593	9.88	5.440	99.266
score=5	80	1.33	0.734	100.000

本研究对 6 位专家的评分结果进行信度检验，Cronbach's Alpha 系数为 0.918，远超 0.7 的可接受标准，说明专家评分具有很高的一致性和可靠性，在此基础上分析是可靠的。表 5 显示，自动提取的搭配质量呈现倒金字塔分布，低质量搭配占比较大，高质量搭配占比较小。自动提取的搭配质量普遍较低，43.280% 的搭配低于 1 分，累积比例更是表明 72.204% 的搭配低于 2 分，这意味着大多数自动提取的搭配可能不适合教学或教学价值有限。相比之下，高质量搭配（4 分及以上）仅占 6.174%，其中只有 0.734% 被认为是必须教学的，反映了高质量搭配的稀缺性。平均搭配数的分布进一步证实了这一点，每个节点词仅有 1.33 搭配词被所有专家认为是需要教学的（评分为 5 分）。

鉴于高质量搭配的稀缺性，实验将不低于 4 分的搭配视为专家搭配，确保搭配合理数量的同时确保搭配的质量。值得注意的是，实验假设专家搭配已经包含在自动提取的搭配集合中，这是后续对比分析的前提。此外，表 5 的评分结果凸显了在国际中文词汇教学中精确筛选搭配的重要性，高分的专家搭配为自动搭配提取时公式的特点分析与性能评估提供了参考。

3　搭配强度计算公式特点分析

3.1　频次特点分析

本研究对高频搭配词的频次均值、搭配的频次均值以及这两个指标的比值进行统计。其中，比值越小表明搭配词的使用越依赖于节点词。窗口搭配和依存搭配的频次信息见表 6 和表 7。

表 6　高频搭配频次信息（窗口搭配）

公式	搭配词频次均值（a）	搭配频次均值（b）	频次比值（a/b）
互信息	3.02	3.23	0.94
MI^3	1 194 087.53	5 386.03	221.70
T 值	1 674 397.94	6 030.35	277.66
对数似然比	1 362 181.66	5 709.44	238.58
Dice 系数	97 522.42	2 603.02	37.47
搭配频次	1 758 344.53	6 049.60	290.65
搭配词频次	2 158 454.80	4 866.52	443.53

表 7　高频搭配频次信息（依存搭配）

公式	搭配词频次均值（a）	搭配频次均值（b）	频次比值（a/b）
互信息	66.70	16.07	4.15
MI^3	444 574.16	2 208.69	201.28

续表

公式	搭配词频次均值（a）	搭配频次均值（b）	频次比值（a/b）
T 值	471 849.50	2 343.96	201.30
对数似然比	591 928.94	2 111.90	280.28
Dice 系数	107 633.21	2 003.73	53.72
搭配频次	1 106 620.85	2 556.61	432.85
搭配词频次	2 145 381.92	1 124.13	1908.48

表 6 和表 7 显示，各计算公式提取的搭配在频次上的特点如下。

（1）互信息公式。在窗口搭配中，该公式的搭配词频次均值（3.02）和搭配频次均值（3.23）都很小，二者比值为 0.94[⑤]，小于 1 表明该公式倾向于选择低频词作为搭配词[⑥]，且这些词的使用高度依赖于节点词。相比之下，在依存搭配中，该公式的搭配词频次均值（66.70）和搭配频次均值（16.07）明显增大，a/b 比值为 4.15，大于 1 说明受依存关系影响选择了更高频的词作为搭配词。然而，与其他公式相比，无论窗口搭配还是依存搭配，互信息公式的 a/b 比值都明显最小，说明互信息公式提取两种搭配时都更倾向于选择在使用上严重依赖节点词的搭配词。这一特点使得互信息公式在提取罕见但可能具有特殊意义的词语搭配时具有独特优势。

（2）MI[3]、T 值和对数似然比公式。三个公式在窗口搭配和依存搭配中均保持较大的搭配频次和频次比值，显示了它们在提取高频且相关性强的搭配方面保持了较好的平衡性。三个公式倾向于选择高频词作为搭配词，同时又确保所提取的搭配在语料中频率较高。这种特性使得三个公式能够有效反映搭配词的普遍使用情况和词语间的紧密关系。

（3）Dice 系数公式。提取的搭配词频次均值和搭配频次均值相对较低，表明其更倾向于提取相对稳定但不一定高频的搭配，从而捕捉传统高频统计方法可能忽略的语言现象。其频次比值在两种搭配类型中均介于互信息公式和其他公式之间，反映了该公式在评估搭配强度时采取了相对平衡的策略，既考虑了搭配词的独立性，又兼顾了节点词与搭配词的共现频率。

（4）搭配词频次公式和搭配频次公式。与互信息公式相反，搭配词频次公式选择最高频的词作为搭配词，搭配频次公式选择与节点词最高频共现的词作为搭配词。两个公式的频次比值最大，说明它们提取的搭配词在使用上不太依赖节点词。

对比窗口搭配和依存搭配的频次信息不难发现，依存搭配的搭配频次均值普遍低于窗口搭配，但频次比值更高。这表明依存搭配虽然出现频率相对较低，却更能捕捉到词语间的某种关系。

3.2　不同计算公式的性能分析

实验使用精确率（P）和召回率（R）评价不同公式提取搭配的质量，使用 $P@n$ 和 $R@n$ 分别表示得分最高的 n 个搭配的精确率和召回率，计算方法见公式（1）和公式（2）。

$$P@n=\frac{\text{前}n\text{个搭配中正确搭配的数量}}{n}\times 100\% \quad (1)$$

$$R@n=\frac{\text{前}n\text{个搭配中正确搭配的数量}}{\text{专家搭配总数}}\times 100\% \quad (2)$$

公式（1）和公式（2）中，正确搭配指专家评分时，平均分不低于 4 的搭配。精确率衡量了提取结果的精确性，即在提取出的搭配中，有多大比例是正确搭配。召回率衡量了提取结果的完备性，即在专家搭配中，有多少被正确提取出来了。计算 n 在不同大小时的精确率和召回率可以更全面反映公式的搭配提取性能。

3.2.1 不同计算公式的精确率

搭配数量 n 从 5 开始，以 5 为步长递增至 50 时，不同公式提取的窗口搭配和依存搭配的精确率见图 2。

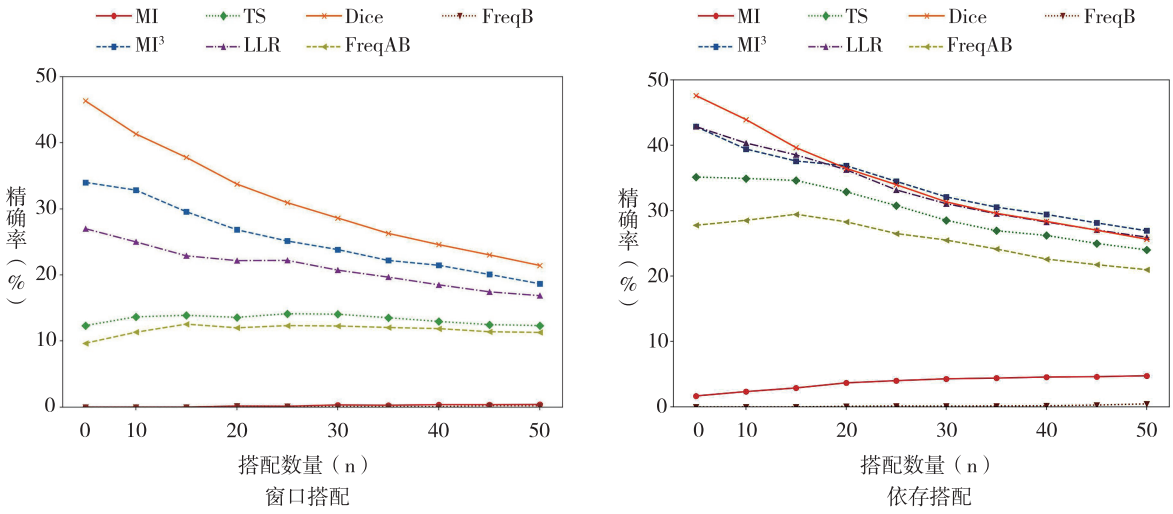

图 2 搭配提取的精确率

图 2 显示，不同公式的精确率差异明显。随着搭配数量的增加，大多数公式的精确率呈下降趋势，但下降程度和模式各不相同。Dice 系数公式整体表现最佳，尤其在搭配数量较小时更为突出。MI^3 和对数似然比公式展现出相似的下降趋势，但 MI^3 的表现略优。T 值和搭配频次公式的精确率较小但相对稳定，精确率始终维持在 9.67%~14.13% 之间。互信息公式和搭配词频次公式在搭配提取任务上的表现最差，精确度远未达到实际应用要求。值得注意的是，通过提高搭配频次的最小阈值，互信息方法的精确率可以得到一定程度的改善。

提取依存搭配时，不同公式的精确率随着搭配数量的增加呈现出与窗口搭配相似的趋势，然而，所有公式的精确率明显更高，充分证明了依存关系在提高搭配提取精确率方面的有效性。具体而言，Dice 系数公式在提取少量搭配（5~20 个）时表现最佳，精确率达到最高水平。随着搭配数量的增加，MI^3 和对数似然比公式在中等数量范围（25~40 个）内保持了较高的精确率，与 Dice 系数一起构成了精确度最高的三个公式。T 值和搭配频次公式虽然略逊于 MI^3 和对数似然比，但其精确度相对稳定，且在各个数量范围内均优于窗口搭配方法。互信息公式和搭配词频次公式随搭配数量的增加精确率有所提升，但仍然是精确度最低的两个公式。

从精确率的角度上来看，推荐 Dice 系数、MI^3 和对数似然比这三个公式，不推荐搭配词频次公式和采用低频阈值设置的互信息公式。

3.2.2　不同计算公式的召回率

搭配数量 n 从 5 开始，以 5 为步长递增至 50 时，不同公式提取的窗口搭配和依存搭配的召回率见图 3。

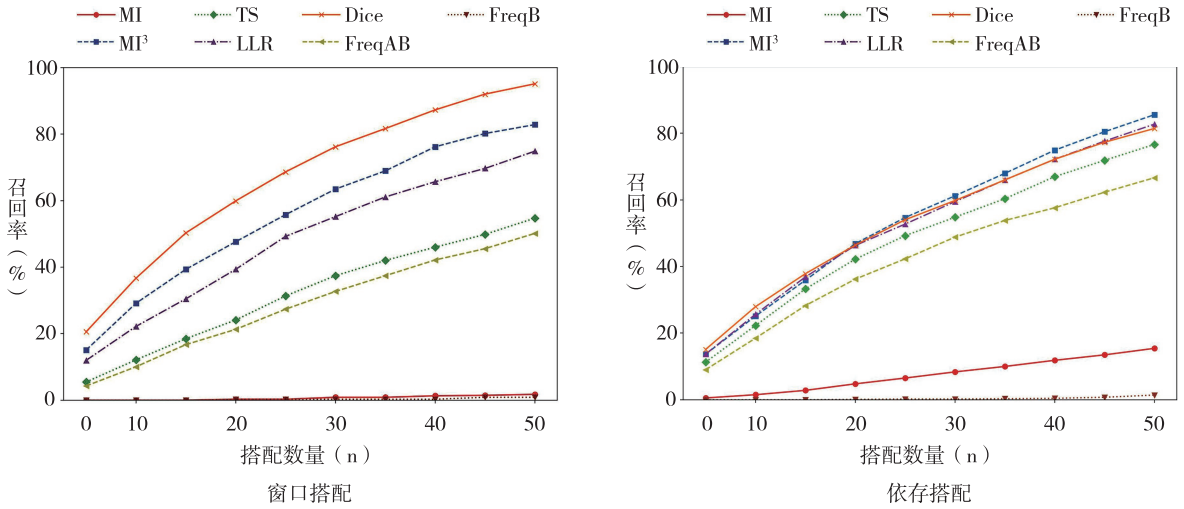

图 3　搭配提取的召回率

图 3 显示，Dice 系数、MI³ 和对数似然比公式在两种搭配类型中都表现出色，随着搭配数量的增加，它们的召回率显著提高。特别是在窗口搭配中，Dice 系数公式表现最佳，而在依存搭配中，MI³、对数似然比和 Dice 系数的表现相近且较好。相比之下，T 值和搭配频次公式表现中等，虽然召回率随搭配数量增加而稳步提高，但整体效果不如前三个公式。值得注意的是，互信息公式和搭配词频次公式在两种搭配类型中都表现较差，即使在搭配数量增加的情况下，其召回率仍然很低。大多数公式的召回率增长趋势随搭配数量增加而上升，但在搭配数量达到 20~30 时，增长速度开始放缓。

提取的全部 50 个搭配中，单一公式在窗口搭配上取得的最高召回率为 95.12%（Dice 系数），在依存搭配上取得的最高召回率为 85.54%（MI³）。从召回率的角度上来看，依然推荐 Dice 系数、MI³、对数似然比这三个公式，不推荐搭配词频次公式或采用低频阈值设置的互信息公式。对比图 2 和图 3 发现，公式不同，搭配自动提取效果差异显著。

3.3　计算公式相关性分析

本研究计算了不同公式提取的搭配之间的一致性，并将其作为衡量不同公式相关性的指标。为了更直观地呈现公式之间的相关性，绘制了公式相关性热力图（图 4、图 5）。图中单元格颜色深浅表示相关性大小，单元格中记录了相关性数值，最小值为 0.00，最大值为 1.00。为分析搭配数量对公式相关性的影响，本研究分别基于排序最高的前 25 个（n=25）搭配和排序最高的前 50 个（n=50）搭配进行绘制。

本研究为窗口搭配和依存搭配分别绘制了公式相关性热力图。窗口搭配的公式相关性热力图见图 4。

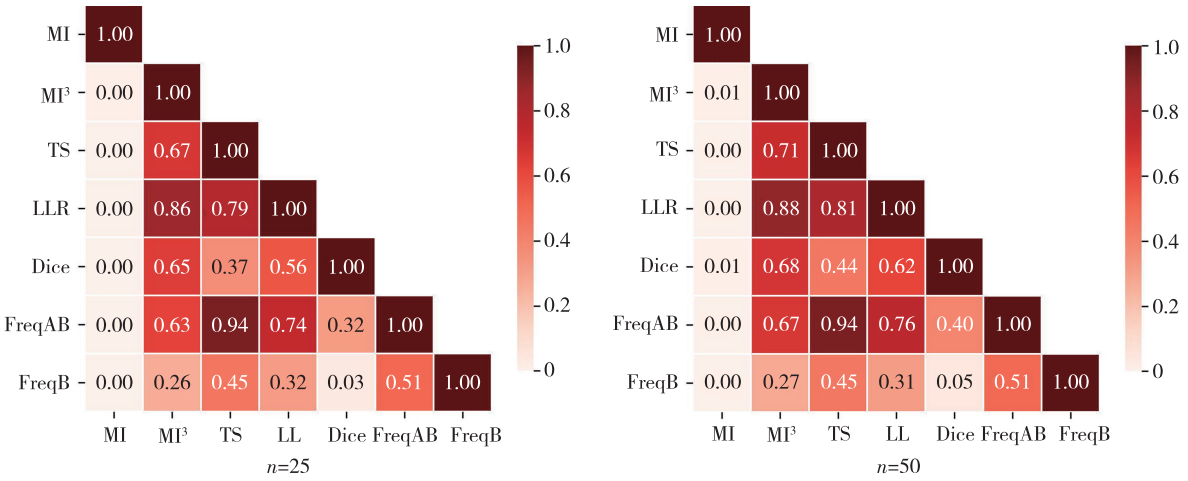

图 4　计算公式相关性热力图（窗口搭配）

图 4 显示，互信息公式相关的区域呈现明显的浅色，表明它与其他公式都有极低的相关性。搭配词频次公式相关的区域大多呈现浅色，仅与搭配频次公式之间有一个相对较深的单元格，显示了它们之间的中等程度相关性。相比之下，热力图的中心区域呈现出一个较为明显的深色区域，主要由 MI^3、T 值、对数似然比和 Dice 系数四个公式组成，表明它们之间存在较高的相关性。特别是 T 值与搭配频次公式之间，以及 MI^3 与对数似然比公式之间的单元格颜色最深，反映了它们之间极高的相关性。Dice 系数公式在热力图呈现出中等深度的颜色，主要与 MI^3 和对数似然比公式有一定程度的相关性。而搭配词频次公式相关的区域大部分呈现浅色，仅与搭配频次公式的单元格颜色稍深，显示出它与大多数公式较低的相关性。总体而言，窗口搭配的热力图的颜色分布模式在搭配数量为 25 和 50 的情况下基本一致，说明提取窗口搭配时，搭配数量的变化对公式间相关性的影响不大。

依存搭配的公式相关性热力图见图 5。

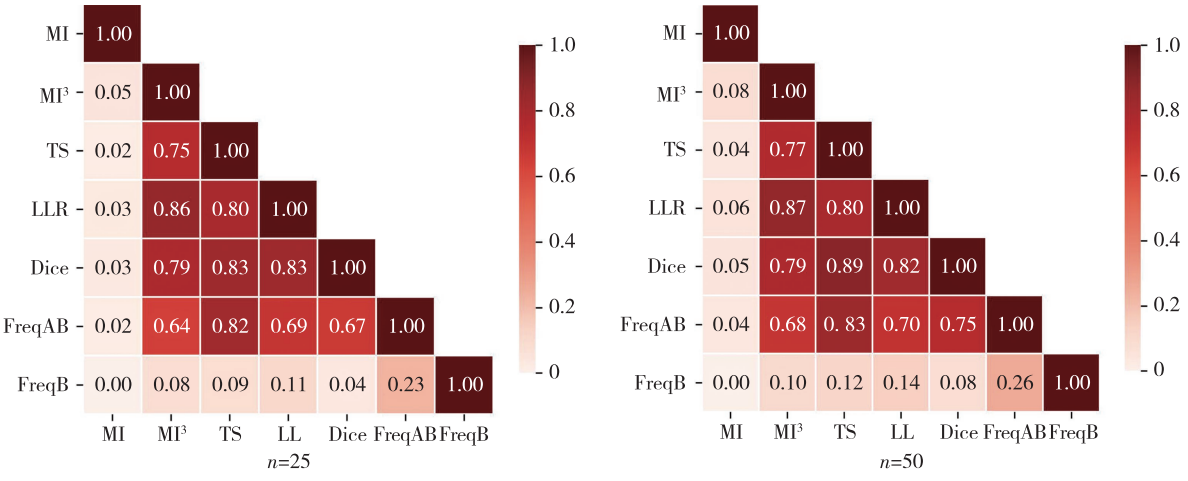

图 5　计算公式相关性热力图（依存搭配）

图 5 显示，在提取依存搭配时，互信息公式和搭配词频次公式与其他公式之间具有极低的相关性，MI³、T 值、对数似然比和 Dice 系数四个公式之间展现出高度相关性。

同窗口搭配提取相比，在提取依存搭配时，搭配词频次与其他公式之间的相关性普遍降低，而 Dice 系数与其他公式之间的相关性普遍提高。这一现象表明，搭配类型的差异对各计算公式之间的相关性具有较大影响。提取搭配时，需要根据具体的研究目的和搭配类型选择合适的公式。

3.4　并用两个计算公式的召回率

单一公式的召回率显示，在搭配数量为 50 时，提取窗口搭配和依存搭配的最高召回率仅为 95.12% 和 85.54%。在实际应用中，为了能快速提取搭配，通常会提前制作一个大的搭配库，为了尽可能多地包含各种搭配，同时使用多个公式提取搭配结果。本研究重点分析了使用两个公式提取搭配的召回率。

由于搭配词频次在提取搭配时表现出极低的召回率，研究未将其纳入分析范围。将其余六个公式两两组合，并将各自提取的搭配进行等量合并。每个公式提取的搭配数量从 5 开始，以 5 为步长递增至 50 时，窗口搭配和依存搭配的召回率见图 6。

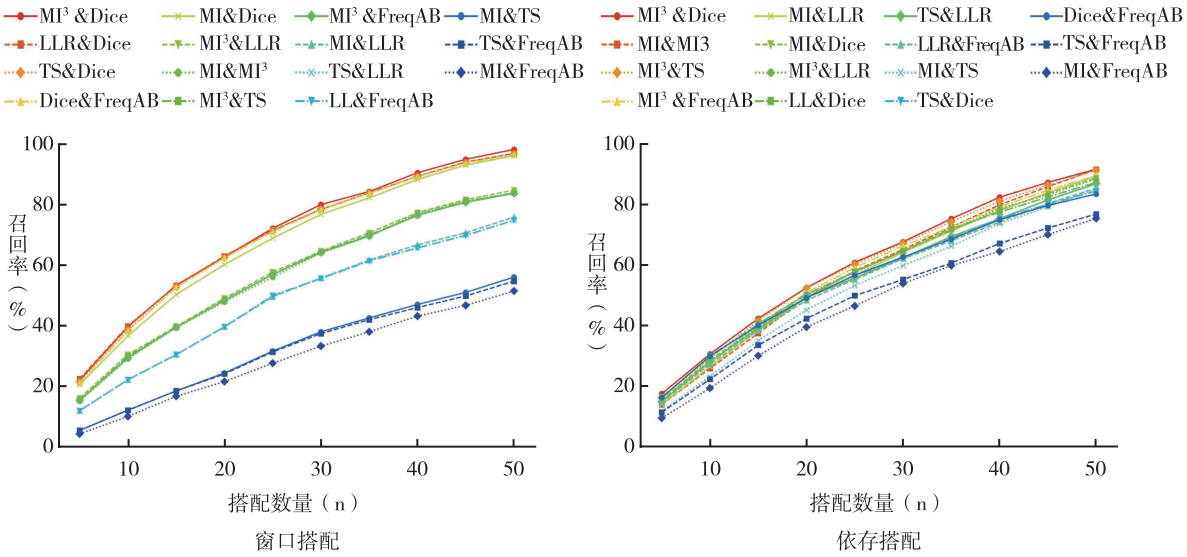

图 6　搭配提取的召回率（并用两个公式）

注：在计算过程中，当两个公式均提取出相同的正确搭配时，该搭配只计算一次，不重复计算。

图 6 显示，对于两种搭配类型，所有曲线都呈现出上升趋势，表明随着提取搭配数量的增加，召回率普遍提高。窗口搭配中 MI³&Dice 组合的曲线最陡，且终点最高，达到 98.22% 的召回率，而依存搭配中同样是 MI³&Dice 组合表现最佳，但最高点仅为 91.41%，较单一公式的召回率分别提高了 3.10% 和 5.87%。这种差异说明了窗口搭配方法在召回率方面具有明显优势。窗口搭配中不同公式组合的曲线之间存在较大的间距，表明公式选择对窗口搭配召回率的影响更为显著。而依存搭配的曲线则相对集中，说明公式选择对依存搭配召回率的影响相对较小。因此，为窗口搭配选择公式组合时需要更加谨慎，因为不同组合可能导致显著的召回率差异。

提取窗口搭配时，包含 Dice 系数公式的召回率普遍较高。提取依存搭配时，包含 MI^3 公式的召回率普遍较高。同时使用二者提取窗口搭配和依存搭配时，可以获得每个搭配类型最高的召回率。此外，少数公式组合在两种搭配类型中表现差异较大，比如 T 值和 Dice 系数公式的组合提取窗口搭配时具有较高的召回率，但提取依存搭配时排名靠后。因此，在选择公式组合时需要考虑具体的搭配类型，不能简单地套用。图 6 同时显示，即使各自提取 50 个搭配，所有公式组合仍未达到 100% 的召回率。这说明增加搭配数量和使用公式组合可能都无法完全解决召回率问题。在实际应用中，需要合并窗口搭配和依存搭配结果，或者探索更高效的公式组合来实现更全面的搭配提取。

4 结 语

本研究通过分析不同搭配强度计算公式在窗口搭配和依存搭配提取中的表现，揭示了它们在提取自动搭配时的特点和性能差异。通过绘制相关性热力图，直观地展示了不同公式之间的关系。此外，本研究还对单一公式提取搭配的精确率和召回率、并用两个公式提取搭配的召回率进行了分析，系统地比较了不同公式以及公式组合的性能表现。

本研究的结果不仅对汉语语料库语言学领域的搭配研究具有重要意义，也可以为国际中文教育领域的搭配教学和研究提供有益参考。通过深入理解不同公式的特点和性能差异，汉语研究者和国际中文教育者等可以更加有针对性地选择和应用这些公式，提高搭配提取效果和词汇教学效果。同时，本研究也为开发更高效、更准确的汉语搭配提取工具提供了理论依据。

未来工作主要有：（1）扩大公式分析范围，更全面地评估不同公式在自动搭配提取中的表现；（2）使用更大规模、更多样化的语料，增加分析的可靠性；（3）扩大词语的选择范围，包括更多词性、不同使用频率的词语，更全面地评估不同公式在各类词语搭配提取中的表现；（4）在分析现有公式的基础上，探索如何改进公式。

【注释】

① http://ccl.pku.edu.cn:8080/ccl_corpus.

② https://brat.nlplab.org/index.html.

③ http://ltp.ai/docs/appendix.html.

④ https://www.sketchengine.eu/documentation/statistics–used–in–sketch–engine/.

⑤搭配由节点词和搭配词组成，除节点词外，搭配词还可以和其他词一起使用，因此搭配词频次通常不小于搭配频次。但在窗口搭配中，搭配词出现在两个相同节点词中间时，将导致搭配词频次小于搭配频次的情况，进而导致二者的比值小于1。比如，对“……重要贸易港口，对于斯瓦希里文化的发展具有重要意义……”文本提取搭配时，以“重要”作为节点词，“斯瓦希里”作为搭配词时，“斯瓦希里”的频次为1，而“重要”和它组成的搭配的频次为2，出现了搭配词频次小于搭配频次的情况。

⑥实验中互信息倾向于选择低频搭配组合与设定的搭配最小频次阈值为2相关。在使用互信息公式提取窗口搭配时，不同的阈值设置会显著影响搭配提取的结果。

【参考文献】

[1] Sinclair J. Corpus Concordance Collocation [M]. Oxford: Oxford University Press, 1991.

[2] 孙茂松，黄昌宁，方捷. 汉语搭配定量分析初探 [J]. 中国语文，1997（1）：29–38.

[3] Firth J R. Papers in Linguistics 1934–1951 [M]. Oxford: Oxford University Press, 1957.

[4] 张永伟，吴冰欣. 基于网络的第四代语料库分析工具核心功能评介 [J]. 当代语言学，2023，25（4）：611–624.

[5] Wong K F, Li W, Xu R, et al. Introduction to Chinese natural language processing [M]. San Rafael: Morgan & Claypool Publishers, 2009.

[6] 张永伟，马琼英. 面向语文辞书编纂的词语依存搭配检索系统研究 [J]. 辞书研究，2022（4）：30–40，125.

[7] 胡韧奋，肖航. 面向二语教学的汉语搭配知识库构建及其应用研究 [J]. 语言文字应用，2019（1）：135–144.

[8] Wermter J, Hahn U. Collocation extraction based on modifiability statistics [C]//Proceedings of the 20th International Conference on Computational Linguistics. 2004: 980–986.

[9] Church K, Hanks P. Word association norms, mutual information, and lexicography [J]. Computational Linguistics, 1990, 16(1): 22–29.

[10] Dunning T E. Accurate methods for the statistics of surprise and coincidence [J]. Computational Linguistics, 1993, 19(1): 61–74.

[11] Oakes M P. Statistics for Corpus Linguistics [M]. Edinburgh: Edinburgh University Press, 1998.

[12] Church K, Gale W A, Hanks P, et al. Using statistics in lexical analysis [M]//Zernik U, ed. Lexical acquisition: exploiting on–line resources to build a lexicon. New York: Psychology Press, 1991: 115–164.

[13] Berry–Rogghe G. The computation of collocations and their relevance in lexical studies [M]//Aitken A, Bailey R, Hamilton–Smith N. The Computer and Literary Studies. Edinburgh: Edinburgh University Press, 1973: 103–112.

[14] Dice L R. Measures of the amount of ecologic association between species [J]. Ecology, 1945, 26(3): 297–302.

[15] Zhang H, Zhang Y, Yu J. Collocation extraction using square mutual information approaches [J]. International Journal of Knowledge and Language Processing, 2011, 2(1): 53–58.

[16] Sproat R, Shih C. A statistical method for finding word boundaries in Chinese text [J]. Computer Processing of Chinese & Oriental Languages, 1990, 4(4): 336–351.

[17] 罗盛芬，孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究 [J]. 中文信息学报，2003（3）：9–14.

[18] 孙健，王伟，钟义信. 基于统计的常用词搭配（Collocation）的发现方法 [J]. 情报学报，2002，（1）：12–16.

[19] 王大亮，张德政，涂序彦，等. 基于相对条件熵的搭配抽取方法 [J]. 北京邮电大学学报，2007，30（6）：40–45.

[20] Qian Y. Dynamism of collocation in L2 English writing: a bigram–based study [J]. International Review of Applied Linguistics in Language Teaching, 2022, 60(2): 339–362.

[21] Su Q, Gu C, Liu P. Association measures for collocation extraction: automatic evaluation on a large–scale corpus [J]. International Journal of Corpus Linguistics, 2024, 29(1): 59–86.

[22] 梁敬芝. 词语搭配强度计算公式的特点及其对国际中文教育的启示 [D]. 北京：中国社会科学院大学，2024.

张永伟，梁敬芝 . 搭配强度计算公式的特点与应用——以国际中文教育为例［J］.
文献与数据学报，2025，7（2）：069–083.

Features and Applications of Collocation Strength Calculation Formulas: Taking International Chinese Language Education as an Example

Zhang Yongwei[1]　Liang Jingzhi[2]

（1. Corpus and Computational Linguistics Research Center, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China;
2. School of Global Education & Development, University of Chinese Academy of Social Sciences, Beijing 100102, China）

Abstract:［**Purpose/Significance**］This study aims to analyze the characteristics and performance differences of collocation strength calculation formulas in the automatic extraction of Chinese window-based collocations and dependency-based collocations, providing references for Chinese collocation studies and international Chinese language education.［**Method/Process**］Seven typical collocation strength calculation formulas were selected to extract window-based collocations and dependency-based collocations for 60 representative words from authentic corpora. Following expert scoring validation, the performance of different formulas was analyzed.［**Result/Conclusion**］Regarding international Chinese language education, formulas such as Dice coefficient, MI^3, and log-likelihood ratio performed well in collocation extraction, while mutual information and collocate frequency showed poor performance. The precision of dependency-based collocation extraction was generally higher than that of window-based collocation extraction. Furthermore, the simulataneous use of MI^3 and Dice coefficient achieved the highest recall rates but still could not reach 100%. These findings provide a basis for selecting collocation strength calculation formulas and developing collocation extraction tools.

Keywords: Window-based collocation; Dependency-based collocation; Collocation strength calculation formula; Corpus

（本文责编：孔青青）