现代汉语语素系统的研发与应用

刘扬

北京大学计算语言学研究所

中文系应用语言学报告 @ 2024/4/18

报告提纲

- **1.** 回望汉语世界中的字词关系
- **2.** 汉语的语素概念表示与语义构词分析
- **3.** 汉语语素系统的应用与展望

报告提纲

- 1. 回望汉语世界中的字词关系
- 2. 汉语的语素概念表示与语义构词分析
- 3. 汉语语素系统的应用与展望

1. 回望汉语世界中的字词关系

- 关于汉语语言事实与特点的讨论
  - 汉语中的形态、句法规则较弱，属于意合型语言
  - 在语言单位形成上，“由字组词、由词造句”大体了遵循同一原则
  - **汉语的字是清晰独立的表义单位**，而词的界限并不分明
  - 在词汇化历程中，**字作为不同的语素出现，其语素义相对稳定**
  - 因此，虽然“词无定形”，但是字和词之间具有很强的推导性
- 汉语语素与构词分析具有突出价值
  - 遵循“从结构形式到意义”的分析方法，并达成对词的理解、计算
- 汉语语素与构词分析相关资源建设
  - 苑春法：汉语语素数据库
  - 周亚民：汉字知识本体（Hantology）
  - 董振东：知网（HowNet）

报告提纲

- **1.** 回望汉语世界中的字词关系
- **2.** 汉语的语素概念表示与语义构词分析
- **3.** 汉语语素系统的应用与展望

2. 汉语的语素概念表示与语义构词分析

- 以《现代汉语词典》刻画的汉语语素及语素义为依据
  - 语素是语言中最小的语法单位，也就是最小的语音、语义结合体

选(選)xuǎn ①动挑选：筛～|～拔|～派|～种。②动选举：～民|普～|～代表。③被选中了的(人或物)：入～|人～。④挑选出来编在一起的作品：文～|诗～|民歌～。

挑 [1]tiāo 动①挑选：～心爱的买。②挑剔：～毛病。

挑 [2]tiāo ①动扁担等两头挂上东西，用肩膀支起来搬运：～担|～水|～着两筐土。②(～儿)名挑子：挑～儿。③(～儿)量用于成挑儿的东西：一～儿白菜。

另见 1354 页 tiǎo。

材 cái ①木料，泛指材料①：木～|钢～|药～|就地取～。②名棺材：寿～|一口～。③资料：教～|题～|素～。④有才能的人。⑤(cái)名姓。

才 [1]cái①名才能：德～兼备|多～多艺|这人很有～。②有才能的人：干～|奇～。③(cái)名姓。

才 [2](纔)cái 副①表示以前不久：你怎么～来就要走?②表示事情发生得晚或结束得晚：他说星期三动身，到星期五～走|大风到晚上～住了。③表示只有在某种条件下然后怎样(前面常常用"只有、必须"或含有这类意思)：只有依靠群众，～能把工作做好。④表示发生新情况，本来并不如此：经他解释之后，我～明白是怎么回事。⑤表示数量小，次数少，能力差，程度低等等：这个工厂开办时～几十个工人|别人一天干的活儿，他三天～干完。⑥表示强调所说的事(句尾常用"呢"字)：麦子长得～好呢|我～不信呢!

2. 汉语的语素概念表示与语义构词分析

- 汉语的语素概念表示工程
  - 工程目标 1：　确认“**语素义编码**”
    - 收集汉字的各个语素义，明确并补齐其所属的语素类（POS），并赋予其唯一的“语素义编码”
  - 工程目标 2：　提取“**语素概念**”
    - 对语素的释文句子，借助语义相似度计算，形成各个“同义（或同类）语素集”，用以表征“语素概念”，或称“语义基元”
  - 工程目标 3：　构建“**语素概念体系**”
    - 对现代汉语的全部“语素概念”，构建“语素概念体系”
  - 工程目标 4：　建立“**语义构词描述**”
    - 通过对构词结构标注及其下的“语素概念”绑定，建立基于“语素概念体系”的“语义构词描述”

2. 汉语的语素概念表示与语义构词分析

- 当前工程进展：目标 *1* 的阶段性数据成果
  - 获得 8514个汉字（TYPE）14个语素类（POS）下的 20855个语素义（SENSE），并按规则给每个语素义赋予唯一的“语素义编码”
  - 这其中，成词的自由语素（free morpheme）和不成词的黏着语素（bound morpheme）大致各占一半
  - “选”字的各个“**语素义编码**”：

    选1_04_01:　动词　　挑选：筛～｜～拔｜～派｜～种
    选1_04_02:　动词　　选举：～民｜普～｜～代表
    选1_04_03:　名语素　被选中了的(人或物)：入～｜人～
    选1_04_04:　名语素　挑选出来编在一起的作品：文～｜诗～｜民歌～
  - “材”字的各个“**语素义编码**”：

    材1_05_01:　名语素　木料，泛指材料：木～｜钢～｜药～｜就地取～
    材1_05_02:　名语　　棺材：寿～｜一口～
    材1_05_03:　名语素　资料：教～｜题～｜素～
    材1_05_04:　名语素　有才能的人
    材1_05_05:　名语　　(Cái)姓

2. 汉语的语素概念表示与语义构词分析

- 当前工程进展：目标 2 的阶段性数据成果
  - 作为主体的汉语的名、动、形语素分别聚合形成了2018、1630、549个相对独立、封闭的“**语素概念**”
    - 这构成了汉语世界中操纵计算的基本意义单元（即“**语义基元**”）
    - 数据成果具备优良特性：明确、可判的覆盖度与准确性
    - 形成了用于表征、操作汉语词型、词义的新的数据基础
  - 动语素“**语素概念**”举例，某概念 X：“选择、挑选”
    - {刷3_01_01, 抡1_01_01, 拔1_08_03, 拣1_01_01, 择1_02_01,
    - 择2_02_01, 挑1_02_01, 擢1_02_02, 调4_02_02, 选1_04_01,
    - 遴1_01_01, 铨1_02_01}
  - 名语素“**语素概念**”举例，某概念 Y：“有才能的人”
    - {匠1_02_02, 哲1_02_02, 器1_05_04, 尖1_09_06, 彦1_02_01,
    - 才1_03_02, 材1_05_04, 杰1_03_01, 模1_04_03, 氏1_05_03,
    - 秀2_04_04, 英1_03_02, 豪1_04_01, 贤1_04_02, 通2_12_07,
    - 骥1_02_02}

2. 汉语的语素概念表示与语义构词分析

名语素“语素概念”覆盖、分布情况

集合字数	概念个数	概念比例	概念示例	概念说明
1866	1	0.05	丁七万三上下与丐丑专且世丘丙业从东丞两严…	姓氏
261	1	0.05	丽亳亹令任侯俦僬兖兹臣单厍厦吓吴喷圬坜坻…	地域的简称、别称
136	1	0.05	匏柊棘稂稗箣篙缩艻艾艿芏芑芒芘芣芥芦芨芩…	草本植物
123	1	0.05	们刺妫婺岷建桂汉汜汧汨汶汾沂沅沌沔沘沣沩…	水域、河流名称
118	1	0.05	乌凫凰凤枭椋翡鸷蔺虎蝶螺隹隼雀雇雁雕鹳…	鸟类
113	1	0.05	朴杉杧杨杻松构枋枞枥枫枸枹梅柃柏柚柜柞柳…	木本植物
110	1	0.05	鲍鲟鲨鲂鲛鲵鲩鲫鲔鲷鲭鲑鲢鲅鲣鳃鲦鳘鲽鳊…	鱼类
91	1	0.05	且伋仙佺佛佚偃偈傕其冒鼹尚吒员嘟嘗埒垓堃…	人名用字
89	1	0.05	龈拉孛萤虮虾虻蚂蚄蚋蚜蚣蚨蚫蚴蛆蛈蛙蛄蛉…	节肢动物
83	1	0.05	金鏈鑞钆钇钋钌钍钐钒钔钕钙钚钛钠钡钨钪钫…	金属
71	1	0.05	元兒冉南卫召吴周唐商夏娀宋巴明晋曹朝杞梁…	朝代或国家名
63	1	0.05	吖吗吲呋哌啉嘧啶嗪嘌巯氨氰溶甙苗砜糖羧羰…	制药的有机化合物
53	1	0.05	玉玘玙玚玟玢玦玫玮珅珏珖珙珣珩珰珺珽琀琅…	各种玉或玉器
52	2	0.10	卜椿瓜蝙笋芋芡芥芦芫芰芹苋苓苣茼荇菇荠荚…	植物中的蔬菜类
41	1	0.05	华台圌峚岈岍岐岘岞峋岱岷峄峒峱崂崃崆崑崙…	山名
40	1	0.05	仡仫佤佧侗俫傈傣僰僳党华回壮夷家彝戎掸擁…	少数民族
39	1	0.05	骞驮驶驹骍骈骥马驴驵驷驸驸驹驼驽骀骃骈…	马类
37	2	0.10	呢哔布帛彩缯絁绩缟红纨纱纶纺纻绒绡绢绤绨…	丝织品
36	2	0.10	云京冀台川广晋桂楚沪豫浙渝港湖湘滇澳琼中…	行政区划的省级单位
28	1	0.05	匏李杏杧枇枣枸柑柚柠柰枥栗桃梳梅梨椑椰榅…	植物的果实
27	3	0.15	丞令侯傧傧僮卿史吏士官宦宰宋尉尉尹弁授畯…	古代官名
26	3	0.15	卣卮卺尊彝斗斝植樽爵瓯瓴盅盉盦盏簋罍角觥…	与酒有关的器皿
24	1	0.05	冈坂阪坳埯埼塬堠壁岕岑岬峡峪崖崿嵴嶂谷麓…	山体的某一部分
23	1	0.05	禾秫秬稆稞稷稷稻籼粟粱粳糜荞莜菽谷麦麻黍…	粮食作物
22	1	0.05	治府寺监部馆所厅院科局股处室课段署家司部…	政府机关部门
21	1	0.05	囤席笆箕筐箪筥筲筹箩篑篓篮篚篰篼篾篦箪篆…	竹或木制成的容器
20	3	0.15	酊酎酒酘酝酤酿酶酷醇醐醑醚醪醪醴醽醾卣坞	酒类
19	5	0.25	帘帜幖幡幢旂旆旃旄旌旎旐旒旗旟纛望蠹麾	旗帜
18	4	0.20	内心牙肝肠肺肾胃胆胰胱胴脏脾膀腺膀膵	动物内脏
17	4	0.20	土地垌垸墒田町畈畦畲畤荒莱苗衍隰圩	土地田地
16	7	0.35	匠哲器彦才材杰氏秀英豪贤通骥模尖	有才能的人
15	6	0.30	体例则制宪彝律法矩科臬规贯轨辟	法律、规章制度
14	11	0.55	刀戈戟戣枪槊矛鏦鎩铍铩铗铤镖	长矛类兵器
13	9	0.45	丫桠本杈条杪枝柯标株梗桠梗	树枝
12	15	0.74	丘冢圹坟墓墦寝穴窀茔阡陵	坟墓
11	14	0.69	上储后君圣帝庙王皇辟驾	古代帝王
10	24	1.19	位品地档流等级职衔阶	社会地位、等级
9	19	0.94	粑糕酥饼饼馃馇馍馕	糕点
8	37	1.83	勋印戳玺章篆钤鼎	印章、图章
7	50	2.48	唾唎沫津涎漦痰	唾液
6	81	4.01	舟舡航舫船艑	船只
5	99	4.91	志标符记识	标志
4	156	7.73	棰箠鞘鞭	用于鞭打的工具
3	212	10.51	仲孟季	兄弟排行位置
2	405	20.07	伞盖	用于挡雨、遮阳的工具
1	821	40.68	驮	牲口驮着的货物
不限	2018	100	N/A	N/A

2. 汉语的语素概念表示与语义构词分析

动语素“语素概念”覆盖、分布情况

集合字数	概念个数	概念比例	概念示例	概念说明
77	1	0.06	俏冲卤卧嘘扒摊氽涮溜溇瀹炆炊炒炖炝炮炸烀…	加工、烹饪食物的方式
34	2	0.12	儡剋吡哔呵咄咎哈嚷埋怨怪批撸斥派熊病诃诮…	责备
32	2	0.12	储厝囤坌垛堆委存存窖寘居庋庤弆攒㩟滀潴积…	存储
29	1	0.06	云侃具叙吭启咧哨唠扯拉提摆曰称聊言讲话语…	言谈交流
28	1	0.06	刳剐剖劈异溃掊溃披敞析撑毂睁解捭展挓巴启…	分开、张开
26	2	0.12	死亡化卒危夭尽徂故殁殂殇殣殪毙没终绝薨逝…	死亡
25	4	0.25	了会博喻审悉悟惺憭懂明晓曝照省知解谂通醒…	理解、明白
24	1	0.06	冤哄唊拐捞欺绷罔肥蒙虞诈诓诖诱诳谩赚逗钓…	欺骗
23	2	0.12	仗手扒执把持挝挠揭捉捞掺提揪搏搦操捱攀攥…	手的抓取、持握的动作
22	3	0.18	上傅刮刷图垩塈墩打扙抹抹抹抿拭揾搽敷油涂…	涂抹、擦刮
21	4	0.25	临促傍凑即压向守就拢挨挨毗濒薄近迫逼贴附…	靠近
20	2	0.12	与丐予付付供发命奉把投授效施献畀给缴致赋	送出、给予
19	10	0.61	会佸参摊撞晤搴碰见觌觏遇赶迕逢逷遘遭邂	遇见
18	6	0.37	上下之到即如幸往往徂抵至莅赴赶蹬适造	去、往、到
17	5	0.31	串为化变变变可失嬗愈成拜构济移迁革	改变、变化
16	4	0.25	下乏亏匮少少差悭拉朒欠短离缺该阙	缺乏
15	8	0.49	上偏到勾及够平底彀满等致臻达齐	达到高度、目标、期限
14	11	0.67	号吆吼呐嗷呼哮啡唱喊喝噪嚎嚷	大声地叫喊
13	17	1.04	冶化化泮溶炀烊焊熔融铄销锢	物体溶解或融化
12	14	0.86	刷抡拔拣择择挑擢调选遴铨	挑选、选拔
11	19	1.17	倒北破砸胜衄负败败败输	失败
10	28	1.72	书作修做写撰泐纂编著	文学写作或学术创作
9	20	1.23	剽劫夺抢掠掳攘越逼	掠夺、劫取
8	39	2.39	假僦租租赁贷赁赁	借取、租取
7	45	2.76	抽生疯结苗起长	植物生长
6	59	3.62	拯斋施舍赍食	出于怜悯的施舍
5	86	5.28	扳斗竞角赌	竞技比赛、争胜
4	121	7.42	抶掠箠鞭	鞭打
3	163	10.00	冤屈歉	感到冤枉
2	305	18.71	俘虏	俘虏
1	645	39.57	萌	植物发芽
不限	1630	100	N/A	N/A

2. 汉语的语素概念表示与语义构词分析

形语素“语素概念”覆盖、分布情况

集合字数	概念个数	概念比例	概念示例	概念说明
46	1	0.18	亮光冏昉明昪晅晞晟晳晶暲曙朏朗杲灼灿炅炜…	光亮、耀眼
31	1	0.18	侗倥僋傻兀冥朦悖惛惝愚愦惷憨懵憃昏昧暗浑…	愚蠢的、糊涂的
27	1	0.18	举俗全到合周圞完尽彻总普毕洽浑溥满漫叆盛…	完整的、完备的
26	2	0.36	令佼佷劭嘉姝姱婉媚嫽嫿峥峬懿旖昌琦璀皖粲…	美好的
25	2	0.36	丰丽俊俏倩冶妍妖姣姹娩娇娈娟嫣婵帅扬昳玉…	样貌美丽的
24	1	0.18	乐休哈娱快忭忺忻怡怿恺悦愉旷昪欢欣歆熙豫…	欢喜的、快乐的
23	2	0.36	仡伟偈剧勇勐壮悍敢朅枭武激烈狂猛矫虎骁骄…	勇猛的、强健的
22	2	0.36	汪沄洞泓洸泠洌浏湉湛湜滉滢漻潇潲澄澈秀粼…	形容水流
21	2	0.36	丰优博厚夥奂富广敞旺殷洋浩海盛腴赡趁饶足…	丰富的、丰盛的
20	2	0.36	冷唷嘿宓寂寂寥幽恬悄泊清澹谧闷阒静默寞愔	寂静的、冷清的
19	3	0.55	丑倦傒厌堵忳恶恼懊憋斁毵烦烦烦絮腻腻臭	厌恶的、厌烦的
18	3	0.55	凶恶悍惨瞀暴横横残犷狠狰猖粗赳豪酷骜	残暴的、凶狠的
17	3	0.55	乖俐卓巧惺憨慧敏智猴睿神秀精聪颖鬼	乖巧的、机灵的
16	4	0.73	凛寒怯怵慄恂恫猲恟怂伈惶觳屏慑慄	胆怯的、害怕的
15	5	0.91	忠恂恳悃悫悾馐挚敦款笃肫忱谆赤	忠诚的、真诚的
14	3	0.55	敻曼杳泂缅辽远迢迥迷逴遐遥邈	距离遥远的
13	3	0.55	切忧怫悄悒惆惙惴慅戚敌荩虑	内心担忧、焦虑的
12	8	1.46	刁嚣奸油滑狡狯猾诡谲贼黠	奸邪的、狡猾的
11	10	1.82	冥刻奥妙微深深玄稀赜邃	深奥的
10	14	2.55	仄挤浅狭紧细蕞迮逼隘	狭窄的
9	10	1.82	个匹单单只孑孤索茕	单独的、孤单的
8	17	3.10	严冷凛峻愀懔狠苛	严肃的、严厉的
7	18	3.28	倥寒拮穷窘窭贫	生活贫穷的
6	29	5.28	夭葶茂蓁萋蕤	植物茂盛
5	45	8.20	芳菲郁香馝	形容味道香
4	46	8.38	暵槁枯萎	植物干枯
3	62	11.29	万千多	形容数量多
2	99	18.03	涩苦	形容味道苦，与甜相对
1	151	27.50	馋	想吃、嘴馋
不限	549	100	N/A	N/A

2. 汉语的语素概念表示与语义构词分析

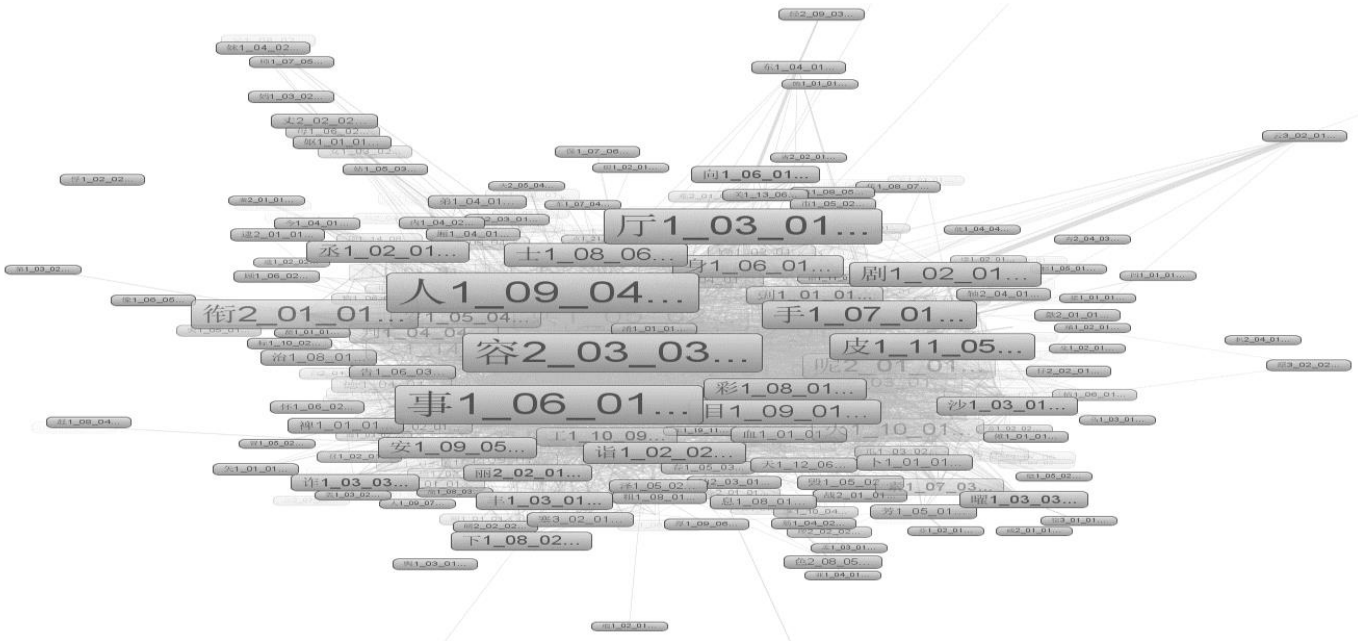
- 当前工程进展：目标 3 的阶段性数据成果
  - 以面向对象思想、生成词库理论等为指导，构建汉语的“**语素概念体系**”（**字义系统**）
  - 汉语的名、动、形语素的层次结构是大致同构、映射的，形成同语素类内的聚合关系以及跨语素类间的组合关系，以方便认知、计算和推理
  - 在多种特征赋值基础上，借助“特征序列”的设定实现多分类树功能
    - 陈刚，刘扬．基于特征序列的语义分类体系的自动构建．中文信息学报，2015，29（3）：52-57

2. 汉语的语素概念表示与语义构词分析

- 当前工程进展：目标4的阶段性数据成果
  - “**语义构词描述**”1：“构词结构”标注，以“选材”为例
    - 语法结构标签：述宾，语义结构标签：受事 [已融合了：对象]
  - “**语义构词描述**”2：“语素概念”绑定，以“选材”为例
    - “选”字的语素义为“选择、挑选”，语素义编码为“选1_04_01”
      - 绑定了“语素概念” X：
      - {刷3_01_01, 抡1_01_01, 拔1_08_03, 拣1_01_01, 择1_02_01,
      - 择2_02_01, 挑1_02_01, 擢1_02_02, 调4_02_02, 选1_04_01,
      - 遴1_01_01, 铨1_02_01}
    - “材”字的语素义为“有才能的人”，语素义编码为“材1_05_04”
      - 绑定了“语素概念” Y：
      - {匠1_02_02, 哲1_02_02, 器1_05_04, 尖1_09_06, 彦1_02_01,
      - 才1_03_02, 材1_05_04, 杰1_03_01, 模1_04_03, 氏1_05_03,
      - 秀2_04_04, 英1_03_02, 豪1_04_01, 贤1_04_02, 通2_12_07,
      - 骥1_02_02}

2. 汉语的语素概念表示与语义构词分析

- 当前工程进展：目标4的阶段性数据成果
  - 首次以“**语素概念**”为节点，基于52108个双字词的“语义构词”信息，在语言系统层面上刻画了**汉语构词过程中基本意义单元之间结合、转移的状态**



报告提纲

- **1.** 回望汉语世界中的字词关系
- **2.** 汉语的语素概念表示与语义构词分析
- **3.** 汉语语素系统的应用与展望

3. 汉语语素系统的应用与展望

- 这些基础性研发工作有望推动**人文领域**和**计算应用**的开展
  - 充分考虑汉语的语言事实与特点，从汉字及其基本意义入手
    - 刘扬, 林子, 康司辰. 汉语的语素概念提取与语义构词分析. 中文信息学报, 2018, 32(2): 12-21
- 一、在**人文领域**方面的应用尝试与探索方向
  - 1. 支持新型汉语电子词典的编撰、出版与浏览使用
    - 字、词意义空间划分、对应的验证与考核，严格的可计算化
    - 不必再依赖字形或拼音了，以电子数据方式，让基于“语素概念”的查询成为可能，并快捷遍历所有（形式或）意义相关的字、词
  - 2. 支持汉语字、词的母语教学以及对外汉语教学
    - 网站查询与浏览服务
    - 手机微信小程序应用
    - 作为插件，对第三方软件（涉及汉语学习理解）的广泛的、无缝的支持

3. 汉语语素系统的应用与展望

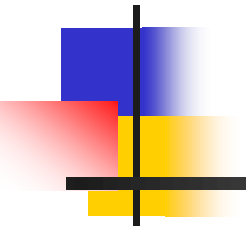
- 一、在人文领域方面的应用尝试与探索方向
  - 3. 支持更加细致、深入的汉语语言认知研究
    - 汉语的字义演变与词汇化过程的解释研究
      - 康司辰，虞梦夏，刘扬．基于平行周遍原则的汉语未登录词的知识表示与预测．中文信息学报，2020，34（8）：23-31
    - 基于语言认知（隐喻、转喻等）理解的实证研究
      - 陈龙，饶琪，刘扬．汉语词的非字面义的表示与应用．中国科学：信息科学，2019，49（8）：1005-1018
    - 汉语基本情感与复合情感的表征、分析与计算
    - 汉语的语义本体构建与应用评估
- 二、在计算应用方面的应用尝试与探索方向
  - 1. 面向理解的汉语未登录词的词义知识表示及语义预测
    - 田元贺，刘扬．汉语未登录词的词义知识表示及语义预测．中文信息学报，2016，30（6）：26-34

3. 汉语语素系统的应用与展望

- 二、在计算应用方面的尝试方向
  - 2. 基于语素概念与语义构词分析等汉语特性的认知计算
    - 康司辰，刘扬．基于语义构词的汉语词语语义相似度计算．中文信息学报，2017，31（1）：94-101
    - 康司辰，虞梦夏，刘扬．基于平行周遍原则的汉语未登录词的知识表示与预测．中文信息学报，2020，34（8）：23-31
    - 郑俪，刘扬，殷雅琦，王悦，代达劢．基于词信息嵌入的汉语构词结构识别研究．中文信息学报，2022，36（5）：31-40，66（第二十届中国计算语言学大会 CCL2021，会议最佳论文）
    - 王悦，刘扬，梁启亮，王涵思．汉语语义构词的资源建设与计算评估．语言文字应用，2023(4)：105-117
  - 3. 汉语等语言中的语素义、词义的表示、应用与可解释性研究
    - Lin Zi, Yang Liu. Implanting Rational Knowledge into Distributed Representation at Morpheme Level. AAAI 2019

3. 汉语语素系统的应用与展望

- 二、在计算应用方面的尝试方向
  - 3. 汉语等语言中的语素义、词义的表示、应用与可解释性研究
    - Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, Yang Liu. Decompose, Fuse and Generate A Formation-Informed Method for Chinese Definition Generation. NAACL 2021
    - Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, Yang Liu. Leveraging Word-Formation Knowledge for Chinese Word Sense Disambiguation. EMNLP Findings 2021
    - Yaqi Yin, Yue Wang, Yang Liu. Chinese Morpheme-informed Evaluation of Large Language Models. COLING 2024
    - Yue Wang, Hua Zheng, Yaqi Yin, Hansi Wang, Qiliang Liang, Yang Liu. Morpheme Sense Disambiguation: A New Task Aiming for Understanding the Language at Character Level. COLING 2024
    - Yue Wang, Qiliang Liang, Yaqi Yin, Hansi Wang, Yang Liu. Disambiguate Words like Composing Them: A Morpheme-Informed Approach to Enhance Chinese Word Sense Disambiguation. ACL 2024 (to appear)



谢谢各位，请批评指正！

liuyang@pku.edu.cn