Journal of Technology and Chinese Language Teaching Volume 6 Number 1, June 2015
http://www.tclt.us/journal/2015v6n1/zhan.pdf pp. 1-15

汉语述补结构数据库的构建及其可视化研究[1]
(Development of an Online Database for Chinese Resultative Verb Compounds and Its Application Based on Data Visualization Technology)

詹卫东
(Zhan, Weidong)
北京大学
(Peking University)
zwd@pku.edu.cn

马腾
(Ma, Teng)
腾讯计算机系统公司
(Tecent, Inc.)
matengneo@gmail.com

田骏
(Tian, Jun)
北京大学
(Peking University)
61347084@qq.com

砂岡和子
(Sunaoka, Kazuko)
早稻田大学
(Waseda University)
ksunaoka@gmail.com

摘要：本文简述了“现代汉语述补结构用法词典”在线数据库的构建情况，提出基于该数据库、大规模语料库及已有的语义知识库，用事件语义相关度计算的方法来度量两个谓词性成分（V1，V2）构成述补结构的可能性，并探讨了述补结构用法词典数据及事件语义相关度计算的可视化呈现及其在对外汉语教学中的应用。

Abstract: This paper describes the development of an online database for Chinese resultative verb compounds. Based on the database and related linguistic resources, including a semantic lexicon and a large-scale corpus, the authors propose a new computing method using semantic correlations to determine if two verbs can be formed as a Resultative Compound. In order to use the database and the computational method to support Chinese teaching and learning as a second language more efficiently, a web-based demo program is developed with the help of data visualization technology.

关键词：述补结构，在线词典，复合事件，语义关联度，可视化

Keywords: Resultative verb compound, Online database, Composite events, Semantic correlation, Data visualization

[1]本文研究工作得到国家社科基金面上项目“语言知识资源的可视化技术研究”（项目号：12BYY061）、教育部人文社会科学研究项目规划基金项目“现代汉语述补结构网络数据库的构建与应用”（项目号：12YJA740104 ）、以及国家社科基金重大项目“汉语国际教育背景下的汉语意合特征研究与大型知识库和语料库建设”（项目号：12&ZD175）资助，特此致谢。

1. 引言

现代汉语述结式由两个谓词性成分（本文记作 V1-V2）黏合而成（比如“吃饱、哭肿、唱红、洗干净、摆放整齐”等）。该结构典型的语义模式是：V1 所表示的动作导致出现 V2 所表示的状态。比如在“吃饱”中，“吃”这个动作行为导致其主体（一般是人或动物）处于“饱”的状态。很显然，V1 和 V2 之间应该有事理上的因果联系，而如果缺乏这种联系，就无法构成述结式，比如“吃饿”一般情况下就不是一个合格的述结式，因为作为 V1 的“吃”按常理不会导致“饿”这个结果状态。对于母语者来说，V1 跟 V2 是否能构成述结式，似乎不是一个问题。但是，对于计算机理解中文信息来说，判断一个 V1-V2 组合是否构成述结式，却并不简单。因为到底如何判断“事理上的因果联系”以及什么样的因果联系能用述结式这样的结构来编码表达——这实际上是一个涉及到深层语义理解的复杂问题。根据砂岡和子（2013）的考察，汉日机器翻译对述结式的翻译就存在很多问题。另外，对于很多非母语者来说，汉语的述结式也是比较独特、不容易掌握的一种结构，在其他语言中可能需要用两个小句或其他复杂的动词性结构来表达的事件因果联系，在汉语的述结式中则可以用一个 V1-V2 黏合型的紧凑的谓词性结构来表达，这种结构和语义上的非常大的错位，往往使得非母语者很难把握其使用条件，因而在阅读理解时很容易误解汉语述结式的意思，而在自己的表达中，则会倾向于避免使用 V1-V2 述结式[2]。

汉语学界以往对汉语述补结构进行过广泛和深入的研究，但主要是集中在 V1-V2 整体的论元结构如何由 V1 和 V2 各自的论元结构导出、述结式与相关句法结构（如“把”字句、重动句等）的互动、述结式的认知研究等方面（参见：Li 1990，黄锦章 1993，王红旗 1995，郭锐 1995,2002，袁毓林 2001，施春宏,2008，宋文辉 2007 等），而对于述结式的能产性问题，即什么样的 V1 和 V2 会构成述结式，却关注的不多。下面四个例子，显示了在实际使用中，汉语 V1-V2 述结式具有很强的能产性。

[1] 他演哭戏很感人，把导演都给哭哭了。
[2] 吃懂法兰西
[3] 别让公共场所的劣质洗手液“洗脏”了你的手
[4] 中国学生吻瘫美国机场一个热吻引发的思考

“哭哭”表面上是一个动词重叠形式（比如“哭哭闹闹”），但在例 1 中，却是典型的述结式，前一个“哭”是“他哭”，后一个“哭”是“导演哭”。两个“哭”分属不同的施事论元。例 2 中的“吃懂”是一个少见的组合，但其语义模式也符合典型述结式的要求，V1“吃”的结果是导致其主体（人）更懂得法国（的

[2]我们曾经考察过母语为日语的汉语初级学习者所写的 58 篇汉语作文，发现述结式使用频率很低，但只要用了，一般都没有用错。暨南大学唐玲的硕士论文《印尼留学生粘合式述补结构习得状况研究》）（2004 年）也注意到了类似的现象。该文考察发现印尼留学生学习汉语粘合式述补结构时，结果补语正确使用率高，但掌握的补语量非常少。

文化）了，即 V2“懂”所指示的状态。例 3 中 V1“洗”导致了一个不合常理的结果“脏”（跟预期的符合常理的结果“干净”相反）。例 4 是文章的标题，其中的“吻瘫”更是表达了极为少见的两个事件之间的因果联系：“中国留学生接吻”（事件 1）导致“美国机场陷入瘫痪状态”（事件 2）。

显然，一方面，V1-V2 述结式的构成形式及其用法特点很像是一般的复合动词（compound verb），另一方面，V1-V2 又像一般的短语结构那样是能产、开放的（参见下文第二节的讨论）。因此，从面向计算机的汉语信息处理以及面向非母语者的汉语教学的需要来说，关于 V1-V2 构成述结式的判别条件，就是一个很值得探讨的问题。就这个问题，本文提出的思路是：对于大量常见的述结式，可以像对待一般复合动词那样，以词典数据库形式来描写其基本构成与用法特点。而对于 V1-V2 临时组合能否形成述结式，则可以从复合事件中两个子事件之间的语义相关度计算的角度，对 V1-V2 构成述结式的可能性进行估计。下文第二节扼要介绍我们在“述补结构用法词典”方面做的工作；第三节讨论对 V1-V2 的语义相关度进行计算的具体方法；第四节介绍利用可视化技术展示述补结构用法词典的相关研究成果及其在对外汉语教学中的可能应用；第五节是结语。

2. 现代汉语述补结构用法词典

由于汉语述补结构的独特性，对外汉语教学界一直把述补结构的教学作为重点和难点。针对述补结构的教学策略之一就是，把具体的用例尽可能多的穷举出来，加以细致的描写。比如北京语言大学王砚农等（1987）和刘月华等（1998）就分别编纂了述补结构的专题词典。前者集中在结果补语述结式，后者则是趋向补语的用法详解。不过，这些传统的纸本工具书在使用便利性、例句的鲜活性等方面还是有一些不足，为了学习者和研究者在互联网环境中能更有效地使用电子化的语言资源，在 2007 年到 2010 年间，北京大学与日本早稻田大学合作，构建了“现代汉语述补结构用法词典”的在线数据库（A Database for Chinese ResultativeVerb Compounds，以下简称 DCRVC），通过互联网供学习者和研究者使用（网址：http://ccl.pku.edu.cn/vc）。目前 DCRVC 收集的述补条目共 21031 条，其中述结式 7942 条。主要描述的信息包括（1）述补结构的释义；（2）述补结构的事件语义角色；（3）述补结构的用例（每条至少 3 个汉语例句，有部分条目还有汉语例句的日语和英语译文）；（4）述补结构的类型（分为“结果补语、趋向补语、可能补语、程度补语、介词补语 5 类），等等。下面是通过网页显示的述结式“吃遍”在数据库中的部分信息。

表 1：现代汉语述补结构用法词典条目示例

吃1 【chi1】	词性：动词	HSK：甲	该述语有 2 个义项，点击下方链接查看各义项 吃1 吃2	
吃遍	chi1 bian4	补语类型：结果补语	CCL频次：24	来源：砂冈

	中文	日文	英文
释义	某个地方的或某类食品、药物等都品尝过。		
例句	1. 小李吃遍了京城的美食楼。 2. 小时候，家里穷，各种野菜都吃遍了。 3. 他从小就是个病秧子，吃遍了各种药也不见成效。	1. 李さんは北京のグルメを食べ尽くした。 2. 幼い頃、家が貧しかったので、あらゆる種類の野草を食べ尽くした。 3. 彼は子供のころから病気が多く、あらゆる種類の薬を服用したが効果はない。	1. Mr. Li has eaten all the delicious dishes of the capital city. 2. When I was a child my family was poor that we ate all kinds of vegetables. 3. He has always been a ill boy since he was little, he has taken various kinds of medicine but did not see the results.

为便于了解 DCRVC 中数据的总体情况，下面给出一些统计数据。

表 2：DCRVC 条目数统计表

类型	数据	
述语	词	1639
	义项	2014
补语	词	494
	义项	580
述补结构	21031	

表 3：DCRVC 各类型补语统计表

补语类型	数目	百分比
结果补语	7942	37.76%
趋向补语	7267	34.55%
可能补语	3390	16.12%
程度补语	1336	6.35%
介词补语	1096	5.21%
总数	21031	100%

表 4：DCRVC 述语分组带不同类型补语条目统计

动词类别	结果补语	趋向补语	可能补语	程度补语	介补补语
HSK(甲)	2110(26.57%)	1516(20.86%)	939(27.70%)	400(29.94%)	238(21.72%)
HSK(乙)	2552(32.13%)	2563(35.27%)	1057(31.18%)	437(32.71%)	364(33.21%)
HSK(丙)	1148(14.45%)	1131(15.56%)	513(15.13%)	199(14.90%)	159(14.51%)
HSK(丁)	736(9.27%)	642(8.83%)	226(6.67%)	90(6.74%)	88(8.03%)

非 HSK	1396(17.58%)	1415(19.47%)	655(19.32%)	210(15.72%)	247(22.54%)
合计	7942(37.76%)	7267(34.55%)	3390(16.12%)	1336(6.35%)	1096(5.21%)

我们同时也考察了北语王砚农等（1987）编的述补词典的条目以及北大计算语言所“人民日报分词和词性标注语料库”[3]中的述结式用例。其中，北语述补词典中共有述语 984 个，补语 321 个，述补条目 4106 个。人民日报语料库中述结式用例中共有述语 1641 个，补语 220 个，述补结构 3835 个。比对 DCRVC 跟这两个述结式数据源，可以发现，尽管 DCRVC 中已经收录述结式近 8000 条，但在这两个述结式数据源中未出现（未登录）的记录条数仍占相当高的比例。如下面表 5、6 所示：

表 5：DCRVC 中未登录北语词典述结式条目的比例统计

	《北语词典》总条数	**DCRVC 中未登录条数**
述语	984	134（13.62%）
补语	321	46（14.33%）
述补结构	4106	1750（42.62%）

表 6：DCRVC 中未登录人民日报语料库中述结式条目的比例统计

	《人民日报》述结式用例数	**DCRVC 中未登录条数**
述语	1641	876（53.38%）
补语	220	34（15.45%）
述补结构	3835	2368（61.74%）

以上考察说明，尽管用穷举的办法可以列出相当数量的述结式，但由于述结式的能产性，以词典列举词条的方式来描写述结式，还是有一定局限的。因此仍有必要对 V1-V2 构成述结式的条件做进一步深入分析。下面就来讨论从可计算的角度判断 V1-V2 构成述结式的定量方法。

3. V1-V2 的事件语义关联度计算

3.1. 基于复合事件语义关联的 V1-V2 述结式分析框架

比较容易想到的一个思路是：DCRVC 数据库中已有的述补结构实例可以看作是比较典型的述结式（范例）的集合。对于一个新出现的“V1-V2”组合，可以通过比较它跟现有的述结式范例的相似程度，来估计这个新的“V1-V2”组合是否构成述结式。

[3]语料包含 1998 和 2000 年两年全年的《人民日报》文字内容，五千多万字。本文统计所用的述补结构用例材料由北大计算语言所段慧明老师提供。特此致谢。

不过，单纯基于相似度来评估 V1-V2 构成述结式的可能性，也有可能造成误判。比如在已有述结式数据库中有“放跑”“放走”等实例，现在要判定“摆 - 走”构成述结式的可能性，基于“摆”跟“放”有相似性，而“走”跟“跑”也有相似性，就容易把“摆-走”看作是述结式，但这显然与一般人语感不符。这实际上就又回到了 V1-V2 构成述结式的语义模式问题，即 V1 跟 V2 之间一般应具有“致使-结果”的事件关系。可见，跟比较 V1-V2 与已知述结式范例之间的相似程度相比，更合理的方法是估计 V1 和 V2 之间是否具有“致使-结果”事件语义联系。

从事件语义关系的角度来看，现代汉语的述结式可以看作是一个复合事件的压缩编码形式（詹卫东 2013）。如下面例子所示：

	事件 1	事件 2	复合事件（压缩编码形式）
[5]	妈妈喂女儿。	女儿饱了。	妈妈喂饱了女儿。
[6]	张三洗衣服。	衣服干净了。	张三把衣服洗干净了。

两例中都是事件 1 的发生导致了事件 2 的发生，并且这两个事件存在着共有事件角色。例 4 中的共有事件角色是“女儿”，例 5 中的共有事件角色是“衣服”。复合事件的语义结构可以用下面图 1 表示。

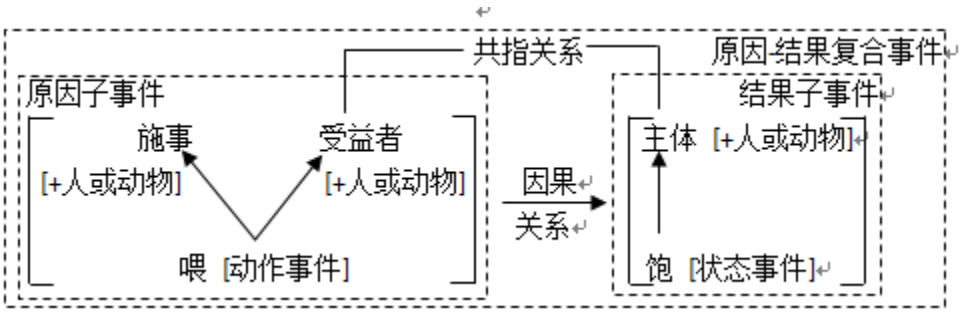


图 1：复合事件“喂饱”的事件语义结构示意图

基于上面这样的分析框架，可以把影响 V1-V2 构成述结式的条件归纳为：

条件 1：事件 1 和事件 2 存在共有事件角色；

条件 2：事件 1 和事件 2 之间存在“致使-结果”的语义关系。

对两个谓词性成分构成述结式的合法性的研究，不仅要回答它们能否构成一个述结式，还要回答所构成的述结式的可接受度有多大的问题。在上面这两个条件的基础上，我们提出关于述结式中复合事件的语义关联度的两个假设。

假设 1：构成述结式的两个谓词性成分（V1，V2）所激活的事件（记作 A，B）之间必须存在共有事件角色，且共有事件角色在这两个事件中的凸显程度影响了述结式中复合事件的语义关联度。

假设 2：构成述结式的两个谓词性成分所激活的事件之间存在“致使-结果”的语义关系，“致使-结果”复合事件的语义相关度影响了述结式中复合事件的语义关联度。

据此，我们提出述结式中复合事件语义关联度的计算公式，这也是 V1-V2 构成述结式的可接受度的定量描述公式：

$$\mathrm{ER} = \alpha * ER_1 + (1-\alpha) * ER_2 \text{（公式 1）}$$

公式 1 中 ER 表示复合事件的语义关联度；ER1 是共有事件角色在两个子事件中的凸显程度。ER2 是两个子事件之间的“致使-结果”语义相关度。α 是权重，用于调节上述两个因素对计算结果的贡献程度，可根据实验情况调整。下面分别说明公式 1 各部分的具体计算方式。

3.2. 共有事件角色凸显度计算（ER1）

计算 ER1 的具体步骤如下：

1) 获取事件角色：以句子为单位，在语料中抽取包含 V1、V2 的句子中共现的名词性成分，构成 V1、V2 所代表的事件 E_1、E_2 的潜在事件角色集合 R_1、R_2；
2) 抽取共有事件角色：集合 R_1 和 R_2 的交集，构成事件 E_1 和 E_2 的共有事件角色集合 R_0；
3) 分别计算 R_0 中的事件角色在整体事件角色集合中所占的比例，取二者的最小值作为结果输出。

具体的计算公式如下所示。

$$ER_1 = \min\left(\frac{\sum_{r\in R_0} C(r)}{\sum_{r\in R1} C(r)}, \frac{\sum_{r\in R_0} C(r)}{\sum_{r\in R2} C(r)}\right) \text{（公式 2）}$$

其中，C(r)是事件角色 r 在事件中出现的频次。这里不妨看一个例子：对于“吃-饱”来说，计算“吃”和“饱”这两个事件的共有事件角色在事件中的凸显程度的过程如下：

1) 分别抽取“吃”和“饱”这两个事件的事件角色，结果如下：

R_1（吃）：{“饭”：7863, “人”：5269, “东西”：2357, “晚饭”：1924, “肉”：1562, “菜”：1187, “药”：1044, “午饭”：992, “时候”：903, “水”：825，……}

R_2（饱）：{“肚子”:380, “饭”：371, “人”：285, “口福”：90, “人们”：64, “墨”：48, “肚皮”：48, “笔”：45, “酒”:42, “书”:42，……}

2) 抽取共有事件角色，结果如下：

R_0：{“饭”，“人”，……}

3) 根据公式 2 进行计算：

$$ER_1 = min\left(\frac{7863 + \cdots + 5269}{7863 + 5269 + \cdots + 825},\ \frac{371 + \cdots + 285}{380 + 371 + \cdots + 42}\right) = 0.6467$$

3.3. V1-V2 复合事件语义关联度计算（ER2）

公式 1 中计算 ER2 可以利用两个资源，分为两个部分进行。一是依赖大规模语料库的基于概率统计的计算；一是依赖现有语言知识资源的基于事件相似度的计算。前者基于大规模语料的计算覆盖率较高，但准确率往往较低；后者基于知识库资源的计算方法准确率高，但覆盖率较低。综合考虑这两种度量方法，可以有效的融合这两种计算方法的优点。ER2 的具体计算公式如下：

$$ER_2 = \beta \cdot ER_{21} + (1 - \beta) \cdot ER_{22}$$（公式 3）

其中，ER_{21} 是基于语料库概率统计的语义关联度计算结果；ER_{22} 是基于语义知识资源的语义相似度计算结果。β 是权值，用于调整两种计算方法的贡献度，可根据实验效果进行调整。下面分别来看公式 3 中 ER_{21} 和 ER_{22} 的具体计算方法。

（一）基于语料库中 V1-V2 共现概率的语义关联度计算

先看 ER_{21} 的计算方法。一般地，A、B 两个事件若存在“致使-结果”语义关系，从时间顺序上来看，A 事件的发生要早于 B 事件。反映到语序上，则是代表 A 事件的谓词 V1 先于代表 B 事件的谓词 V2 出现。因此，对两个事件之间的“致使-结果”语义关系的计算，可以简化为 V1、V2 在实际语料中顺序出现时的相关度计算。这里，我们用 V1-V2 的点式互信息（PMI，pointwise mutual information）来估计二者所代表事件的相关程度。ER_{21} 的计算公式如下：

$$ER_{21} \equiv \log\frac{\#(v_1,v_2)/\#}{\frac{\#(v_1)}{\#}\cdot\frac{\#(v_2)}{\#}} = \log\frac{\#(v_1,v_2)\cdot\#}{\#(v_1)\cdot\#(v_2)}$$（公式 4）

其中，#(v1,v2)为两个谓词性成分在语料中前后共现于一个句子中的频次，#(v1)，#(v2)分别为两个谓词性成分在语料中出现的频次，#为语料总频次。

（二）基于词典知识库中词语相似性的语义关联度计算

再来看 ER_{22} 的计算方法。ER_{22} 是对任意两个 V1-V2 组合，计算其与已有的典型述结式的最大相似度。具体计算公式如下：

$$ER_{22} = \max_{w_1 \in \|V1\|, w_2 \in \|V2\|}\big(\mathrm{Sim}(\mathrm{w}_1, \mathrm{V1}), \mathrm{Sim}(\mathrm{w}_2, \mathrm{V2})\big)$$（公式 5）

其中：

1) $\|V1\|$ 是 DCRVC 中所有带 V2 补语的述语集合，对于 $\|V1\|$ 中的每一个词语 w1，

计算其与 V1 的词语相似度 Sim(w_1,V1)；

2) $\|V2\|$是 DCRVC 中所有给 V1 作补语的词语集合，对于$\|V2\|$中的每一个词语 w2，计算其与 V2 的词语相似度 Sim(w_2,V2)；

3) 最终结果为 Sim(w_1,V1)和 Sim(w_2,V2)中的最大值。

4) Sim(w_1,V1)和 Sim(w_2,V2)的计算则直接采用了刘群、李素建（2002）基于《知网》语义资源的词语相似度计算方法。

下面以“吃-懂”组合为例说明 ER_{22} 的计算步骤：

（1）在 DCRVC 中查找“懂”的述语集合，记作$\|V1\|$={看、读、……}；

（2）在 DCRVC 中查找“吃”的补语集合，记作$\|V2\|$={饱、光、遍、急、……}；

（3）计算“吃”跟$\|V1\|$中每个元素（w1）之间的相似度，计算“懂”跟$\|V2\|$集合中每个元素（w2）的相似度[4]。在所有相似度中取最大值作为结果输出。

$$ER_{22}(\text{吃，懂}) = Max(Sim(w1,\text{吃}),(Sim(w2,\text{懂})) = Max(1, 0.2424,...) = 1$$

显然，若 V1，V2 跟已有的典型述结式越相似，则 V1 和 V2 的语义关联度越高，二者越有可能构成述结式。如果 V1、V2 都是 DCRVC 中已有的述语词和补语词，且二者构成述结式，则 ER_{22}(V1,V2)的值为 1。如果 V1、V2 在 DVRVC 中均未出现，则规定 ER_{22}(V1,V2)的值为 0。

至此，V1-V2 复合事件的关联度计算得以落实。根据实验，上面公式 1 和公式 3 中的权值 α 和 β 分别取 0.1 和 0.7，所得 V1-V2 事件语义关联度计算结果在准确率（precision)和召回率（recall）两个指标上可以达到最优。因此，V1-V2 事件语义关联度计算公式最终可确定为：

$$\text{ER} = 0.1 * ER_1 + 0.9 * (0.7 * ER_{21} + 0.3 * ER_{22})\text{（公式 6）}$$

4. 述补结构词典的可视化应用

为了更直观地展示 DCRVC 数据库中的述补结构数据，为对外汉语教学与研究提供更好地计算机辅助，本文尝试借鉴数据可视化（data visualization）技术，在网页环境下为用户提供 DCRVC 数据库的查询以及文本中述补结构的自动识别服务。这一节介绍目前开发的原型系统（以下简称 VisualRVC[5]）的主要功能。

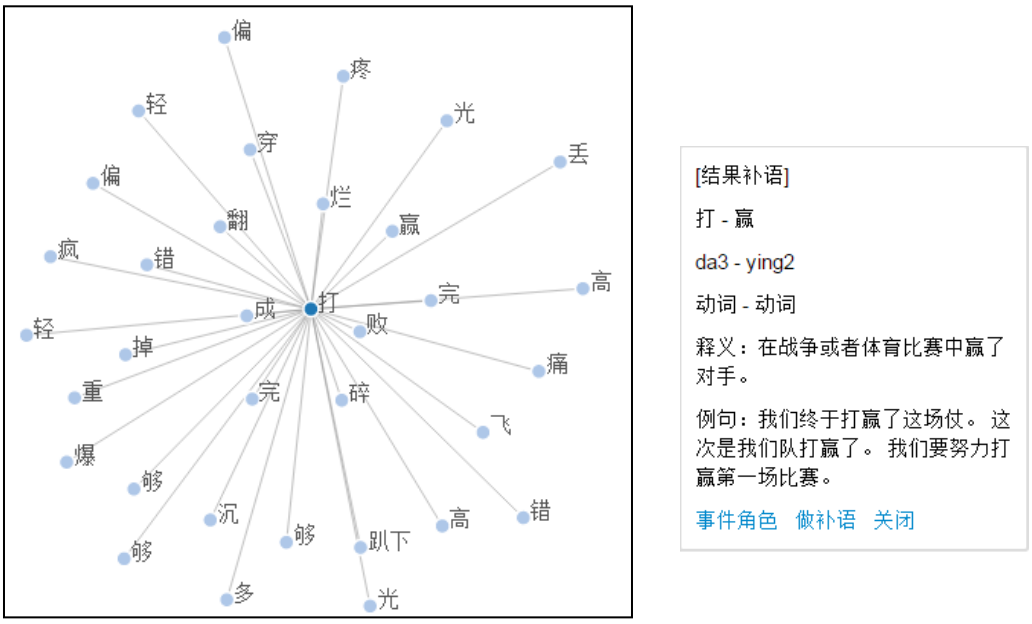
4.1. 查询述语和补语的相关信息

[4]基于 DCRVC 中的述语和补语条目，以及刘群、李素建（2002）基于《知网》数据库的相似度计算程序，“吃”跟“看”的相似度值为 1；“懂”跟“急”的相似度值为 0.2424。

[5]http://ccl.pku.edu.cn:8080/visualization

VisualRVC 实现的述补结构可视化页面布局以某一个述语（或补语）为中心，其所能搭配的全部补语（或述语）环绕四周。在这种展示布局下，我们用四周的节点到中心节点的距离来表示该述补结构的使用频率或聚合程度，使用频率或聚合程度越高的离中心节点越近。此外，四周节点的不同颜色可以区分补语类型的差异。通过这种展示方式，用户可以很直观地看到一个述语跟哪些补语结合的更紧密（组配凝固度高），跟哪些补语的关系较为松散（临时组配）。

除静态的环绕式布局展示外，用户用鼠标点击中心节点四周的某一节点时，会显示该节点相应的述补结构完整信息，包括补述语及补语的拼音、词性、补语类型、释义及例句等。如下面图 2 所示：



2-a　　2-b

图 2：述补结构基本信息可视化页面

图 2-a 是用户查询述语“打”所带的全部结果补语；图 2-b 是用鼠标悬停在补语“赢”节点上，然后点击鼠标左键后弹出的框图。

4.2. 查询述语和补语关联的事件角色

上文 3.1、3.2 两个小节讨论了从复合事件语义关联来分析述结式的基本框架。在 VisualRVC 系统中，也相应地提供了查询 V1-V2 各自作为单个事件的事件角色查询以及二者共有事件角色的展示。对单个事件的事件角色采用“文字云（Word Cloud）”的布局[6]进行展示。一个动词所关联的事件角色简单的定义为语料中跟该动词共现的名词性成分。下面图 3 是从实际语料中抽取的"打"的不同的事件角色[7]，

[6]“文字云”布局中具体词语的显示方式可以有多种，既可以按照常规的从左到右线性排列的方式，也可以有左右横排和上下竖排混合的模式，后者可以使“文字云”图有更富于动感的效果。

[7]这里的“打”并没有区分义项，因而实际上代表了很多不同的事件。

即语料句子中与“打”共现的名词。各个名词的出现频次多少对应到“画布”上的字号大小和颜色的不同。

图 3:“打”所代表事件的事件角色文字云图

而两个事件的共有事件角色（以“打”和“死”为例）的展示如下面图 4 所示，每个事件的事件角色放置在椭圆形布局内，共有事件角色位于两个椭圆的交汇处。

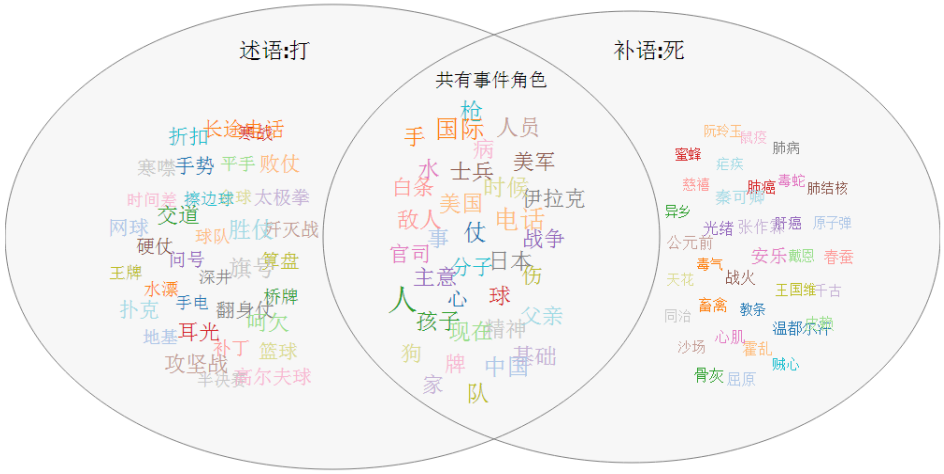


图 4:共有事件角色示意图

4.3. 文本中述结式的自动标识

利用第三节提出的事件语义关联度计算方法，VisualRVC 系统实现了一个从文本中自动抽取 V1-V2 述结式的功能模块。具体工作流程分为以下四个步骤：

1) 用户通过浏览器页面提交文本到服务器；
2) VisualRVC 系统对文本进行自动分词和词性标注，抽取其中的谓词性组合（V1-V2）；
3) 系统对抽取出的 V1-V2 进行事件语义关联度计算，按照设定的阈值，筛选出其中的述结式。
4) 对于被判定为述结式的 V1-V2 实例，鼠标点击后显示其事件语义关联度计

算结果及相关的述语、补语、事件角色等基本信息。

下面图 5 和图 6 是上述步骤的示意图：

图 5:“抽取述结式”结果界面示意图

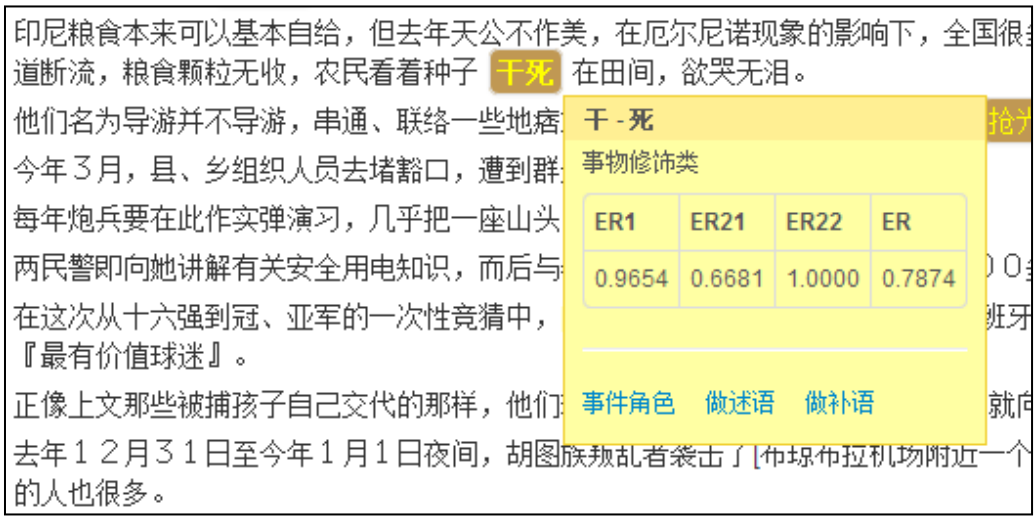


图 6:述结式具体信息及其他相关操作

图 5 中，程序对本文中自动识别出来的述结式做了高亮显示（highlight），并用不同颜色区分了述结式的不同小类。文本上方提供了一个平滑条，由用户自主设定阈值，可以综合考虑准确率和召回率的平衡。在这个页面上可以看到：滑动条上滑块位置变化引起阈值发生变化，文本中述结式的识别结果也相应发生变化，其中未达到阈值（关联度不够高）的述结式将不被高亮显示。

图 6 展示了当鼠标悬停在文本中某个述结式实例上，程序会弹出一个框图，显示该述结式的事件语义关联度计算结果，并可以链接到 DCRVC 数据库，进一步查询述语（V1）和补语（V2）的相信信息，也可以从语料库中查询各自所代表事件的事件角色（即如图 3、图 4 显示效果）。

5. 结语

本文介绍了构建现代汉语述补结构用法词典的背景及在线数据库的现况，探讨了在复合事件语义分析框架下，通过计算 V1-V2 的事件语义相关度，来判断 V1-V2 构成述结式的可能性（条件）。尽管基于本文提出的计算公式，对一定规模的 V1-V2 组合进行初步试验，结果表明这种计算方法具有一定的可行性（马腾，2014），但本文所提出的计算方法存在的问题也是比较明显的。除了公式 6 中各组成部分的细节可以再完善外，还应该更全面地考虑 V1-V2 组合的外部环境因素。以本文开头举的例 1 中比较极端的述结式例子“哭哭”来说，它的上文中有“把”“给”这类标志词，是可以提示这一环境中的 V1-V2 更倾向于解读为述结式的，但如果仅从事件语义关联度的角度来计算“哭”跟“哭”的关联，二者因为完全同形，无法代表两个本质上不同的事件（两个“哭”的参与角色是不同的），这时候用本文的计算方法，就有点缘木求鱼的味道了。当然，考虑的特征因素越多，计算的复杂性就越高。本文所提出的计算公式的优点是计算简单。在有了述补结构数据库以及事件语义相关度计算方法的基础上，本文利用数据可视化技术，开发出一个原型系统，来直观展示述补结构数据库的内容及在文本中自动标识 V1-V2 述结式。我们期望，对于汉语述结式的教学和研究，本文提出的这一思路以及初步的工作成果，是一次有意义的尝试。

参考文献

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*. *32*(1), 13-47.

Card, S., Mackinlay, J., & Schdiderman, B. (1999), Readings *in information visualization: using vision to think*. San Diego, CA: Morgan Kaufmann

Eick, S. G. (1994). Graphically displaying text. *Journal of Computational and Graphical Statistics*, *3*(2), 127-142

Eppler, M. J., & Burkhard, R. A. (2004). Knowledge visualization: towards a new discipline and its fields of application. *ICA-Working Paper #2/2004*, University of Lugano, Lugano, Switzerland.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. 1606-1611.

Guo, R. (1995). Valency and integration of resultative verb compound in Chinese. In Y. Shen, & D. Zhen (Eds.), *Studies on Chinese valency grammar* (pp.168-191).

Beijing, China: Peking University Press. [郭锐. (1995). 述结式的配价结构和成分的整合. 沈阳, & 郑定欧 (编). (1995). 现代汉语配价语法研究, 168-191. 北京, 中国: 北京大学出版社.)

Guo, R. (2002). Argument structure of resultative verb compound in Chinese. In L. Xu, & J. Shao (Eds.), *Proceedings of the First Conference on Grammar of Modern Chinese in the 21st Century* (pp.169-186). Hangzhou, China: Zhejiang Education Publishing House. [郭锐. (2002). 述结式的论元结构. 徐烈炯, & 邵敬敏(编). 汉语语法研究的新拓展（一）——21 世纪首届现代汉语国际研讨会论文集 (pp. 169-186). 杭州, 中国: 浙江教育出版社.]

Huang, J. (1993). Logical structure and syntactic features of V-R predicates. *Linguistic Research*, *2*, 57-62. [黄锦章. (1993). 行为类可能式 V-R 谓语句的逻辑结构与表层句法现象. 语文研究, 2, 57-62.]

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING X, International Conference Research on Computational Linguistics* (pp. 19-33).

Li, Y. (1990). On V-V compounds in Chinese. *Natural Language and Linguistic Theory, 8*(2), 177-207.

Liu, Q., & Li, S. (2002). Word Similarity Computing Based on How-net. *Computational Linguistics and Chinese Language Processing*, *7*(2), 59-76. [刘群, &李素建. (2002). 基于《知网》的词汇语义相似度计算. 中文计算语言学期刊, *7*(2), 59-76.]

Liu, Y. (Ed.). (1998). *A Dictionary of Directional Verb Complements*. Beijing, China: Beijing Language and Culture University Press. [刘月华(主编). (1998). 趋向补语通释. 北京, 中国: 北京语言大学出版社.]

Ma, T. (2014). *A study of resultative verb compound in modern Chinese based on event semantics* (Unpublished master's thesis). Peking University, Beijing, China. [马腾. (2014). 基于事件语义学的现代汉语述结式判别研究(硕士学位论文). 北京大学, 北京, 中国.]

Resnik, P., & Diab, M. (2000). Measuring verb similarity, In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, 399-404.

Sahami, M., & Heilman, T. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, 377-386.

Sunaoka, K. (2013). Design of an online database for Chinese resultative verb compounds, In *Proceedings of The 63rd Annual Conference of The Chinese Linguistic Society of Japan* (pp. 23-32). [砂岡和子. (2013). 关于汉语动词补语用法的网上词典的设计. 2013 年日本中国語学会第 63 回全国大会论文集(pp. 23-32.]

Shi, C. (2008). *Syntax and semantics of resultative verb compound in Chinese*. Beijing, China: Language and Culture University Press. [施春宏. (2008). 汉语动结式的句法语义研究. 北京, 中国: 北京语言大学出版社.]

Song, W. (2007). *A Cognitive Approach on Resultative Verb Compound in Chinese*, Beijing, China: Peking University Press. [宋文辉. (2007). 现代汉语动结式的认知研究. 北京, 中国: 北京大学出版社.]

Tenny, C., & Pustejovsky, J. (2002). A History of events in linguistic theory. In. C. Tenny, & J. Pustejovsky (Eds.), *Events as grammatical objects: The converging perspectives of lexical semanticslogical semantics and syntax* (pp.3-37). Stanford, CA: CSLI Publications.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press:

Tufte. E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Yang, D., & Powers, D. M. W. (2006). Verb similarity on the taxonomy of WordNet. In *Proceedings of GWC-06* (pp. 121-128).

Yuan, Y. (2001). An analysis of control and restore on valency of resultative verb compound in Chinese. *Chinese Linguistics*, *5*, 399-479. [袁毓林. (2001). 述结式配价的控制——还原分析. 中国语文. *5*, 399-479.]

Wang, H. (1995). A Study of valency of resultative verb compound in Chinese, In Y. Shen, & D. Zheng (Eds.), *Studies on Chinese valency grammar* (pp.144-167). Beijing, China: Peking University Press. [王红旗, (1995). 动结式述补结构配价研究. 沈阳, &郑定欧(编), 现代汉语配价语法研究(pp.144-167). 北京, 中国: 北京大学出版社.]

Wang, Y., Jiao, Q., & Pang, Y. (Eds.). (1987). *A collocation dictionary of Chinese resultative verb compounds.* Beijing, China: Beijing Language and Culture University Press. [王砚农, 焦群, 庞颙(编). (1987). 汉语动词-结果补语搭配词典. 北京: 中国: 北京语言学院出版社.]

Zhan, W. (2013). Combinational conditions of Chinese resultative verb compounds: A perspective from conceptual structure of composite events. *Research on Chinese as a Second Language*, *9*. 111-141. [詹卫东. (2013). 复合事件的语义结构与现代汉语述结式的成立条件分析. 对外汉语研究, *9*, 111-141.]

Zhan, W., & Ma, T. (2013). *Calculation of semantic correlation between composite events and combinational conditions of Chinese resultative verb compounds*, Paper presented in the 25th Annual North American Conference on Chinese Linguistics (NACCL–25), June 21-23,2013. University of Michigan, Ann Arbor, USA. [詹卫东, 马腾. (2013). 从复合事件语义相关度看现代汉语述结式的成立条件. 第 25 届北美汉语语言学年会, 2013.6.21-23，密西根大学, 密西根，美国.]

