

文章编号: 1003-0077 (2017) 00-0000-00

汉语的语素概念提取与语义构词分析*

刘扬^{1,2}, 林子^{1,3}, 康司辰^{1,3}

(1. 北京大学 计算语言学教育部重点实验室, 北京 100871;

2. 北京大学 计算语言学研究所, 北京 100871;

3. 北京大学中国语言文学系, 北京 100871)

摘要: 作为基础的表义单位, 语素及此上的构词分析, 既是汉语作为意合语言进行语义分析的起点, 也是认知、理解词义的关键。本文提出了一种探寻汉语语义基元和分析语义构词的新方法和视角: 基于语素义相似度计算形成“同义语素集”, 用来表征“语素概念”, 并借鉴生成词库理论和面向对象思想形成“语素概念体系”; 建立在这些基础上的汉语语义构词分析, 在全局性语义分析、数据挖掘等方面也有新的进展。这些思路、做法及语言资源建设, 有望推动人文领域和计算应用等相关工作的开展。

关键词: 语素; 语素义; 语素概念; 语义基元; 语义构词

中图分类号: TP391 **文献标识码:** A

Towards a Description of Chinese Morphemic Concepts and Semantic Word-Formation

Liu Yang^{1,2}, Lin Zi^{1,3}, Kang Sichen^{1,3}

(1. Key Laboratory of Computational Linguistics (Ministry of Education), Peking University, Beijing 100871, China;

2. Institute of Computational Linguistics, Peking University, Beijing 100871, China;

3. Department of Chinese Language and Literature, Peking University, Beijing 100871, China)

Abstract: Morphemes and their Word-Formation Analysis are both the starting point for the semantic analysis of Chinese as the parataxis language, and also the key to understanding the meaning of words. This paper presents a novel approach to exploring the Chinese Semantic Primitives and using them for Semantic Word-Formation Analysis: first form the Synonymous Morpheme Sets, used for denoting the Morphemic Concepts, based on similarity calculation of Chinese morpheme glosses; then form the Morphemic Concept Hierarchy, serving as a systematic description of the Chinese Semantic Primitives, by principles of the Generative Lexicon Theory and the Object-Oriented Ideas; built on these, Chinese Semantic Word-Formation Analysis has made new progress from overall consideration and data mining. These ideas, practices and language resources construction are expected to promote the humanities and computing applications as well.

Keywords: morphemes; meaning of morpheme; morphemic concepts; semantic primitives; semantic word-formation

1 引言

在汉语中, 存在着“语素、词、短语、句子”等由小到大的语言单位和层级结构, 而语素构词更是汉语的特点。作为基础的符号单位, 语素及其意义, 以及此上的构词分析和意义表达, 既是汉语语义分析的起点, 也是计算机理解词义的关键。

在研究构词结构时, 人们很早就注意到了汉语词法与句法的平行性, 汉语中的由字组词、由词造句过程遵循同一原则。赵元任^[1]认为构词成分之间存在造句关系, 此后, 陆志韦^[2]、朱德熙^[3]、王洪君^[4]等指出, 复合词内部的结构关系和句法结构是类似的。这在汉语词的历时形成过程中亦可找到解释, 董秀芳专门提到^[5], 现代汉语中的多字词多是古代汉语中单字词短语词汇化的产物, 一些复合词的前身即是自由的句法组合; 另一方面, 考虑构词结构下的成分与整体, 语素义与词义在某种程度上显然是关联的。徐通锵^[6]分析汉语社团的思维方

*收稿日期: 定稿日期:

基金项目: 国家重点基础研究发展计划资助项目 (2014CB340504)、国家社科基金一般项目 (16BYY137)、
国家社科基金重大项目 (12&ZD119)

式与编码机制,强调汉语作为语义型语言,字的表义性是其内在结构基础。此外,符淮青^[7]、周荐^[8]等人也注意到了汉语词的意合特征,认为汉语中的语素义(字义)和词义之间具有很强的推导性。

这表明,探究汉语的语素构成及其意义系统,以及在此基础上的语义构词分析有扎实的理论基础和潜在的应用价值。

从自然语言处理的实践看,此前,句法及语义分析一直居于主流地位,典型工作如汉语的短语结构规则研究^[9]、短语结构句法分析^[10]、依存结构句法分析^[11]、句子语义计算^[12]等开展较为充分。但是,对汉语语素、词法和意义的系统化的构建和分析工作还比较欠缺。目前,关于语素与构词分析方面的研发工作主要包括以下几项:

清华大学苑春法的“汉语语素数据库”^[13],以语素描写和构词分析为核心,覆盖常见汉字的语素项信息,包括语法类、语素义的刻画,并对语素项构成的汉语词进行了结构描述和意义绑定。但不同的语素项之间是彼此孤立的,缺乏面向整个语言系统的意义关联,只以离散的语素项集合的面貌出现,没有形成体系结构,无法满足基于意义比较的计算需求;

鲁东大学亢世勇的“汉字义类信息库”和“汉语语义构词信息库”^[14],前者描写了常见汉字的字位(不妨理解为语素的义项),后者在此基础上对二字合成词进行标注,对字位和合成词均进行了归类并形成了积极的意义关联。归类以前已有的《同义词词林》为标准,存在语素义与词义的本原、因果参照问题,结构合理性有待商榷;

中国台湾周亚民的汉字知识本体(Hantology)^[15],分析了取自许慎《说文解字》中的540个部首汉字的基本义符所刻画的概念,并映射到IEEE SUMO上层共用知识本体上,形成了与世界通用概念(该通用概念由英语词汇来承担)对应的层次结构。该本体在分类上同样存在先天的参照问题,且只考虑少数部首汉字的粗粒度的基本意义,也难以对汉语的语素义认知、计算提供足够的支撑;

中科院董振东的知网(HowNet)^[16],认为任何一个概念(语言中的词所表征的义项)均能够分解为一组义原并以此为基础来加以定义,并且,在不同语言中存在同样的义原集合。基于对汉字的考察、分析,董先生归纳、提取了1600多个义原,采用人工给定的英-汉词汇序列(如“i11|病态”)表示并在其间形成了层次结构。这些义原均没有特定的语素载体,其定位近于抽象的语素义。知网注意到了汉语的意合特征,为汉语的词义计算做出了贡献,但它并没有走语素和构词分析的路,且义原的形成和认定也带有较强的主观性。

这些先驱工作开拓、丰富了人们的视野,值得思考和借鉴。与此同时,它们在汉语语素及其意义的构建客观性、数据覆盖度、结构体系化以及汉语构词的全局性语义分析、数据挖掘与可视化等方面,还有期待改进的地方。

我们希望在WordNet理论^[17]、生成词库理论(GLT理论)^[18]、面向对象思想^[19]等观点指导下,以《现代汉语词典(第5版)》(以下简称《现汉》)刻画的全部的汉语语素及语素义为客观依据,基于语素义的相似度计算形成“同义语素集”,用来表征“语素概念”并建立“语素概念系统”,以描述汉语世界中的语义基元状况。在此基础上,进一步描述汉语词的构词结构,实现构词结构下的构词成分与“语素概念”的严格绑定,系统研究汉语的语义构词现象并做数据获取、挖掘和可视化呈现,推动人文领域和计算应用等相关工作的开展。

2 汉语语素概念提取方法

2.1 语义基元理论基础

语言中的语义基元揭示了人们思维中的核心语义概念,在语言认知与计算等诸多领域扮演着重要角色^[20]。上世纪30年代,Sapir在其系列著作中探究了“基本语义单元”的概念^[21,22],Morris随后表达了对于以后出现该类系统的期望^[23]。到70年代,Wierzbicka等人认为“复合词的语义能够被一组意义更简单、更易理解的词语来解释”,并称其为语义基元(semantic primitives)^[24],这是重要的思路 and 提示。然而,在各种语言中,目前还没有找到表征和生成语义基元的有效方法。

在英语中,语素处于相对弱势的地位,语言中的概念意义主要由词来承载和体现,WordNet率先采用“同义词集”来表征“词汇概念”。值得注意的是,汉语是一种意合语言,语素作为最小的字符单位具有很强的表义性,对更大单位的词义的贡献十分明显。结合Wierzbicka等人的观点,并考虑汉语构词的显著特点,我们希望以“同义语素集”来表征

“语素概念”，一个“同义语素集”包含了语言中大致同义或同类的所有语素，代表了汉语世界中的一个基本的“语素概念”，即语义基元。在全部“语素概念”之上，还可以做结构化工作，进而形成“语素概念体系”。

2.2 语素类区分与语素义编码

考虑现有词典的权威性和应用的影响力，我们的汉字语素取自于商务印书馆《现代汉语词典（第5版）》（以下简称《现汉》）中的定义。

尹斌庸指出^[25]：“语素不一定是词，但却明显地具备着词性……从词的词性能够类比出语素的词性”。仿照词类，我们引入语素类的概念，将语素分为名语素、动语素、形语素、副语素、数语素、量语素、代语素、介语素、助语素、连语素、拟声素、叹语素、缀语素等13类。目前，在语素的义项区分的基础上，《现汉》只为成词语素（即单字词）标注了词类，可直接视为成词语素的语素类；对不成词语素，我们用人工标注的方式补齐了语素类。这样，共获得8514个汉字（包括繁体、异体字）的20855个语素类信息。其中，名、动、形语素分别占46.90%、30.59%、11.25%，共计88.74%，构成主体；而其余10类语素共计11.26%，形成补充。

在此基础上，我们对上述13类语素下的不同语素义项（即语素义）做释义文本的提取，并赋予唯一的“语素义编码”。比如，“材”字有多个语素义，其中的一个释义文本为“有才能的人”，其“语素义编码”为“材1_05_04”，依次表明：它是该字在《现汉》中的第1次条目出现，该条目下共有5个语素义，当前为第4个语素义。

2.3 语素义的相似度计算与语素概念生成

为了获得可靠的“同义语素集”，需要对《现汉》中的不同语素义的释义文本进行语义相似度计算。

吕叔湘^[26]在《〈现代汉语词典〉编写细则（修订稿）》中提出了一些释义原则，“同属一类条目，注释措辞必须一致，避免分歧”。例如，对于“雏”和“仔”，词典中表示“幼小的”概念的形语素的释义文本如下：①雏_形幼小的（多指鸟类）；②仔_形幼小的（多指牲畜、家禽等）。考虑释义文本的这些特性，在语义相似度计算方法的选择上，我们采用“字共现”模型，该方法兼具实现简单、效果显著等优势。对于不同语素义的释义文本 S_1 、 S_2 ，语义相似度Sim值计算公式如下，其中， A 、 B 分别表示 S_1 、 S_2 中的汉字集合。

$$\text{Sim}(S_1, S_2) = \frac{|A \cap B|}{|A \cup B|}$$

对于特定语素类的任一语素义的释义文本，按照它与其它语素义的Sim值降序排列，并按设定阈值将意义相近的语素义推荐给专家。经人工检验，每确定追加一条即对其做Sim值计算的迭代，如此反复补充、过滤，形成一个“同义语素集”（其中的语素是大致同义或同类的），亦即一个“语素概念”，或称一个语义基元。对剩余语素义的释义文本，重复此过程，直至覆盖该特定语素类的全部语素义为止。然后，选择新的语素类，重复如上过程。

2.4 语素概念的结构化与系统描述

在获得汉语的“语素概念”全集之后，有必要进一步为这些语义基元之间建立起层次结构，让离散的概念维持基本的语义关联，形成义场，以方便认知、推理和计算。

受WordNet启发，名语素的“语素概念”主要依据上下位关系进行结构化建设。例如，表示“草本植物”的“语素概念”{艸1_02_01, 柃1_01_01, 棘1_03_02, 稂1_02_01, …}和表示“木本植物”的“语素概念”{朴3_01_01, 杉2_02_01, 柘1_02_01, 杨1_02_01, …}均属于名语素层次结构“物—具象物—生物类—生物个体—植物”下的节点，形成同语素类内的聚合关系。

在跨语素类的语义关联方面，则借鉴生成词库理论，对动语素和形语素，分别建立起以名语素层次结构为中心和参照的对应体系。其中，动语素表达名语素所指事物的事件，或者说，动语素的主体是对应的名语素；形语素表达名语素所指事物的性质，或者说，形语素修饰的对象是对应的名语素。由此，名、动、形等不同语素类的层次结构是大致对应的、同构的，并形成同语素类内的聚合关系以及跨语素类间的组合关系。例如，表示“植物发芽、生长、枯萎等行为”的动语素节点和表示“植物茂盛、干枯等特征”的形语素节点，它们都与表示“植物”的名语素节点相关联，分别挂靠在动语素、形语素层次结构“物—具象物—生物类—生物个体—植物”下。实际上，语言中的生成词库理论与计算机中的面向对象思想

有相似的内涵，该知识表示有利于各类“语素概念”的组织 and 计算。

基于以上方案，我们对汉语的“语素概念”全集建立了层次结构，并对内部节点进行了特征描写和赋值，形成了图 1 所示的“语素概念体系”，这也是对汉语世界中的语义基元的系统描述。

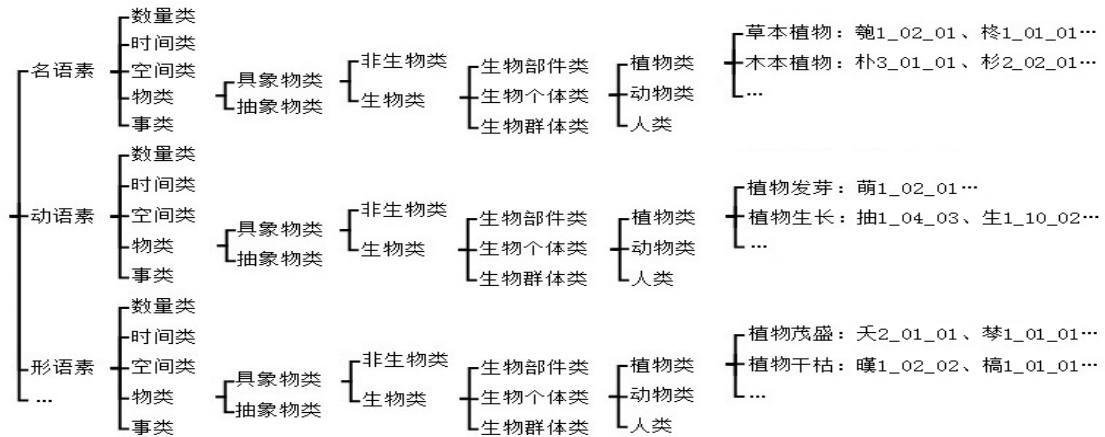


图 1 汉语“语素概念体系”示例

3 汉语语义构词分析方法

3.1 语义构词理论基础

对于汉语的构词结构性性质，语言学界一般有语法构词^[1,2,3]、语义构词^[6,27]等两种不同观点。前者强调构词成分之间的语法关系，如主谓、述宾等语法结构标签的认定，而后者强调构词成分之间的语义关系，如主体、客体等语义结构标签的认定。

考虑语言计算、应用的状况和需求，傅爱平^[28]指出：虽然语义构词在表示词义时有天然、直观的优势，但是其结构产生依据过于复杂，标签集难以统一，并不利于机器处理。相比之下，语法构词的结构体系较为简单，标准统一，且与句法结构有某种天然的相似性。苑春法的研究也表明^[13]，语法构词结构与构词语素类、词性之间存在一定的相关性，采用语法构词体系更有利于计算的开展。

在借鉴前人观点的基础上，我们选择语法构词体系以方便工程展开，这也遵循了自然语言处理中从形式到意义的主流路线。值得注意的是，事实上，由于后续环节要求构词成分对“语素概念”的严格绑定，我们获得的依然是广义的语义构词知识。

3.2 构词结构类型界定与标注

语法构词结构体系，大多沿用朱德熙^[3]的方案。杨梅^[29]在借鉴了语法构词和语义构词两派的观点后，提出了一套以语法标签为主的构词结构体系，并兼顾了语义构词派的部分观点。考虑到这套体系的兼容特点以及合适的标签数量，它成为我们工程开展的首要参考。

在杨梅标签基础上，我们进行了适当的增添和删减。增加“单纯式”标签用于表示成分义与词义之间缺乏明显关联，并将“附加式”细分为前附加、后附加。同时，删除了一些缺乏计算价值或结构类别实例过少的结构标签，如截取式、虚配式、指量式、数构式。最终确定的构词结构标签集包括 16 种，即：主谓式、连谓式、联合式、述宾式、述补式、定中式、状中式、介宾式、重叠式、名量式、数量式、方位式、复量式、前附加、后附加、单纯式。

界定了构词结构体系之后，在词的义项区分的基础上，我们为《现汉》中的所有二字词依规范标注了构词结构，共计 52108 个。为保证构词结构知识的可靠性，请三位专家对同一词项进行标注，标注结果的两人或以上一致率为 93.46%。

3.3 构词成分与语素义绑定

在构词结构标注的基础上，对二字词中的构词成分，即前后语素，我们继续标注它们在《现汉》中的语素义，并依据语素义的标注获得对应的“语素义编码”。

注意到，一个语素义对应一个“语素义编码”并进入一个“同义语素集”，这一过程实际上是将构词成分与特定“语素概念”建立了绑定关系，并受整个“语素概念体系”意义系统的表达和制约。这样一来，单一的语素义就携带了丰富的、便于计算的多种内容，包括了其在“语素概念”中的“同伴”信息、在“语素概念体系”的“位置”信息以及由此取得的基于继承链条的一系列“特征取值”信息。

3.4 词义知识表示与意义序列输出

符淮青^[7]等语言学家指出：语素义的组合在一定程度上体现词义。因此，利用语义构词知识进行词义知识表示是一种新的选择。这种表示具有简单、直观的特点，并反映构词成分对词义的贡献。例如，在“选材”中，“选”的语素义为“挑选、选拔”，“材”的语素义为“有才能的人”，其结构关系及成分义能够较为准确地反映“选材”的语义。

基于上述工作，我们获得的语义构词知识涵盖词性、构词结构、前后语素类、前后语素义等广义知识，其中前三个属于语法层，最后一个属于语义层。例如，“选材”的语义构词知识如表 1 所示。需要指出的是，前后语素义的“语素义编码”已经携带了丰富的、便于计算的多项信息。

表 1 语义构词知识示例

例词	词性	构词结构	前语素类	后语素类	前语素义	后语素义
选材	动词	述宾结构	动语素	名语素	选 1_04_01	材 1_05_04

为了对各种构词结构提供表达词义的一般指导，并获得系统化的诱导词义的方法，还需要在构词结构和词义表达之间搭建意义关联的模式。

亢世勇^[14]曾给出包括 $A+B=A+B$ 、 $A+B=A=B$ 、 $A+B=A$ 、 $A+B=B$ 、 $A+B=C$ 、 $A+B=A+B+D$ 、 $A+B=A+D$ 、 $A+B=D+B$ 等形式的意义结构体系，其中 A、B 分别表示二字词的前语素义和后语素义，C 代表转义意义，D 代表附加意义。这种体系分类详细，但相关知识难于表示和获取。考虑转义和附加义的占比不高，从应用的角度出发，我们采取了一种相对简单、方便计算的意义结构体系，如表 2 所示。在本文中，暂时只考虑词的字面意义，即本义。关于词的引申义问题，在其它后续环节中将给出新的解决方案和计算手段，目前并不涉及。

表 2 意义结构与构词结构的对应关系

意义结构	语素义和词义的关系	构词结构	例词
11 型	词义与前后语素相关性均较高	主谓、连谓、联合、述宾、述补、定中、状中、介宾、数量、施、受	选材、红旗
10 型	词义只与前语素相关性较高	后附加式、名量	忘却、船只
01 型	词义只与后语素相关性较高	前附加式	老虎、仔细
00 型	词义与前后语素义相关性均较低	单纯式	沙发、名堂

在此基础上，我们依据意义结构与构词结构的对应关系给出词的“意义序列”输出形式。该序列为构成语素的“语素义编码”的排列，内容和顺序基本由构词结构决定。仍以“选材”为例，其“意义序列”为“<选 1_04_01, 材 1_05_04>”。此外，允许在需求中依据约定改变序列顺序，以表达计算应用的灵活性，如“<材 1_05_04, 选 1_04_01>”也可认为是一个合法的“意义序列”。更多不同构词结构类型下的二字词的“意义序列”输出如表 3 所示。

表 3 二字词的“意义序列”示例

例词	词性	例词	构词结构	意义结构	词的“意义序列”
年轻	形容词	年轻	主谓	11 型	<轻 1_09_01, 年 1_11_04>
进攻	动词	进攻	连谓	11 型	<进 1_06_01, 攻 1_04_01>
丰满	形容词	丰满	联合	11 型	<丰 1_03_01, 满 1_07_01>
选材	动词	选材	述宾	11 型	<选 1_04_01, 材 1_05_04>
提高	动词	提高	述补	11 型	<提 2_10_02, 高 1_08_05>
红旗	名词	红旗	定中	11 型	<旗 1_06_01, 红 1_06_01>
热爱	动词	热爱	状中	11 型	<爱 1_05_01, 热 1_10_05>
从小	介词	从小	介宾	11 型	<从 2_04_01, 小 1_09_06>
哥哥	名词	哥哥	重叠	11 型	<哥 1_04_01>
一些	数量词	一些	数量	11 型	<些 1_02_01, 一 1_10_01>

野外	名词	野外	方位	11 型	<外 1_08_01, 野 1_07_01>
场次	量词	场次	复量	11 型	<场 2_09_07, 次 1_08_05>
船只	名词	船只	名量	10 型	<船 1_01_01>
忘却	动词	忘却	后附加	10 型	<忘 1_01_01>
老虎	名词	老虎	前附加	01 型	<虎 1_04_01>
克隆	动词	克隆	单纯词	00 型	<>

4 数据结果分析

4.1 关于汉语语素概念的分析

名、动、形语素构成汉语语素的主体，其它语素类只占很小的一部分。根据计算结果和工程进展，目前，名、动、形语素分别形成了 2018、1631、550 个“语素概念”，共计 4199 个“语素概念”（注意：不包括“语素概念体系”中的内部节点）。表 6、7、8 依据“同义语素集”的大小、多少等信息，分别展示了名、动、形语素“语素概念”覆盖、分布的一般情况。

表中一行描述了同种规模（即“同义语素集”的大小，或称语素个数、集合字数）的“同义语素集”的大致情况。每行有 5 项信息：1、“同义语素集”的大小，即集合字数；2、该大小下的集合个数，即概念个数；3、占同语素类“语素概念”的比例，即概念比例；4、该大小下的集合例子，即概念示例；5、该例子所代表的概念意义，即概念说明。例如，在名语素“语素概念”中，语素个数为 16 的“同义语素集”共有 7 个，占名语素“语素概念”总数的比例为 0.35%，其中的一个“语素概念”包含了特定语素“匠哲器彦才材杰氏秀英豪贤通驥模尖”（基于可以理解和简化描述的原因，这里均省略了相应的“语素义编码”，仅以语素字的形式出现，且不排除相同字的出现），其概念意义为“有才能的人”。

表 4 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	丽毫毫令任侯僭僭窳兹匿单屋厦坏吴喷坊坳坻...	地域的简称、别称
136	1	0.05	匏柛棘稗稗篇缩芳艾芴芒芒芭芥芥芦芨芨...	草本植物
123	1	0.05	们剌妨葵岷建桂汉汜汜汜汜汜汜汜汜汜汜汜...	水域、河流名称
118	1	0.05	乌崑凰罔泉掠翡莺衡虎蟻蟻住隼雀雁雁雅鸚...	鸟类
113	1	0.05	朴衫杉杨枞松枞枞枞枞枞枞枞枞枞枞枞...	木本植物
110	1	0.05	鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑鮑...	鱼类
91	1	0.05	且伋伋伋伋伋伋伋伋伋伋伋伋伋伋伋伋伋...	人名用字
89	1	0.05	蝦拉孛蜚蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶...	节肢动物
83	1	0.05	金鏵鏵钷钷钷钷钷钷钷钷钷钷钷钷钷钷钷...	金属
71	1	0.05	元兒冉南卫召吴周唐商夏宋巴明晋曹朝杞梁...	朝代或国家名
63	1	0.05	吡吗吗咪咪咪咪咪咪咪咪咪咪咪咪咪咪咪...	制药的有机化合物
53	1	0.05	玉玊玊玊玊玊玊玊玊玊玊玊玊玊玊玊玊玊...	各种玉或玉器
52	2	0.10	卜椿瓜稗笋苜苜苜苜苜苜苜苜苜苜苜苜苜...	植物中的蔬菜类
41	1	0.05	华台圃麥蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶蚶...	山名
40	1	0.05	乞佻佻佻佻佻佻佻佻佻佻佻佻佻佻佻佻佻...	少数民族
39	1	0.05	蹇馱馱馱馱馱馱馱馱馱馱馱馱馱馱馱馱馱馱...	马类
37	2	0.10	呢帛布帛彩縠縠縠縠縠縠縠縠縠縠縠縠...	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申...	行政区划的省级单位
28	1	0.05	匏李杏杜柑枣柑柑柑柑柛柛柛柛柛柛柛...	植物的果实
27	3	0.15	丞令侯僚僚僚僚僚僚僚僚僚僚僚僚僚僚僚...	古代官名
26	3	0.15	卣卣卣卣卣卣卣卣卣卣卣卣卣卣卣卣卣...	与酒有关的器皿
24	1	0.05	冈坂阪坳坳坳坳坳坳坳坳坳坳坳坳坳坳坳...	山体的某一部分
23	1	0.05	禾秝秝稈稈稈稈稈稈稈稈稈稈稈稈稈稈...	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	匱席笆篾篾篾篾篾篾篾篾篾篾篾篾篾篾...	竹或木制成的容器
20	3	0.15	酹酹酒酹酹酹酹酹酹酹酹酹酹酹酹酹酹酹...	酒类
19	5	0.25	帘帜幟幟幟幟幟幟幟幟幟幟幟幟幟幟幟...	旗帜
18	4	0.20	内心牙肝肠肺肾胃胆胰胱脏脾脾脾脾脾脾...	动物内脏
17	4	0.20	土地坳坳坳田町畷畦畷畦畦畦畦畦畦畦畦...	土地田地
16	7	0.35	匠哲器彦才材杰氏秀英豪贤通驥模尖	有才能的人
15	6	0.30	体例则制宪彝律法矩科臬规轨辟	法律、规章制度
14	11	0.55	刀戈戟戣枪槊矛鏃鏃鏃鏃鏃鏃鏃鏃鏃鏃...	长矛类兵器
13	9	0.45	丫帮本杈条抄枝柯标株榧榧榧	树枝

12	15	0.74	丘冢圻坟墓壙寝穴窀穸阡陵	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	位品地档流等级职銜阶	社会地位、等级
9	19	0.94	粍糕酥饼饼饅馐饅	糕点
8	37	1.83	勋印戳玺章篆铃鼎	印章、图章
7	50	2.48	唾喇沫津涎痰痰	唾液
6	81	4.01	舟舡航舫船舩	船只
5	99	4.91	志标符记识	标志
4	156	7.73	橇箠鞘鞭	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

表 5 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	俏冲卤卧嘘扒摊余涮溜淩淩炊炒炖炆炮炸烩…	加工、烹饪食物的方式
34	2	0.12	僂剋毗毗呵咄咄咄哈嚷埋怨怪批摺斥派熊病诨…	责备
32	2	0.12	储厝囤垒垛堆委存存寔寔居度庑券攢囊溜滞积…	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	剝剝剖劈异渍搯搯披敞析撑毅睁解掸展扞巴启…	分开、张开
26	2	0.12	死亡亡卒危夭尽徂故歿殒殒殒殒殒殒殒殒殒…	死亡
25	4	0.25	了会博喻审悉悟惶惶懂明晓曝照省知解谗通醒…	理解、明白
24	1	0.06	冤哄啖拐捞欺绷罔肥蒙虞诈诬诳诱逛谩赚逗钓…	欺骗
23	2	0.12	仗手扒扒把持挝撓揭捉拈拈拈拈拈拈拈拈拈…	手的抓取、持握的动作
22	3	0.18	上傅刮刷图罔壁墩打拉抹抹抹抹抹抹抹抹抹…	涂抹、擦刮
21	4	0.25	临促傍湊即压向守就拢挨接毗濒薄近迫逼陆附…	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	会佻参摊撞晤寔碰见覩覩謁赶逢逦逦遭遭避	遇见
18	6	0.37	上下之到即如幸往徂徂至莅赴赶踵适造	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	下乏亏匮少少差慳拉脑欠短匮乏缺该阙	缺乏
15	8	0.49	上偏到勾及够平底馥满等致臻达齐	达到高度、目标、期限
14	11	0.67	号咄吼呐嗷呼哮嗥唱喊喝噪嚷嚷	大声地叫喊
13	17	1.04	冶化化泮溶汤烱焊熔融铄销铄	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	倒北破砸胜负败败败败输	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	剽劫夺抢掠掠擄越逼	掠夺、劫取
8	39	2.39	假借租租赁赁赁赁赁	借取、租取
7	45	2.76	抽生疯结茁起长	植物生长
6	59	3.62	捭斋施舍赈食	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	扶掠箠鞭	鞭打
3	163	10.00	冤屈讫	感到冤枉
2	305	18.71	俘虏	俘虏
1	645	39.57	萌芽	植物发芽
不限	1630	100	N/A	N/A

表 6 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	亮光同昉明晃晒晞晞晞晶晔曙肚朗杲灼灿炅炜…	光亮、耀眼
31	1	0.18	侗倥傯傻兀冥瞶悖悖悖悖悖悖悖悖悖悖悖悖…	愚蠢的、糊涂的
27	1	0.18	举俗全到合周圉完尽彻总普毕洽浑溥满漫熨盛…	完整的、完备的
26	2	0.36	令佼俊劭嘉姝姘媚媚媚媚媚媚媚媚媚媚媚…	美好的
25	2	0.36	丰丽俊俏俏冶妍妖姘媚媚媚媚媚媚媚媚媚…	样貌美丽的
24	1	0.18	乐休哈娱快怵欣忻怏怏怏怏怏怏怏怏怏怏…	欢喜的、快乐的
23	2	0.36	伉伟倜劬勇勐壮悍敢竭臬武激烈狂猛矫虎骁骄…	勇猛的、强健的
22	2	0.36	汪沅洞泓泚泚冷冽冽冽冽冽混混混混混混混…	形容水流
21	2	0.36	丰优博厚夥夥富富广敞旺殷洋浩海盛腆腆趁饶足…	丰富的、丰盛的
20	2	0.36	冷暗嘿必寂寂寥幽恬恬恬恬恬恬恬恬恬恬恬…	寂静的、冷清的
19	3	0.55	丑倦傴仄堵怵恶悒悒悒悒悒悒悒悒悒悒悒…	厌恶的、厌烦的
18	3	0.55	凶恶悍惨警暴横横残狰狞狰猖粗藏蒙酷鸩	残暴的、凶狠的
17	3	0.55	乖俐卓巧惺惺慧敏智猴睿神秀精聪颖鬼	乖巧的、机灵的
16	4	0.73	凛寒怯怵懔懔恟恟恟恟恟恟恟恟恟恟恟…	胆怯的、害怕的
15	5	0.91	忠恂恳悃恹恹恹恹恹恹恹恹恹恹恹恹恹…	忠诚的、真诚的
14	3	0.55	复曼杳洞缅辽远迢迢迢迢迢迢迢迢迢迢…	距离遥远的

我们首次将“语素概念”作为节点刻画构词过程中基本意义单元之间的结合情况。如图2所示，图中的每一个矩阵节点代表一个“语素概念”，节点的大小代表“语素概念”中的各个语素（已确定了语素义）在构词过程中贡献的能产性的加和，而节点之间的边代表两个“语素概念”中的某两个语素依确定的语素义参与了构词过程，参与次数体现为边的权重，即边越粗，表明两个“语素概念”结合的可能性越大。该图依据前述4199个名、动、形语素的“语素概念”和52108个二字的语义构词知识绘制，充分、客观地反映了汉语世界中的语义基元的能产性分布情况。需要指出的是，由于全局数据覆盖面大，图中相对较细的边由于像素问题并不能清晰显示，需要在机器上放大后才能拾取。

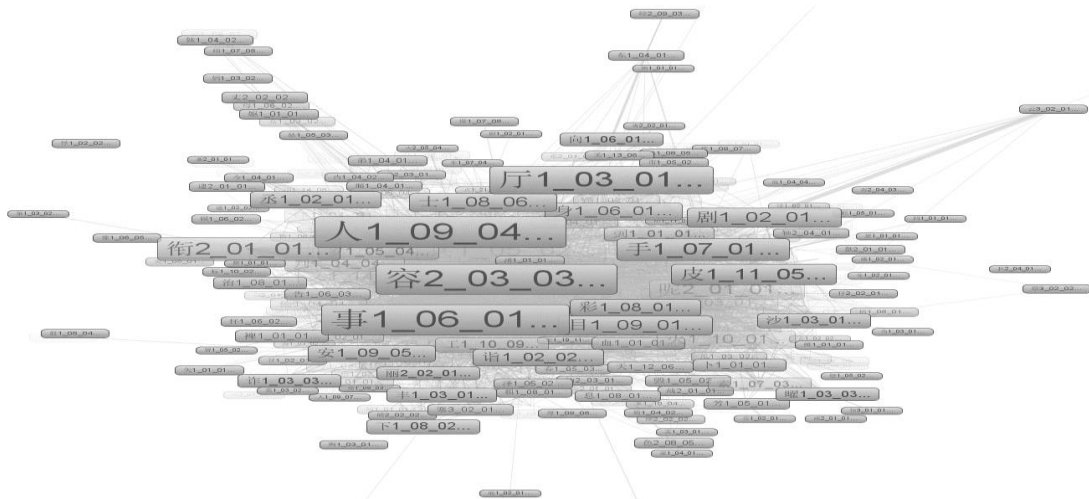


图2 基于“语素概念”的构词能产性示例

通过对可视化数据的考察，易发现{人1_09_04, 号2_13_08, 民1_05_02, 郎1_05_02}表示“某种人”、{事1_06_01, 务1_05_01}表示“某种事”、{厅1_03_01, 厝1_03_03, 厦1_02_01, 宅1_01_01, 宇1_05_01, 室1_06_01, 窠1_01_01, 窠1_02_01, 寓1_04_02, 居1_08_02, 屋1_03_01, 屋1_03_02, 廛1_01_01, 房1_07_01, 房1_07_02, 斋2_01_01, 庖_02_01, 舍2_05_01, 舍2_05_01}表示“房屋住宅”等“语素概念”在构词过程中具有显著的高能产性，表明这类意义属于常用的基本概念。当然，具体、微观一些，也可以考察这些“语素概念”中的特定语素（已确定了语素义）的能产性和搭配特征。这是以往基于字符、语素类、构词规则统计或语义构词个例剖析等不能得到的语言学结论，也显示了对汉语构词做全局性语义挖掘与可视化分析的比较优势。

在未来，语义构词模式的分析可以加深对词的结构和意义的理解，并用于未登录词识别和意义判定。在机器学习领域，这也是一项极其重要的特征和任务，而我们开发的基于语义基元的构词能产性数据给出了语义构词的转移概率，能为相关的算法开发提供支持。

5 结语

综上所述，我们提出了一种探寻汉语语义基元和分析语义构词的新的方法和视角，并表现出明显的优势：一、语素及其意义作为抽象概念难以表达、计算，借鉴 WordNet 理论引入“同义语素集”做表征，使得原本抽象的“语素概念”有了相应的承担实体，架起了汉语的语素及其意义和构词分析的天然联系，契合了汉语的意合特征。在此之上，借鉴生成词库理论和面向对象思想，形成“语素概念体系”，这也是对汉语世界中的语义基元的系统描述；二、建立在这些基础上的汉语构词分析，在全局性语义分析、数据挖掘等方面也有新的进展；三、从语言知识工程的角度看，面向《现汉》中的全部语素和二字词，在“语素概念”提取等环节采取人机结合、自底向上的策略，尽量排除主观因素的干扰，这些做法也保障了研发数据的覆盖度和完备性，提升了语言资源建设的质量。

这些创新的思路、做法以及获得的数据成果,在人文领域和计算应用等方面都有潜在的应用价值。前者如(电子)词典编纂与出版、汉语教学、语言本体研究等,对于后者,我们的工作已经有初步的验证,在汉语未登录词的词义知识表示与语义预测^[34]、汉语词语语义相似度计算^[35]等方面进行了探索和尝试。

在此前阶段,汉语的语义构词分析主要针对词的本义,但部分合成词的词义存在转义、隐喻等现象,如何有效表达和处理这类现象,将是后续工作的一项重点。此外,“语素概念”及其体系的考核、优化以及多字词的词义知识表示的拓展也在扎实推进中。在此基础上,我们希望尽快推出包含全集数据和 API 接口的北京大学《汉语概念词典》(英文名称 Chinese Object-Oriented Lexicon, 简称 COOL)。

参考文献

- [1]赵元任. 中国话的文法[M]. 丁邦新译. 香港: 香港中文大学出版社, 1980
- [2]陆志韦. 汉语的构词法(修订本)[M]. 北京: 科学出版社, 1964
- [3]朱德熙. 语法讲义[M]. 北京: 商务印书馆, 1982
- [4]王洪君. 汉语语法的基本单位与研究策略[J]. 语言教学与研究, 2000, 02: 10-18
- [5]董秀芳. 词汇化: 汉语双音词的衍生与发展(修订本)[M]. 北京: 商务印书馆, 2011
- [6]徐通锵. 核心字和汉语的语义构词法研究[J]. 语文研究, 1997, 03: 2-16
- [7]符准青. 词义和构成词的语素义的关系[J]. 辞书研究, 1981, 01: 98-110
- [8]周荐. 论词的构成、结构和地位[J]. 中国语文, 2003, 02: 148-155, 192
- [9]詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[M]. 北京: 清华大学出版社, 2000
- [10]周强. 汉语短语的自动划分和标注[J]. 中文信息学报, 1997, 01: 1-10
- [11]JiangGuo, WanxiangChe, Haifeng Wang. A Universal Framework for Inductive Transfer Parsing across Multi-typed Treebank[C]. In Proceedings of the 26th International Conference on Computational Linguistics (Coling 2016).
- [12]袁毓林. 语言的认知研究和计算分析(增订本)[M]. 北京: 商务印书馆, 2014
- [13]苑春法, 黄昌宁. 基于语素数据库的汉语语素及构词研究[J]. 世界汉语教学, 1998, 02: 8-13
- [14]亢世勇, 李毅, 孙道功, 等. 汉语系统语料库的建设与词典编纂[C]//上海辞书学会. 2004年辞书与数字化研讨会论文集. 上海辞书学会, 2004:7
- [15]Ya-Min Chou. Hantology: The Knowledge structure of Chinese Writing System and Its Applications[D]. Taiwan: National Taiwan University, 2005.
- [16]董振东, 董强, 郝长伶. 知网理论发现[J]. 中文信息学报, 2007, 21:4
- [17]Fellbaum C. WordNet: An Electronic Lexical Database[M]. Mass: MIT Press, 1998
- [18]Pustejovsky, J. The Generative Lexicon[M]. Mass: MIT Press, 1995
- [19]GradyBooch, Robert A. Maksimchuk, Micheal W. Engle, etc. Object-Oriented Analysis and Design with Applications, 3rd Edition[M]. Addison-Wesley Professional, 2007.
- [20]Pesina S, Solonchak T. Semantic Primitives and Conceptual Focus [J]. Procedia - Social and Behavioral Sciences, 2015, 192:339-345.
- [21]Sapir E. The Expression of the Ending-Point Relation in English, French, and German[M]// The expression of the ending-point relation in English, French, and German. Linguistic Society of America, 1932
- [22]Sapir E. Grading, A Study in Semantics[J]. Philosophy of Science, 1944, 11:93-116.
- [23]Alice Morris. Editorial note to Sapir and Swadesh's The Ending-Point Relation[Z], 1944
- [24]Wierzbicka, A. Semantic Primitives[M]. In J Frawley (ed.), International Encyclopedia of Linguistics (2nd ed). New York: Oxford University Press, 2003
- [25]尹斌庸. 汉语语素的定量研究[J]. 中国语文, 1984 (5): 340.
- [26]吕叔湘. 《现代汉语词典》编写细则(修订稿)[M]. 《现代汉语词典》五十年. 北京: 商务印书馆, 2004.
- [27]刘叔新. 汉语描写词汇学[M]. 北京: 商务印书馆, 1990
- [28]傅爱平. 汉语信息处理中单字的构词方式与合成词的识别与理解[J]. 语言文字应用, 2003, 04: 25-33
- [29]杨梅. 现代汉语合成词构词研究[D]. 南京: 南京师范大学, 2006
- [30]YoshuaBengio, RéjeanDucharme, Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research. 2003, 03:1137
- [31]Plag, I. Word-formation in English[M]. Cambridge, UK: Cambridge University Press, 2003
- [32]现代汉语频率词典[M]. 北京语言学院, 语言教学研究所编. 北京: 北京语言学院出版社, 1986
- [33]Harington, Dennis. Input-drive language learning[J]. Studies in second language acquisition, 2002,24:261-268
- [34]田元贺, 刘扬. 汉语未登录词的词义知识表示及语义预测[J]. 中文信息学报, 2016, 06: 26-34
- [35]康司辰, 刘扬. 基于语义构词的汉语词语语义相似度计算[J]. 中文信息学报, 2017, 01: 94-101



刘扬 (1971-), 博士, 副教授, 主要研究领域为语言知识工程、中文信息处理. E-mail: liuyang@pku.edu.cn

林子 (1997-), 本科生, 主要研究领域为应用语言学、语言知识工程、中文信息处理. Email: zi.lin@pku.edu.cn

康司辰 (1993-), 硕士生, 主要研究领域为应用语言学、语言知识工程、中文信息处理. E-mail: 1008_frank@sina.com