基于平行周遍原则的汉语未登录词的知识表示与预测*

康司辰[1]，虞梦夏[1,2]，刘扬[1]

（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 中国语言文学系，北京 100871.）

摘要：汉语未登录词的知识表示与预测，包括词性、构词结构、词义等项目，是计算语言学领域中的基础性问题。该文依据“平行周遍”原则，从现有的语义构词知识中提取“平行条件”，将未登录词潜在的构词因素与这些“平行条件”进行适应性匹配，从而对其知识表示进行相对完整的预测。该方法将新的语言学理论与未登录词的理解应用问题结合，取得了显著的效果，其解释能力、便捷性和精细程度优于此前方法。这些研究，除了在自然语言处理领域有实用价值，也有望推动词典编撰、语言研究与教学等人文领域的进展。

关键词：汉语未登录词 平行周遍条件 语义构词 知识表示 知识预测

中图分类号：XXXXX　　**文献标识码：**X

Knowledge Representation and Prediction of Chinese Unknown Words by Using Parallel Conditions

KANG Sichen[1], YU Mengxia[1,2], LIU Yang[1]

(1. Key Laboratory of Computational Linguistics (Ministry of Education), Peking University, Beijing 100871, China;
2. Department of Chinese Language and Literature, Peking University, Beijing 100871, China)

Abstract: Knowledge representation and prediction of Chinese unknown words, including which of parts of speech, word-formation structure and word meaning, is a fundamental problem in computational linguistics. According to the Principle of Parallel Circumference, this paper extracts Parallel Conditions from the existing semantic word-formation knowledge, matches the potential word-formation factors of unknown words with these Parallel Conditions adaptively, and then makes a relatively complete prediction of their knowledge representation. This method combines the new linguistic theory with the understanding and application of unknown words, and achieves remarkable results. Its explanatory ability, convenience and precision are better than those of previous methods. These studies not only have practical application value in the field of natural language processing but also are expected to promote the progress of computational lexicography, language research and teaching and other humanities fields.

Keywords: Chinese Unknown Words; Parallel Conditions; Semantic Word-Formation; Knowledge Representation; Knowledge Prediction

0　引言

随着社会交往和科技手段的发展，大量新的可独立运用的汉语片段出现在语料中，给信息处理带来了极大的麻烦。这些在语料中出现频繁、结合紧密并且还没有进入词典的词，被称为“未登录词”。这些未登录词的知识表示与预测，包括词性、构词结构、词义等项目，是计算语言学领域中的基础性问题。

未登录词的知识表示与预测工作，在汉语语言学界处于探索阶段。周洪波[1]对此做过理

*收稿日期:　　定稿日期:
基金项目：国家社科基金一般项目（16BYY137）、国家社科基金重大项目（18ZDA295，12&ZD119）

论性探讨，认为：“潜在的未登录词的预测应该是基于语素与语素之间的排列组合；新义的预测应着眼于语素义与语素义之间的排列组合。”他认识到语素在未登录词预测中的重要作用，但没有将语素“排列组合”的实质和应用讲清楚。换句话说，未能在语义构词层面上解决汉语语素的聚合与组合问题，不能清晰地说明未登录词的“新义”与语素义“排列组合”规则之间的联系。王东海[2]等对未登录词语义构词的条件和依据做了进一步研究，提出了 8 条依据并做了一些预测。在这些依据中，考虑到了构词义场、隐喻、方言等因素，对语素义的聚合问题做了初步探讨，但没有注意到构词结构所反映的语素组合问题。此外，该方法由于可获得的语义构词知识极其受限，只能在部分语义场上做出示例性的预测，这种情况在自然语言处理领域难以广泛采用。

在计算语言学界，对未登录词的知识表示与预测，近来主要集中在“语义类别”方向上[3-5]，而此类预测往往不能满足精细化的词义表达需求。至于其他方法，基于概念图[6]的知识表示又过于复杂，不够精简。注意到这些问题，在稍早的工作中，田元贺等[7]给出了一种新的词义表达方式与预测方法，该方式、方法建立在汉语的语义构词分析的基础之上，其词义知识表示简单有效，但所采用的神经网络系统的解释性不够，且性能存在一定的提升空间。瞿健菊[8]在此基础上，对未登录词的知识表示开展了进一步研究，但由于缺乏系统性的语素义知识，无法利用语素的聚合信息形成覆盖词典的类比条件，也无法对这些类比条件进行有效合并从而找到语义构词规则，其预测的未登录词数据规模依然偏小。

另一方面，在相对宏观的语言学理论层面上，陈保亚等[9]对于语言单位的“组合性”与“聚合性”问题做了一般性的探索。在徐通锵[10]字本位的研究基础之上，陈保亚提出了基于“平行周遍”原则的“平行周遍对比”方法，指出对满足平行条件的语言单位的组合可进行规则性地理解。该方法与“同形替换法”都运用了类比的方法，但是相较于后者，强调系统地利用词性、构词结构、前后构词成分等信息作为“平行条件（Parallel Conditions，简称 PC）”的特征，对类比的实质阐述得更为详实。该理论还有需要完善的地方，比如，周上之[11]、金朝炜[12]认为“平行条件”的确立缺乏实证数据的支持，主观性较强，无法形成系统性的“平行条件”集合，而这恰恰是“平行周遍对比”方法需要着力解决的一大问题。

对以上视角、观念、方法的借鉴，构成了我们对汉语未登录词知识表示与预测的新的语言学理论基础。

在可用的知识库方面，我们对《现汉》词依据义项先期开展了系统化的语义构词分析研发工作[13-14]。通过大规模的专家标注，获取了《现汉》词的语义构词知识，并对《现汉》中语义构词的所有构词成分（即语素）的义项，采用“同义语素集”来表征“语素概念”，建立起结构化的“语素概念系统”，旨在获得汉语世界中详尽、完备的语义基元集合。这些相对完整、丰富的语义构词知识、语素概念知识，在很大程度上，可以为“平行周遍对比”方法提供批量的数据支持，能够相对客观地找出全体“平行条件”的集合，从而更好地解决汉语未登录词的知识表示与预测问题。

1　汉语的语义构词描写

1.1 汉语的构词结构分析

受西方结构主义的影响，汉语语言学家在 20 世纪中期开始了对“词”内部结构的语法分析，如赵元任[15]、吕叔湘[16][17]、高明凯[18]、朱德熙[19]等，强调各个构词成分之间的语法关系。在稍后的时间，刘叔新[20]、傅力[21]、周荐[22]等将视线聚集到“复合词”内部的语义关系，力图发掘并利用汉语独特的“意合”特征。

我们希望兼顾这两大流派的优势，并结合自然语言处理的实际需要，形成面向计算应用

的构词结构体系。杨梅[23]参考之前“语法构词”和“语义构词”的诸多体系，对此前的标签集进行了进一步的删改与讨论，形成了一套以“语法构词”标签为主的构词体系，包括：定中式、数量式、方位式、名量式、状中式、联合式、述宾式、述补式、主谓式、连谓式、复量式、介宾式、虚配式、截取式、指量式、数构式、重叠式、附加式，共 18 类。在此基础之上，我们通过对《现汉》词的抽样考察，删去了一些实例过少或计算意义不大的构词标签，如截取式、虚配式、指量式、数构式，将“附加式”分为了“前附加”与“后附加”，并增加“单纯式”以表示缺乏内部结构的单纯词，最终形成了一套面向自然语言处理应用的标签集。

在此规范基础之上，对《现汉》中的 52108 个二字词，我们按义项区分进行了构词结构分析与标注。为方便表述，我们称每个词及其一个义项为一个词项（同理，在后文中，将每个语素及其一个义项称为语素项，以避免可能的歧义）。对二字词的每个词项，请三位专家对其进行结构标注，标注结果两人及以上的一致率达 93.46%。

1.2 汉语的构词成分分析

构词结构体现了构词成分之间的组合关系，同时，这些构词成分所构成的义场，体现了构词成分之间的聚合关系。符淮青认为[24]：“语义单位依次包括：句义、词组义、词义、语素义。”语素义是汉语语义构词中的基本单位，汉语的构词大体上是以语素为基础的。通过对“同义语素”的聚类，并对得到的聚类做高层次的泛化与分类，可以得到汉语语素的义场，从而解决构词成分基于意义的聚合问题。

参考 Wordnet 采用“同义词集”的方式来表征词义概念，我们采用“同义语素集”来表征“语素概念”。对《现汉》中所有的基本构词单位——语素项进行语义相似度考察，来获得可信的“同义语素集”。为了便于知识工程开展，避免语素项与语素发生混淆，我们为每一个语素项赋予唯一的“语素义编码”，如“雨 2_02_01”式样。通过批量数据处理，我们获得了《现汉》中 8514 个汉字（包括繁体、异体字）的 20855 个语素项信息。考虑《现汉》释义文本的特点，在语义相似度计算方法的选择上，我们采用“字共现”模型，对于某一特定语素项的释义文本，按照它与其他语素项的语义相似度值将这些语素项降序排列，并按设定阈值，将意义相近的语素项推荐给专家遴选。在人工校验的基础之上，经过反复核对、补充、过滤，语言知识工程覆盖《现汉》中所有的语素项，最终形成了 4450 个“同义语素集”（或称“语素概念”），这也是汉语世界的语义基元。如表 1 所示，《现汉》中所有表示“雨”的语素经过上述操作、整理，形成了一个特定的“语素概念”，其中包括了 6 个不同的语素项。

表 1 表示“雨”的语素概念

语素义编码	《现汉》中的语素释义
雨 2_02_01	<名>从云层中降向地面的水。云里的小水滴体积增大到不能悬浮在空气中时，就落下成为雨。
霡 1_01_01	<名><书>〖霡霂〗(màimù)小雨。
澍 1_01_01	<名><书>及时的雨。
霈 1_02_01	<名><书>大雨：甘～。
溦 1_01_01	<名><书>小雨。
霖 1_01_01	<名>霖雨：秋～丨甘～。

进一步的，我们对这些“语素概念”建立层次和网络结构，形成“语素概念系统”，增添面向应用的语义推理和计算能力。参考 WordNet 体系，我们依上下位关系对名语素概念做结

构化梳理。例如，表示“雨”的语素概念{雨 2_02_01、霖 1_01_01、澍 1_01_01、霈 1_02_01、潎 1_01_01、霖 1_01_01}和表示“潮汐”的语素概念{潮 1_03_01、汐 1_01_01、汛 1_02_01}均属于“物-具象物-非生物-天然物-水”结构下的结点。对于其他类别的语素概念，借鉴生成词库理论，在形语素概念和动语素概念中分别建立起相应的结构。大体说来，可以认为，动语素表达了名语素所指事物的事件，形语素表达了名语素所指事物的性质。由此，名、动、形等不同语素概念是大致对应的、同构的，形成同语素概念内的聚合关系以及跨语素概念间的组合关系。

在构词结构标注和语素概念系统构建完成后，我们对《现汉》词按义项区分对构词成分标注其在《现汉》中的语素义，这样一来，每个构词成分就获得了相应的“语素义编码”，也就绑定了相应的语素概念，并在语素概念体系中占有唯一确定的位置，为词义的表达和应用奠定了扎实的基础，提供了丰富多样的“语义构词”知识[13][14]。

1.3 汉语的语义构词描写

符淮青[25]等指出：“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。这种表示方法简单、直观，能够全面、充分地反映构词成分对整体词义的贡献。例如，在“植树”中，“植”的语素义为“栽种”，“树”的语素义为“木本植物的通称”，以“述宾”结构关系对构词成分进行语义上的组合，能够较为准确地反映“栽种树木”的词义。以“植树”为例，我们得到的语义构词知识如表 2 所示。

表 2 语义构词知识示例

例词	词性	构词结构	前语素义	后语素义
植树	动词	述宾结构	植 1_04_01	树 1_04_01

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，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[26]曾给出包括八种形式的意义结构体系，该体系分类详细，词义知识的表示和获取有较大的难度。从实际应用的角度出发，我们采取了一种相对简单、方便计算的意义结构形式，如表 3 所示。在这里，暂时只考虑词的字面意义，即本义。关于词的引申义问题，见此前的相关工作[27],本文不做进一步探讨。

表 3 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	词例
00 型	词义与前后语素义相关性均较低	单纯	沙发、名堂
01 型	词义只与后语素相关性较高	前附加	老虎、仔细
10 型	词义只与前语素相关性较高	后附加、名量	忘却、船只
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、方位、复量	植树、红旗

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”的特定排列，其内容和顺序基本由构词结构决定。还以“植树”为例，其“意义序列”为“<植 1_04_01,树 1_04_01>”。此外，允许在需求中依据强调与约定改变序列顺序，以表达计算应用的灵活性，如“<树 1_04_01，植 1_04_01>”也可认为是一个合法的“意义序列”。

2 汉语语义构词平行条件的获取

基于“平行周遍”原则的“平行周遍对比”方法力图把语言单位间类推的实质描写清楚并给出严格的标准，其核心在于“平行条件”的确定。陈保亚[28]对于“平行条件”的性质做

了详细的阐述:“被替换的部分具有平行特征;在被替换部分保持平行特征的前提下,组合关系平行;整个组合在分布上平行。” 从这三条“平行条件”性质中,我们看到,“平行周遍对比”相较于之前的研究,强调语法和语义两个层面上的要求,在这些要求之下,推导单位的平行条件是有约束的,并且有效地反映了类推的规则。相同平行条件下的词是满足于相同的语义构词规则的。对于未登录词而言,只要与既有的特定平行条件进行适应性匹配,就可以找到其语义构词规则。这种看法,为未登录词的知识表示预测带来了很大的便利。

平行条件的获取,很大程度上依赖于平行特征的选取。陈保亚[29]认为平行特征可以在语音、语法、语义等层面展开,但也没有给出系统性的准则。在本研究中,我们实证性地选定“词性”、“构词结构”、“前语素义”、“后语素义”作为特征,用来归纳平行条件。基于之前的汉语语义构词研究成果,我们将在《现汉》中具有相同特征取值的词归拢到一起,自然地形成了平行条件,这些平行条件构成了平行条件集合。例如:“民办、民营、私营”等词的词性为“形容词”,构词结构为“状中”,前、后语素义分别表达“民间、私有”、“管理、经营”的含义。通过对词的以上这些特征取值的归纳,便获取到了如表 4 最后一个记录所示的一个平行条件。显然,表 4 里总共列举了 6 个不同的平行条件,“词性”、“构词结构”、“前语素义”、“后语素义”即为这些平行条件下的平行特征。

表 4 二字词的平行条件用例示意

平行条件				满足平行条件的《现汉》词
词性	构词结构	前语素义	后语素义	
名词	定中	农历一季度月的名称	季节	仲冬、仲夏、仲春、仲秋、孟冬、孟夏、孟春、孟秋、季冬、季夏、季春、季秋
名词	定中	中国地方简称	戏剧	京剧、京戏、川剧、晋剧、桂剧、楚剧、沪剧、湘剧、滇剧、琼剧、粤剧、苏剧、豫剧、赣剧、闽剧、陇剧、黔剧
动词	联合	集中、汇集	集中、汇集	会聚、凑集、团聚、团聚、屯聚、屯集、归总、拼凑、攒聚、攒集、汇集、聚会、聚集、萃聚、集聚、集萃、麇集
动词	述补	说明、解释	清楚、明白	标明、申明、说明、阐明
形容词	后附加	安静	形容词后缀	寂然、悄然、阒然、默然
形容词	状中	民间、私有	管理、经营	民办、民营、私营

构词结构反映了语素项间的组合关系,语素义体现了语素概念内的聚合关系,相应的平行条件表达了汉语世界中语素成词的客观、现实规则。易见:平行条件是对语义构词知识的形式化描写,处于同一个平行条件下的词,表达的语义往往相同或相近,其形成的词集刻画了确定的概念含义。通过对不同平行条件的探寻,实际上是在汉语的语素层面上,对词的“词性”、“构词结构”、“前语素义”、“后语素义”的再组合。这些先验性的词义合成规则,对于未登录词的知识表示与预测具有显著的指导价值。通过对现有语义构词成果的此类归纳,我们在《现汉》词中,一共得到了 36141 个平行条件。

获取的这些平行条件,有助于进一步探寻汉语构词中的语法与语义状况。

我们在“词性”、“构词结构”的层面上,将平行条件的分布与二字词的分布进行对比,研究不同语义构词规则的覆盖情况。在二字词的词性中,“名词”、“动词”、“形容词”占比最高,在构词结构中,“联合”、“定中”、“状中”占比最高。我们对上述词性、构词结构进行分析,二字词的词性-结构分布统计见图 1,平行条件的词性-结构分布统计见图 2。可以看到,二字词的词性-结构分布和平行条件的词性-结构分布大体一致。这说明,对于特定词性-结构的词来说,这些词所对应的平行条件数量没有大的差异,换句话说,汉语词的构词规则数量在词性和构词结构的联合分布上是基本均匀的。这也是通过对现有语义构词成果的实证分析,得到的一些新的看法。

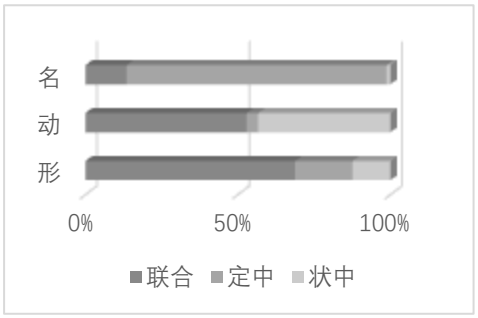


图 1 二字词的词性-结构分布

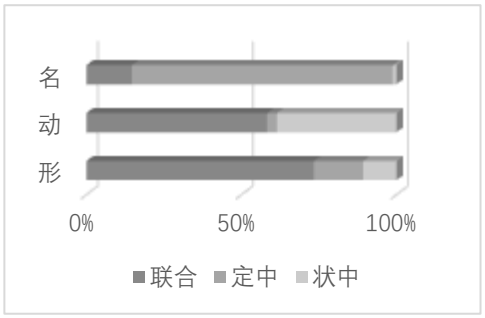
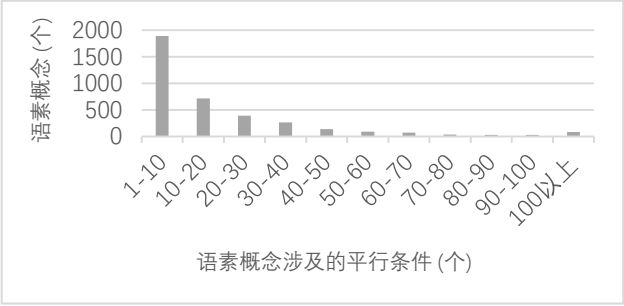


图 2 平行条件的词性-结构分布

除了上述语法层面的探讨，我们也在语义层面上进行了研究。例如，对这 36141 个平行条件中的语素概念进行分析，可以观察特定语素概念所表达的义项在汉语世界中的重要性。如果一个平行条件中涵盖某个语素概念，则称该语素概念涉及了该平行条件。我们统计每个语素概念所涉及的平行条件，其分布统计见表 5。从表中可以看出，大体上，语素概念所涉及的平行条件数量成快速递减的趋势，说明那些覆盖相对较多构词规则的语素概念仍是少数，而这些语素概念，往往表达了汉语中较为重要或显著的语素级义项。

表 5 语素概念涉及平行条件统计



以涉及平行条件数量排名靠前的语素概念及其所表达的义项为例，情况见表 6。这些语素概念涉及更多的平行条件，在更多的语义构词规则中出现。而那些相对不常用的语素概念虽然数量较多，但涉及的语义构词规则相对较少。这在一定程度上，反映了语素概念的构词能力。需要说明的是，在表 6 中，基于可以理解和简化描述的原因，在语素概念示例里均省略了相应的“语素义编码”，仅以字代表它的某个语素项，不排除相同字的出现，表中相同的字表示的是不同的语素项。

表 6 语素概念涉及平行条件示意

语素概念涉及平行条件数量	语素概念含义	语素概念示例	所含语素个数
229	表示聚集、汇集	醵空归麋会譧并团搭拼捏攒接挤注撮汇汪总丛辏簇匀凑存槚屯综纠集聚萃致钟	34
187	人称代词	卬予余吾咱侬俺俺咱咱侬朕我我菲窃敢辱忝卿卿妾仆晚臣钦尊职贫愚愚己舍家鄙敝贱小拙奴自身	42
185	表示道路	垄垣峤嶝径歧涂畛程行路蹊远途逵道阡陌	18
185	表示军队	兵军士师戎旅卒帅兵	9

通过对系统化、实证性获取的诸多平行条件的分析，不难看出，这些平行条件中蕴含着

大量的语法、语义甚至外部世界知识，这对汉语研究以及理解应用都是大有裨益的。

3 汉语未登录词的知识表示与预测

对于一个二字未登录词，找到其前、后字分别可能涉及的语素项集合 $M_1=\{m_{11}, m_{12}, \cdots\cdots, m_{1n}\}$、$M_2=\{m_{21}, m_{22}, \cdots\cdots, m_{2m}\}$，它们构成的全部组合有$|M_1M_2|=|M_1|*|M_2|= n * m$ 个，对于其中的任意组合 $m_{1i}m_{2j}$，如果 m_{1i} 处于某一条平行条件 P 的前语素义中（即：选定了语素项，依据“语素义编码”的绑定，也落入相应概念语素中。下不赘述），m_{2j} 处于相同平行条件 P 的后语素义中，那么我们称构成这个未登录词的语素组适应性地匹配上了平行条件 P。理论上，$m_{1i}m_{2j}$ 有可能对应多个平行条件，相应的，未登录词则可能符合多个语义构词规则，这些规则也对未登录词的多义性问题了提供帮助。

我们以《人民日报》（1998 年 1 月）分词并标注词性的语料为例，未登录词指那些不包含在《现汉》中但出现在语料中的词。其中，词性标注为“名词”、“动词”、“形容词”的二字未登录词共 5255 个。在这 5255 个未登录词中，有 1990 个约四成的词可以直接从现有平行条件集合中直接找到，获得知识表示与预测，我们以这 1990 个未登录词作为考察对象。

对于那些不满足既有平行条件的词，则需要做进一步的扩展研究，比如采取构词结构下的语素概念相似度分析与判定策略。实际上，未登录词的产生可能伴随随着新的构词模式产生[30]，因此，对于这一部分未登录词的研究，有助发现新的语义构词模式，即新的平行条件。这些新的平行条件也可以追加到既有的平行条件集合中去，此类工作，在语言知识工程上应该是迭代进行的。

3.1 未登录词的词性预测

考察满足既有平行条件的 1990 个未登录词，以语料中的词性标注作为标准答案，我们正确预测出了其中 1683 个词的词性，具体结果如表 7 所示。可以看到，本文提出的基于“平行周遍”原则的方法，在名词上正确率最高，动词次之，形容词最低。在总体上，正确率达到了 84.6%。

表 7 未登录词词性预测统计

词性	预测词数	预测正确词数	预测正确率
名词	980	847	86.4%
动词	933	776	83.2%
形容词	77	60	77.9%
总计	1990	1683	84.6%

在未登录词的词性预测方面，基于统计的方法主要依赖于词的外部知识，部分方法也用到了内部知识[31]。基于统计的方法目前得到的最好结果为 85.48%[32]。本文方法仅仅利用了现有语义构词成果，不依赖于任何外部语料，其结果略逊于统计方法，但是本文方法更加简洁自然，采用了易于理解的语言学规则，不需要复杂的算法和海量的语料。此外，张海军[31]表明基于统计的方法如果加入词的内部信息，可以提高词性预测的准确率。将规则方法和统计方法结合，在未登录词词性预测上达到更好的效果，这也是我们未来的工作。

如前所述，本文方法也可以预测出未登录词可能存在的多个词性，解决实际应用中的兼类问题。比如未登录词“著书”，在预测结果中就有“名词”和“动词”两个词性。之所以会出现这种兼类的情况，是因为在《现汉》中，“著作”一词的前、后语素与“著书”的前、后语素分别处于相同的语素概念中，满足相同的平行条件。“著作”是兼类词，相应的，“著书”也有资格作为兼类词。在未登录词的兼类问题上，本文方法也存在着不尽如意的地

方。比如“造成”一词虽然正确地预测出了“动词”词性，但是也错误地预测出了“名词”词性，其原因在于“造化”和“造成”满足相同的平行条件，“造化”一词具有名词性，但显然“造成”不具有名词性。对于这一类错误预测的处理，还需做进一步的研究与修正。

3.2 未登录词的构词结构预测

对上述 1990 个未登录词，按本文方法做出的构词结构预测结果如表 8 所示。未登录词的构词结构预测，因为相关标注语料和知识库系统的欠缺，以往的工作鲜有涉及，而且难有标准答案做评判。本文仅以未登录词的构词结构预测分布略作分析，各构词结构的分布与我们标注的《现汉》词的分布大致相同，见表 8。据此，大致上可以认为构词结构的预测是基本准确的。

表 8 未登录词构词结构预测分布及对比

构词结构	数量	结构预测分布	《现汉》结构标注分布
定中	724	36.38%	40.41%
联合	459	23.07%	20.76%
述宾	275	13.82%	14.22%
状中	207	10.40%	7.38%
连谓	98	4.92%	3.24%
述补	76	3.82%	1.57%
后附加	61	3.07%	4.19%
主谓	30	1.51%	2.21%
介宾	15	0.75%	0.25%
重叠	12	0.60%	0.48%
复量	7	0.35%	0.04%
前附加	7	0.35%	1.10%
单纯	6	0.30%	3.59%
方位	6	0.30%	0.33%
名量	4	0.20%	0.13%
数量	3	0.15%	0.11%

3.3 未登录词的词义知识表示与预测

本文方法能够在新的知识表示下，对未登录词给出相对精准的词义预测结果。以未登录词“支局”为例，在确定了其所对应的平行条件之后，可以迅速地输出其语义构词知识表示：词性为“名词”、构词结构为“偏正”，前语素义为“支 2_02_01”，后语素义为“局 2_04_02”。这种表示结果，一方面人很容易理解，另一方面也方便基于构词结构和成分意义的计算应用。

对于多义词，本文方法也同样适用。例如，“著书”一词对应多个平行条件，依据每个平行条件，均可以对“著书”进行词义知识表示的预测。结果见表 9，表明“著书”一词有可能是一个潜在的多义词，可输出不同的表示结果。其多义性源于词性、构词结构以及构词成分意义上的些微差异。比如，在词性上，可以描述“著书”是一个客观存在的“名词性”物件或是一种“动词性”行为；在“动词”中，通过对构词结构的进一步区分，可以表达“著书”一词是描述抽象的写作行为，还是表达具体写作某种特殊文体的意义；在最细微的相同构词结构“述宾”之下，还可以表达其写作文体的不同。能够预测并表达词义的潜在的细微差别，这无论对计算问题，还是对语言学本体研究都有一定的参考价值（当然，该过程也有

可能引入错误，和本文 3.1 中所述的情况类似，其原因和解决办法在此不做赘述）。这也是先前的研究由于方法和数据所限而无法达到的效果。

表 9 “著书”词义知识表示与预测结果

平行条件				满足条件的《现汉》词
词性	构词结构	前语素义	后语素义	
名词	联合	写作	写字；记录	写作
动词	述宾	写作	装订成册的著作	修书
动词	述宾	写作	文件	修书
动词	联合	写作	写字；记录	写作、撰写、撰著、编写、编撰、编著、著作

此外，本文方法还能够迅捷地获取与未登录词意义形同或相近的所有词及其知识表示，这对理解、处理未登录词的语义是十分有益的，而以往的方法同样很难达到这种效果。比如，未登录词“静寂”满足如表 10 所列平行条件，我们在获取其知识表示的同时，顺便也得到了《现汉》中具有相同知识表示的多个汉语同义词。

表 10 “静寂”平行条件用例示意

平行条件				满足条件的《现汉》词
词性	构词结构	前语素义	后语素义	
形容词	联合	安静，无声	孤单，冷清	冷寂、冷清、寂寞、寂寥、悄寂、清冷、清寂

4 结语

未登录词的词法和语义预测是计算语言学领域中的基础性问题。我们依据“平行周遍”原则，从现有词的语义构词知识中提取“平行条件”，将未登录词潜在的构词因素与这些“平行条件”进行适应性匹配，对未登录词的词性、构词结构、词义等知识表示进行相对完整的预测。该方法将新的语言学理论与未登录词的理解应用问题结合，取得了显著的效果，其解释能力、便捷性和精细程度优于此前方法；换个角度看，这些面向计算的工作，也首次在大规模汉语语义构词成果的基础上，对“平行周遍”语言学理论和观点，进行了系统化、实证性的梳理和校验。这些研究，除了在自然语言处理领域有实际应用价值，也有望推动词典编撰、语言研究与教学等人文领域的进展。

目前，对于涉及转义的未登录词，其资格判定以及语义预测容易出现偏差，在后续，我们希望和此前相关工作[27]结合，从而提高准确率；在“平行条件”获取的问题上，未登录词的前、后语素或许也可以不限于铆定单一语素概念，在语素概念系统中适当引入相似度计算的手段，以增加“平行条件”的弹性，扩大对未登录词的覆盖；结合现有语义构词成果，除了对语料中出现的未登录词做知识表示与预测外，我们也可以在语素概念上直接合成“新词”，这也是新催生的研究课题；此外，我们采取分层迭代分析的技术路线，对《现汉》中的三字及以上词，语义构词知识标注的工作也在开展中，并将应用于未登录词的知识表示与预测。

参考文献

[1]周洪波. 新词语的预测[J]. 语言文字应用, 1996, 2: 73-78.

[2]王东海,王丽英. 谈新词语预测的依据[J].长江学术, 2010(02):118-124+117.

[3]TSENG H. Semantic classification of Chinese unknown words.[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2003.

[4]LU X. Hybrid models for semantic classification of Chinese unknown words[C]//Human Language Technologies 2007: The Conference of the North American Chapter of the Association for

Computational Linguistics; Proceedings of the Main Conference. 2007: 188-195.

[5]尚芬芬,顾彦慧,戴茹冰,等. 基于《现代汉语语义词典》的未登录词语义预测研究[J]. 北京大学学报(自然科学版), 2016, 52(1):10-16.

[6]张瑞霞,杨国增,闫新庆. 基于知网的汉语普通未登录词语义分析模型[J]. 计算机应用与软件, 2012, 29(8): 126-130.

[7]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, 30(6): 26-34.

[8]瞿健菊,冯敏萱. 基于知识库的汉语未登录词语义预测[J]. 中文信息学报, 2018, 32(1): 34-42.

[9]陈保亚. 论平行周遍原则与规则语素组的判定[J]. 中国语文, 2006, 2: 99-108.

[10]徐通锵. 徐通锵自选集/著名中年语言学家自选集[M]. 河南教育出版社, 1993.

[11]周上之. 辞研究和复合词研究的共同任务, 载《世纪对话——汉语字本位与词本位的多角度研究》[M], 北京大学出版社, 2013.

[12]金朝炜. 字的性质和类推, 载《世纪对话——汉语字本位与词本位的多角度研究》[M], 北京大学出版社, 2013.

[13]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[J], 2018, 32(2): 12-21

[14] LIN Z, LIU Y. Implanting Rational Knowledge into Distributed Representation at Morpheme Level[C]//AAAI. 2019.

[15]赵元任. 汉语词的概念及其结构和节奏[J]. 见: 中国现代语言学的开拓和发展——赵元任语言学论文选. 北京: 清华大学出版社, 1992.

[16]吕叔湘. 汉语语法分析问题[M]. 商务印书馆, 1979.

[17]吕叔湘. 中国文法要略[M]. 商务印书馆, 1990.

[18]高名凯. 汉语语法论[M]. 商务印书馆, 1986.

[19]朱德熙. 语法体系和语法分析[J]. 中国语文, 1982, 1.

[20]刘叔新. 汉语复合词内部形式的特点与类别[J]. 中国语文, 1985, 3: 186-188.

[21]傅力. 复合式合成词中应该有〝同位型〞的地位[J]. 汉语学习, 1989 (4): 11-13.

[22]周荐. 复合词词素间的意义结构关系[J]. 语言研究论丛第六辑, 天津教育出版社, 1991.

[23]杨梅. 现代汉语合成词构词研究[D]. 南京: 南京师范大学, 2006.

[24]符淮青. 义项的性质与分合[J]. 辞书研究, 1981, 3: 86-94.

[25]符淮青. 词义和构成词的语素义的关系[J]. 辞书研究, 1981, 1: 98-110.

[26]亢世勇, 李毅, 孙道功, 等. 汉语系统语料库的建设与词典编纂[J]. 上海辞书学会, 2004.

[27]陈龙, 饶琪, 刘扬. 汉语词的非字面义的表示与应用[C]. 第十七届中国计算语言学大会(CCL2018), 《中国科学 信息科学》待刊.

[28]陈保亚. 再论平行周遍原则和不规则字组的判定[J]. 汉语学习, 2005 (1).

[29]陈保亚. 论平行周遍原则与规则语素组的判定[J]. 中国语文, 2006, 2: 99-108.

[30]郑家恒, 李文花. 基于构词法的网络新词自动识别初探[J]. 山西大学学报（自然科学版）, 2002, 25(2):115-119.

[31]张海军, 冯冲, 史树敏, 等. 一种应用组合特征的中文未登录词词性猜测研究[J]. 小型微型计算机系统, 2010, 31(7): 1402-1406.

[32]洪铭材, 张阔, 唐杰, 等. 基于条件随机场 (CRFs) 的中文词性标注方法[D]. 2006.