中文计算技术与语言问题研究

Chinese Computing Technologies and Related Linguistic Issues

——第七届中文信息处理国际会议论文集

Proceedings of the 7th International Conference on Chinese Computing

◆ 主编：萧国政 何炎祥 孙茂松

◆ 执行主编： 姬东鸿 刘礼堂

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

中文计算技术与语言问题研究

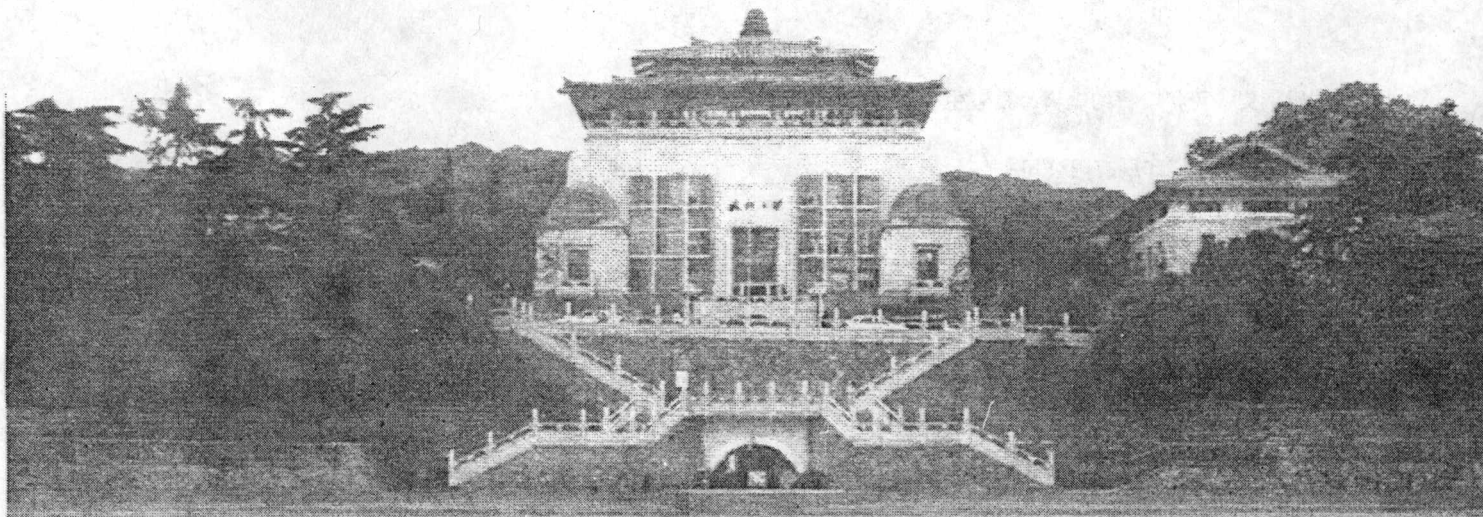
Chinese Computing Technologies and Related Linguistic Issues

——第七届中文信息处理国际会议论文集

Proceedings of the 7th International Conference on Chinese Computing

◆ 主编：萧国政 何炎祥 孙茂松

◆ 执行主编：姬东鸿 刘礼堂



電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

本书是武汉大学召开的第七届中文信息处理国际会议（ICCC2007）的论文集。共收录中外作者论文136篇。分为七个系列：① 词法、句法、语义；② 语言资源建设及相关技术；③ 智能检索（信息检索、抽取、文本分类、自动文摘等）；④ 机器翻译；⑤ 语音学与语音处理；⑥ 语言学研究；⑦ 其他。

本书是当前中文信息处理研究成果的一个缩影，展示了国内外中文信息处理及其应用研究的最新进展和发展动向，对广大中文信息处理理论研究者和相关产品、技术的研发人员具有重要的参考价值。本书可供计算机专业、语言学专业、对外汉语专业等领域的科研人员、工程技术人员和高校教师、研究生参考选用。

图书在版编目（CIP）数据

中文计算技术与语言问题研究：第七届中文信息处理国际会议论文集/萧国政，何炎祥，孙茂松主编．—北京：电子工业出版社，2007.9
ISBN 978-7-121-02563-1

Ⅰ．中… Ⅱ．①萧…②何…③孙… Ⅲ．汉字信息处理—国际学术会议—文集 Ⅳ．TP391.12-53

中国版本图书馆 CIP 数据核字（2007）第 149522 号

责任编辑：史 涛 特约编辑：葛春生
印 刷：
装 订：北京季蜂印刷有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编 100036
开 本：880×1 230 1/16 印张：50 字数：1520 千字
印 次：2007 年 9 月第 1 次印刷
定 价：160.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

第七届中文信息处理国际会议（ICCC2007）

会议日期：2007 年 10 月 13～15 日
会议地点：武汉大学

主办单位：中国中文信息学会
新加坡中文与东方语言信息处理学会
武汉大学语言与信息研究中心
承办单位：武汉大学语言与信息研究中心
赞助单位：国家语言资源监测与研究中心网络媒体分中心
富士通研究开发中心有限公司
湖北省语言学会
清华大学计算机系智能技术与系统国家重点实验室
江汉大学人文学院语言学及应用语言学学科
湖北师范学院汉语言文字学重点学科

大会名誉主席：刘经南　曹右琦
大 会 主 席 ：李宇明　李生　赖金锭　萧国政

（以下按音序排列，为便于排版，中文在前，西文在后）

程序委员会
主　　席：何炎祥　　孙茂松
执行主席：姬东鸿　　于　浩
委　　员：陈群秀　冯志伟　付国洪　傅爱平　关　毅　胡　悌
黄河燕　黄萱菁　靳光瑾　亢世勇　李海洲　李涓子
李文捷　李耀勇　刘德喜　刘　群　刘　挺　马　斌
马　青　孟　遥　齐　冲　史晓东　宋　柔　穗志芳
孙　乐　王海峰　王　惠　徐　杰　杨尔弘　袁毓林
詹卫东　张桂平　张　民　张　全　赵　军　赵铁军
周国栋　周　强　朱靖波　宗成庆
Choi Key-Sun　　Jean-Pierre Chevallet
Kumiko Tanaka-Ishii　　Olivia Kwong

组织委员会
主　　席：吴泓渺
执行主席：刘礼堂
委　　员：陈国恩　陈　敏　陈月屏　段曹林　董明会　杜青钢　郭婷婷　胡勇华　李向农
刘会胜　卢烈红　卢江滨　沈壮海　涂险峰　翟　汛　赵世举　周建民

前言

2007 年的秋天我们迎来了又一个丰收的季节。10 月的武汉，秋高云淡，桂子飘香，气候宜人。来自海内外的专家学者济济一堂欢聚在山清水秀的珞珈山畔，东湖之滨，参加第七届中文信息处理国际会议，交流在中文信息处理领域取得的最新成果，展望中文信息处理的发展远景，研讨中文信息处理新的发展模式和进军方策。

本次学术会议共征集到海内外论文 285 篇，经大会程序委员会严肃认真的多轮匿名评审，最终录用论文 136 篇，汇编成集，由电子工业出版社正式出版。这些论文的作者，既有学界知名的资深学者，也不乏初出茅庐的后起新秀。论文分为七个部分：

1）词法、句法、语义，43 篇；

2）语言资源建设及相关技术，11 篇；

3）智能检索（信息检索、抽取、文本分类、自动文摘等），41 篇；

4）机器翻译，9 篇；

5）语音学与语音处理，13 篇；

6）语言学研究，12 篇；

7）其他，7 篇。

论文数量的构成比例，基本上反映了学者们当前关注的热点和中文信息处理未来的发展方向。第 1）、3）部分论文数量占绝对优势。词法、句法和语义研究作为语言信息处理的基础研究，一直方兴未艾。近年来随着语言信息处理技术和成果应用的不断推广和深入，大规模自然文本理解的社会需求日益凸显，作为后起之秀，基于内容的文本处理迅速崛起，成了语言信息处理领域的中流砥柱。书中这一部分的论文数量与作为基础研究的词法、句法、语义研究部分不相仲伯。机器翻译和语音处理是计算语言学领域传统的分支，一直有学者们在此领域默默耕耘、精益求精，并且将研究视野扩展到了蒙、维、藏等少数民族语言，取得了可喜的成绩。延续上届 ICCC 的传统，本届会议中适当选收了少量与信息处理息息相关的语言本体研究的成果，体现了传统语言学界对中文信息处理的热情关注。此外有个别文章涉及文字、篇章处理和语言教学等技术层面，一并归入“其他”类。相对而言，语言资源建设作为语言信息处理的重要课题，没有得到足够的关注。资源建设是语言信息处理的立根之本，这一部分的工作往往需要耗费巨大的人力物力，我们期待着更多的学者来关注、投入到这一领域。我们也期望中文信息处理继成功完成汉语自动分词、词性（词类）自动标注的学术和技术攻坚之后，在词义自动标注的进军中有更多建树。

中文信息处理国际会议（ICCC）是一个具有 20 年历史和颇高学术水准的连续性的国际学术会议，最早起于 1986 年，代表着中文信息处理的国际学术前沿与动态，在中文信息处理学界享有盛誉。此前，会议一直在新加坡召开，本次会议第一次在新加坡以外的国家召开。

第七届中文信息处理国际会议将于 2007 年 10 月 13～15 日在武汉大学如期召开。为了开好这次会议，在海内外中文信息处理领域的专家学者和各级领导的大力支持下，中国中文信息学会、武汉大学语言与信息研究中心和新加坡中文与东方语言信息处理学会经过了两年半的精心筹备和不懈努力，藉此，我们谨向给予本次会议以支持和帮助的所有领导、专家学者，慷慨解囊的单位，支持论文集出版的电子工业出版社和编辑，为这次会议做了许多具体工作的武汉大学的老师和同学，一并致以最衷心的感谢。

目 录

第 1 编　词法、句法、语义

第 2 编 语言资源建设及相关技术

第 3 编 智能检索（信息检索、抽取、文本分类、自动文摘等）

第 4 编 机 器 翻 译

第5编 语音学与语音处理

第6编 语言学研究

第7编 其 他

第1编

词法、句法、语义

动词“打”本义的结构描写及其同义词群建构
——一种人机共享的“词群－词位变体”研究初探

萧国政

武汉大学语言与信息研究中心　武汉大学文学院　武汉　430072

摘　要：本文认为词义自动标注是语言处理新的奋斗目标，而实现这一目标的首要任务是同义词群的建构。同义词群建构面临词义科学描写与词群构成模式两大难题。本文通过汉语动词“打”的本义（打 $_1$）词义的分离、描述，介绍了在特征性义素分析基础上，对多义词词义进行“语法－语义”分析的方法；通过建立打 $_1$ 语义词位，描述打 $_1$ 的上位变体、下位变体、同位变体、邻位变体及其同义词群系统，展示了新的同义词群构成模式理论。

关键词：词义自动标注；动词；同义词群构建；“词群－词位变体”理论

The Description of the Lexical Meaning Structure of the Verb “da” and the Construction of its Synset —— A Study on “Synset-Lexeme Anamorphosis” Shared by Human and Computer

Xiao Guozheng

Center for the Study of Language & Information / School of Chinese Language and Literature, Wuhan University

Abstract: The paper considers the automated labeling of acceptation is the new aim of NLP, and the chief task of realizing this aim is the construction of the synonymy set, which is faced with two big problems: the scientific description of word meaning and the construction mode of synonymy set. By separating and describing the original meanings of the Chinese verb “da”, this paper introduces the “grammar-semantics” method of acceptation analysis of polysemic words based on the characteristic sense analysis. It sets up the semantic lexeme of “da$_1$”, and describes its superordinate anamorphosis, hyponymy anamorphosis, compeer anamorphosis and adjacent anamorphosis, and its synonymy set system. Meanwhile, it lays out the new theory of the construction mode of synonymy set.

Keywords: Automated Label of Word Meaning; Verb; Construction of Synset; “Synset-Lexeme Anamorphosis” Method

WordNet 问世及其研究发展，使越来越多的人认识到，一个大的词库对自然语言理解和人工智能各方面研究的重要价值，以词为基点研究和表述语言知识，按词语的义项建构语言的若干同义集合——同义词词群（synset），进而描述词语之间、词群之间各种语义、语法关系，这种研究不仅能显示出一片一片解决语义问题的便捷和高效，而且能发现其他对象研究方式无法发现的问题。

同义词词群是围绕同一词义形成的若干词的集合，同义词词群的建构是一项巨大的语言工程，是语言信

息处理、人工智能，尤其是机器翻译进一步发展所面临的新的时代性任务，是自然语言处理和人工智能继词语的自动切分、词性自动标注之后，实现新的奋斗目标——词义自动标注的攻坚性工程。并且如果这个工程完成不好，不管计算机有多先进，计算技术和算法有多高超，其努力都只能是在沙滩上构建高楼大厦。

汉语同义词群的建构，没有什么捷径可走，需要在借鉴其他语言研究有关理论、经验和技术以及得失的基础上，从汉语的实际出发，老老实实地一个一个同义词词群地建构。

建构同义词群，词义的描述和刻画非常关键，它决定着所建词群的客观性、合理性以及智能性。如果把语言研究简单地二分为传统的基础研究和面向语言信息处理的工程研究的话，那么面向工程研究的词义刻画，是不能简单地搬用或化用字词典的研究成果的，更不能依赖它。因为很多字词典的词义描写是就不同词分别考虑的，不仅系统性不足，义项分割具有极大的随意性，而且精确性和一致性也比较差。

不少做过同义词群或词网研究的同行，大多都有过这样一种困惑：汉语是我们的母语，自己的学历基本都在大专及以上，一个汉语常用词有几个意义，每个意义是什么，在没查词典之前一般都觉得是清楚的。但是在查了词典之后，尤其是比较了几个义项之后，就觉得不怎么清楚了，当进而查了几本词典进行系统整理之后，就感到无可适从了。这不是故弄悬殊和耸人听闻，请看具体例子。

比如动词“打”是我们日常生活中使用得相当多的词，或者说是现代汉语里使用平度相当高的词，在《现代汉语频率词典》（北京语言学院出版社，1986）的最高使用频度的8000词中排名第94，在使用频度最高几个动作动词中位居第5。[1]这里我们看看三部字词典对汉语动词“打”多义义项、义项分割与词义的刻画。为节省时间和篇幅，只截取其前5个义项。

（1）《现代汉语词典》（2005）列了动词“打”24个义项，前5个是：（～代表“打”，下同。）

①用手或器具撞击物体：～门｜～鼓；②器皿、蛋类等因撞击而破碎：碗～了｜鸡飞蛋～；③殴打；攻打：～架｜～援；④发生与人交涉的行为：～官司｜～交道；⑤建造；建筑：～坝｜～墙。

（2）《应用汉语词典》（2005）列了动词“打”的27个义项，前5个是：

①敲打；敲击：～门｜～鼓｜～碎；②因撞击而破碎：～了个花瓶｜挺好的一个碗给～了｜鸡飞蛋～；③殴打；攻打：你怎么～人了｜两个人～了起来｜这一仗要～好；④人际之间的交涉：～官司｜～交道；⑤建造：～地基｜～一道墙。

（3）《国际标准汉语大字典》列了动词“打”的19个义项，前5个是：

①击，敲，攻击：～击｜殴～｜～杀；②放出，发出，注入，扎入：～炮｜～雷｜～信号｜～电报；③做，造：～首饰｜～家具；④拨动：～算盘；⑤揭，破，凿开：～破｜～井。[2]

（1）（2）两部词典都是商务印书馆出版的，义项分割基本一致，第三部词典则不然。就是从具有一致性的词典看，这里起码有两个问题：①同是“打门”的“打”，《现代汉语词典》释义用的关键词是“撞击”，若把词典释义与所举例子联系起来看，《现代汉语词典》描述的现象非常粗暴罕见。因为根据其释义，撞门的工具有二，一是手，二是其他器具。若是用手撞门，只能用胳膊肘，这种打门方式不是没有，十分罕见；若是用器具撞击，那器具肯定不小，其行为几近救火、救人或入室抢劫。将其词义解释用于“打鼓”的“打”，用手撞，用器具撞，一般不常见。《应用汉语词典》用的关键词是“敲打”“敲击”，与其所举的前两例联系起来，解释打鼓没有问题，但解释打门只是偏指了一种方式。第三例也是同样的问题，敲碎可以说打碎，但是“打碎”不一定只指敲碎。②在释义方式上，两部词典一用句子刻画，一用同义词语注释，但都涉嫌用下位概念注释上位概念。不过这绝不是这些词典的编者不懂逻辑，应是不得已而为之，因为像这两部词典这样分割多义词“打”的词义义项，谁也无法把其第一义项描述清楚，再高级的专家也难以幸免。因此，多义词词义的正确认识和合理分割，是准确描述词义的前提和基础。

“打”是多义动词，多义动词是同义词群建构的重点和难点。不论是义项的分割，还是不同意义的确定和描述，都不是一件很容易的事。那么我们能否找到一方式，既能系统简明地描述词不同意义之间的关联，又便于研究词群的若干人像一个人一样能保持工作标准的一致性？研究发现，“词群－词位变体”研究是达

1 其前4个动作动词依次是：来、走、到、想。

2 此词典资料摘自《金山词霸》2005。

到这个目标可以一试的方法。这个方法不仅可用来为计算机建立同义词群，也可以用于从事对外汉语教学，是一种人机共享的研究模式。这个研究模式的基本内容是：在对词义构成进行义素分析的基础上，用语法一语义结构模式来描述和揭示其语义结构，显示一义跟他义的语法语义对立及联系，按词义一个一个词义地建立词的语义“词位”（类同音位），再根据词位和词义鉴别式确立其同义词，通过变体理论建立起其同位次词群、下位次词群以及邻位次词群，最终形成“打”某个意义的多层次多侧面的立体词群系统。

限于时间和篇幅，本文仅报告动词“打”本义（记作：打$_{1[击]}$或打$_{1}$）的词义结构描述和同义词群建构的基本思路和初步成果，以向各位专家学者求教。

1 “打”的本义及同义词群的基本构成

“打”是动词，表动作，《说文》曰“打，击也，从手,丁声。”如果采用词典注释的形式描述词义，“打$_{1}$”一方面可仿《说文》用“击”描述其义，另一方面也可如《现代汉语词典》那样用描述性语言刻画其词义，但其表述应修改为：用手等肢体或同时使用器物击向某人某物。同时在词典所举之例外，再补充如下（包括人和动物）语例，如：～手｜别～孩子｜别～我的小狗等。

为了研究的操作方便，我们需要揭示一个词词义的语义构成。从语义构成看，“打$_{1}$”涉及施动者、受动者（动作对象）和工具“手”（或方式——用手），并且［＋施动］［＋动作］［＋用手］［＋所击对象］是该词义不能或缺的因素。但是在“打”的内部，“打”的本义“打$_{1}$”是表施动者自控性的行为或动作（［＋自控性］）。当“打$_{1}$”与指人的对象组合起来形成“‘打’＋指人词语”的结构时，就表达一种具蓄意性的行为。当有些动词（如撞、推等）的所击对象为“人”其行为性质具有多可性时，[3]只有具有故意性特征 V+NP 才能给以“打人”的定性。可比较：

（4）后面有人推他，他站不住，不幸把小山子撞倒，还踩了一脚，这不是打人吗？（×）

（5）后面是有人推他，但他借势把小山子撞倒，还踩了一脚，这不是打人吗？（√）

因此，下面三个义素是“打$_{1}$”的必备要素（或特征性义素）：①动作性质——施动者自控性的行为或动作（［＋自控性］）；②动作条件——使用手或身体其他部件（［＋手］）；③动作目的——击中某人某物（［＋对象］）。

词义是词长期应用的结果。多义词的不同词义是该词应用于不同的对象和环境的历史沉淀。并且词性（词类）不同，词义的组成结构、沉淀方式与激活形式也不相同。打$_{1[击]}$是二价动词，其语义构成可用语法一语义结构来描述为：

（6）××自控性地用手及身体其他部分（或同时使用器具）击某人某物，且某人某物为动词的语法宾语。

（7）不仅是词义结构的综合描述，而且是“打$_{1}$”词义及同义词的鉴别式。所谓鉴别式就是说，只有符合该式条件的“打”才是“打$_{1}$”及同义词。在操作上，“打$_{1}$”同义词的鉴别还要遵循两个原则：一是足量原则，二是等量原则。

所谓“足量原则”是指“打$_{1}$”同义词必须满足（6）的全部条件。具体讲就是，看一个动词是不是“打$_{1[击]}$”的同义词，首先要看该用法的“打”是不是“击”的意思，其次看其宾语是不是“打”的对象，第三要看手及身体部件的参与作用。比如“打墙”一语有两个意思：①装修之前的砸墙；②建房子时的筑墙。不同意义上“打”的同义词分别是“砸”和“筑”。根据语义鉴别式（6），其第一个意思的“打”是击的意思，“墙”是动作所“击”的对象，故其“打”是该词的本义用法，“砸”是“打$_{1}$”的同义词。而第二个意思的“打”也有击的动作，筑墙也要用手或再加上工具击打泥土什么的，但是泥土不是“打”的语法宾语，不满足鉴别式的要求，故第二个意义上的“打”的同义词“筑”不是“打$_{1}$”同义词。

所谓“等量原则”是指由“打”构成的动宾结构，意义上不能增加多于（6）的语义内容或出现第三论元（（6）只有施事和受事两论元）。比如“打电话”的“打”完全符合语义足量原则，但是不符合语义等量

3 “撞”是“打”本义的变体，参看2.2。

原则。因为“电话”这里是指代电话机，打电话的意思并不是止于动宾结构的意思——敲击电话机的按键，而是通过敲击电话机的按键拨通电话，进而通话。故“打电话”的“打”负载的语义不符合等量原则的要求，故其“打”的同义词“拨”不是“打$_1$”的同义词。[4]

相反，像“踢”、“杀”符合同义鉴别的足量原则和等量原则，是“打$_1$”义素$_2$、义素$_3$的不同表现而形成的下位同义词。（参看第二、三、四节）在“打$_1$”的三个基本义素中，因义素$_2$、义素$_3$的各种情况及变化，形成了“打$_1$”的两大类同义变体——同义词：“打$_{12)x}$”和“打$_{13)x}$”。[5]

一个“打”的词义是一个义位，“打$_1$”这个义位可仿音位用方括号或斜线表示，记作：“/打$_1$/”或“打$_{/1[击]/}$”。为建构同义词群和说明问题，其义位变体分为4类：

1）同位变体，记作：打$_{10)x}$。如“击”就是“打$_1$”同一层级上的同义词，“打$_1$”的同位书面变体。比较：打头部——击头部｜打落——击落。“击”的词位系统编号“打$_{10)1}$”，“0)”表示跟其义素构成无关。

2）邻位变体，记作：$_{Lx}$打$_1$。如近义词“打架”、“打斗”、“斗殴”等，就是/打$_1$/ 词位同一层级上的邻居，为方便检索，可表述为邻位变体。其词位系统编号可分别为“$_{L11}$打$_1$”、“$_{L12a}$打$_1$”、“$_{L12b}$打$_1$”。“打$_1$”是着意表单方出击的动词，而“打架”则表双方出手的意思。比如张三打李四，你说李四打架他肯定不服，只有当李四还手了，继续打才是打架。而“撕打”“扭打”则是“打架”的下位变体，跟“打$_1$”不一个系列。

3）下位变体，记作“打$_{1m)x}$”，m是基本义素的标号，且m>0。如“踢”“杀”就是“打$_1$”第二、三两个基本义素的变体——同义词，因此其词群系统编号分别是“打$_{12)22}$”和“打$_{13)31}$”。“打$_1$”同义词群的建构主要是其下位同义变体的发掘、描述及系统整理。

4）上位变体，记作“打^{1x}”。理论上，某个词义上的词是有上位变体的，应留下理论和操作接口，但是在实际操作中易采取上位自动形成的策略，因为你把一种语言所有词的下位都描述完了，其上位在电脑里就自然生成了。

不同类型的变体构成的集合，都是围绕某个词义建立的同义词群。其最大的词群（或基点词群）用大括号标示，与之具有变体关系的其他词群用中括号标示。因此/打$_1$/的同义词群就是：{/打$_1$/：打$_1$+[打$_{1同位变体}$]+[打$_{1下位变体}$]+[打$_{1邻位变体}$]+[打$_{1上位变体}$]}，这里[打$_{1上位变体}$]暂不考虑，故/打$_1$/的同义词群可描写为：{/打$_1$/：打$_1$+[打$_{10x}$]+[打$_{1m)x}$]+[$_{Lx}$打$_1$]}。以下分别描述/打$_1$/下位词群成员及整理/打$_1$/词群系统。

2 “打$_1$”的义素2变体：打$_{12)X}$

“打$_1$”是一个二价动词，在其受事论元（击中的目标）不变的前提下，施行动作和行为的工具“手”的形式和接触目标的部位、方式或及替代物（身体的其他部分）不同，“打$_1$”有不同的词语变体“打$_{12)x}$”。“打$_{12)x}$”的特点是除“打”加名词性宾语的意义是击向某个目标的意义外，还具有固化在其同一变体动词中的“方式义”。即：打+NP_o=击NP+击的方式 。因手及手的替代物不同，“打$_{12)x}$”分为两小类：自身变体“打$_{12)1}$”和换身变体“打$_{12)2}$”，并且就目前来看，这两类变体都是成员可穷尽列举的封闭的类。[6]

（1）“打$_1$”的义素$_2$因“手”击目标的样式、部位或方法不同，“打$_1$”有6个常见的第2）要素的自身变体同义词：

1）使用手掌击人面部等的“打$_{12)11a}$”：掴（如：～了他两巴掌）[7]

2）使用手掌击物某个面的“打$_{12)11b}$”：拍（如：～桌子|有人～门）

---

4 “打电话”的“打”的语法语义结构可描述为：××为达到某目的，用手（或同时使用器具）击某物，且某目的为动词的语法宾语。

5 脚标除前面的“1”是词义的代号外，其余的数字及符号标志该词的词位系统位置。其中“2)”表示该变体跟“打$_1$”的第2个义素有关，“3)”类推。X表此词是该类变体的第几个词，即次类成员序号。

6 由于时间关系，本文穷尽与否尚不知道。

7 为醒目和简便，例中“打”以“～”代之。或者说，举例中的“～”读“打”。一般可用括号前面的词替换。下同。