文章编号：1003-0077（2018）00-0000-00

汉语词的非字面义的表示与应用*

陈龙[1,2]，饶琪[1,3]，刘扬[1,3]

（1.北京大学 计算语言学教育部重点实验室，北京 100871；

2.北京大学中国语言文学系，北京 100871；

3.北京大学 计算语言学研究所，北京 100871）

摘要：作为一种意合型语言，汉语由字组词的特性明显，字面义词的词义大体可由其构词结构和语素概念来表达，但对非字面义词的处理存在偏差，这也是语言深度理解中的一个棘手问题。本文从语言认知的角度出发，提出了适用于汉语词的非字面义的知识表示方式：全面发掘《现代汉语词典》中的非字面义二字词，判定它们作为隐喻或转喻现象的非字面义类型，标注其在《同义词词林》中的源域、目标域，并选取面向计算的适合的字面义词承担者。该工作首次在词汇级别上，系统地揭示了汉语隐喻和转喻现象的数量、类型及语义域映射分布状况，并且在算法框架不变的情况下，显著改进了词义相似度计算效果。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深入开展。

关键词：语义构词 字面义 非字面义 隐喻 转喻

中图分类号：TP391 **文献标识码：**A

Knowledge Representation of Non-literal Meanings of Chinese Words and Its Applications

CHEN Long[1,2]，RAO Qi[1,3]，LIU Yang[1,3]

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;

3. Institute of Computational Linguistics, Peking University, Beijing 100871, China.

Abstract: As a parataxis language, Chinese is characterized by its word-formation. The meanings of Chinese literal words can somehow be expressed by their word-formation patterns and morphemic concepts, but the meanings of non-literal words are quite deviated, which causes a problem in the deep understanding of language. From the perspective of language cognition, this paper puts forward knowledge representation of non-literal meanings of Chinese words. We comprehensively excavate the non-literal meanings of disyllabic words in the Modern Chinese Dictionary, by labeling the meanings as specific metaphorical or metonymic types and annotating their source and target domains in Tongyici Cilin and their synonyms with literal meanings. This work, for the first time at the lexical level, systematically reveals the distribution of types and semantic domain mapping of Chinese metaphor and metonymy, and greatly improves the performance of word similarity calculation under the same algorithmic frame. These ideas, practices and language resources construction are expected to promote the in-depth development of humanities and computing applications as well.

Keywords: Semantic Word-Formation; Literal Meanings; Non-literal Meanings; Metaphor; Metonymy

1 引言

在自然语言理解中，特定语言单位的意义通常由其结构和组成成分的意义导出，这也是语言计算遵循的主流途径[1]。作为一种意合型语言，汉语的构词分析表现出鲜明的特点。在词汇方面，汉语的合成词占了绝大多数，赵元任等学者[2]还指出，汉语复合词的构词规则与句法规则类似。因此，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，构成了汉语语义分析的起点[3]。汉语词的词义若为其字面义，大体可以由它的结构类型及构词语素来做表示、获取[3]。但在语言处理中，也会遇到“上台”、“铁窗”这样的非字面义词，它们的词义和字面义相差较大。在当前，考虑深度理解的任务要求[4]，有必要将词的字

*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）

面义和非字面义关联起来进行研究，并考察、评估其现实的应用效果。

在词的级别上，不同语言中的字面义和非字面义现象，表现出不同的特点，采取了不同的路线。

在印欧语系中，由语素到词的构词法比由词到词组、由词组到句子的句法要来得复杂。比如，英语复合词占的比例不大，除复合词外，其它类型的词中的语素形式多变且构词规则更为复杂[5]。因此，其句子可以通过词和短语推导出字面义来，但词义很难从更小的语言单位出发推导得到，基于字面义的构词分析和语素表示并不是印欧语系语言资源研发工作的关注点。

从认知的角度看，语言中非字面义的产生有隐喻、转喻等不同方式[6,7]。Lakoff 和 Johnson 认为，隐喻和转喻不只是一种修辞，而是人们通过一种事物去认识、去代表另一种事物的通用方式，在思维和语言体系中都有显著体现[8]。相应地，涉及非字面义表示的英语资源多采取了这种角度的描述，典型工作包括 Berkeley 的 MasterMetaphorList[9]、Fass 的 SenseFrame[10]、Martin 的 MetaBank[11]、Goatly 的 MetaLude[12]、VU Amsterdam 的 Metaphor Corpus[13]等。考虑印欧语系语言的特点，这些知识库多是基于句子建立的，只有少数是基于词的，也难以从语素构词的角度出发来表达非字面义及建构资源。

和印欧语系不同，汉语是一种意合型语言，它以汉字这种普遍的音义符号作为语言中的基本构建单位。一个汉字依据意义的不同分属不同的语素[14]，比如“桃花”中的语素“花 $_1$”和“花钱”中的语素“花 $_2$”代表了不同的字义。换言之，汉语中的构词结构和语素形式是较为固定的，由语素构词的特征更为明显。

目前，着眼于汉语词的字面义的表示，在构词分析和语素表示方面的研发成果，主要有清华大学“汉语语素数据库”[15]、鲁东大学“汉字义类信息库”及“汉语语义构词信息库”[16]、北京大学“汉语概念词典”[3]等，它们都尝试分析、计算汉语复合词的意义。其中，“汉语概念词典”以《现代汉语词典（第五版）》（以下简称《现汉》）刻画的汉语语素为依据，在此基础上描述汉语词的构词结构，实现构词成分与“语素概念”的严格绑定，以此来表达汉语词的字面义。

而在汉语词的非字面义方面，目前还缺乏较为深入的研发工作。厦门大学构建了汉语隐喻标注句库[17]，包含了有隐喻现象的句子及其句法分析，但是并没有指明句子中的隐喻部分，也没有描写源域、目标域等信息。北京大学构建了名词性隐喻标注语料库[18]，包括源域词语、目标域词语、例句等信息，不过对其它词类没有考虑。Lu Xiaofei[19]按照 VU Amsterdam 的 Metaphor Corpus 方式，建设了汉语语料库，但是对隐喻知识的描写依然匮乏。在以上资源中，只有北大标注语料库包括了源域、目标域信息，此外，现有的工作没有考虑汉语的特点，和英语一样，多在词或句子的级别展开，尚未以语素构词的视角来记录、表达隐喻和转喻信息。

基于如上考虑，在现有“汉语概念词典”依据语义构词知识表达词的字面义的工作基础上[3]，本文从认知角度出发，探索非字面义词的意义表示。我们希望系统地发掘汉语词的字面义和非字面义之间的区分与关联，描写词的非字面义的产生类型，即隐喻、转喻的确认及它们自身的一些性质，描写隐喻和转喻发生的源域、目标域，以及该非字面义的字面义词承担者，并考察、验证其在人文领域和计算理解等方面的应用。

2 汉语词的字面义与非字面义探索

朱德熙[14]、赵元任[20]等学者指出，汉语复合词的词义不一定都能从构词语素的意义推演得到，部分词可能是不透明的。李晋霞[21]按语义透明度把汉语词分成如下几类：有的词语义完全透明，如“哀叹”的词义基本可以通过语素义相加直接得到，这类词采用是它的字面义；有的词语义比较隐晦或比较透明，如“裁缝”的词义可以在结构信息和语素义的基础上引申得到；有的词语义完全隐晦，如“牺牲”的词义无法从字面义“祭祀用的牛羊”得到。这些不完全透明的词义往往就是非字面义。

依据符淮青对词义和语素义关系的总结[22]，汉语词的释义包括语素义（记为 c）、词的暗含内容（记为 a）、为表述需要而补充的内容（记为 b）和知识性附加内容（记为 s），如果将词义记为 Z，那么有如下类型：Z=c1+c2、Z=c1=c2、Z=c1+c2+a、Z=c1+c2+b、Z=c1+c2+a+b、Z=c1+c2+(a)+(b)+s、Z= (c1+c2)的比喻或引申义。在这些类型中，前六种一般认为属于字面义范畴，而后一种是比喻或引申义，即需要关注的非字面义。

从认知的角度看，非字面义产生的手段主要是隐喻和转喻[6,7]。Ullmann[6]、Waldron[7]等指出，隐喻和转喻在语言变化尤其是词义引申中，起着重要作用。对于两者的关联，有些学者认为隐喻是转喻的一部分，有些则认为转喻是隐喻的一部分[10]。目前多数研究认为，隐喻和转喻是有显著区别的[9,12]。Lakoff 和 Johnson 认为[8]，隐喻是通过一种事物来理解另一种事物，而转喻是通过一种事物来指代另一种事物。Ullmann[6]还指出，隐喻与两个语义范畴之间的相似性相关，而转喻与相关性相关。

考虑主流的语言资源多是从认知角度出发建设，我们也采用隐喻和转喻角度对汉语词的非字面义进行分类，并将其视为非字面义产生的不同方式。通过判断字面义与非字面义之间是相似还是相关关系，并参照字面义和非字面义的语义域，来断定其非字面义产生的具体方式是隐喻还是转喻。

例如，“上台”一词的非字面义为“出任官职或掌权”，字面义为“登上舞台”，这两者之间有相似性，都有“显露出来、向众人展示”的意味。其字面义的语义域是“现实动作”类别，非字面义的语义域是“政治活动”类别，字面义和非字面义处于不同的语义域中，因此转义方式判断为隐喻；而“铁窗”一词的非字面义为“监狱”，字面义为“铁质的窗户”，这两之间没有相似性、但有相关性，“铁质的窗户”往往是“监狱”的组成部分。其字面义和非字面义都处于“建筑物”的语义域中，因此转义方式判断为转喻。

3 汉语词的字面义表示方法

此前，“汉语概念词典”已经给出了汉语词的字面义表示方法[3]。考虑字面义和非字面义存在一系列的关联特征，这里只作简单介绍和铺垫，以形成辅助于非字面义表示的基础认识。

3.1 汉语的语素概念提取

考虑词典的权威性和应用的影响力，汉语语素取自《现汉》中的定义。我们对全部 20855 个语素义做释义文本的提取，并赋予唯一的“语素义编码”。例如，“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：这是该字在《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。

我们对《现汉》中相同语素类的释义文本进行语义相似度计算，经人工检验、迭代，形成一个个“同义语素集”，由此获得了 4199 个“语素概念”。之后，进一步在这些语义基元之间建立起层次结构，形成了“语素概念体系”，以方便后续的认知、推理和计算。

3.2 汉语的语义构词分析

对于汉语的构词结构性质，语言学界一般有语法构词[2,16]、语义构词[23,24]等不同观点。在综合、借鉴前人经验的基础上，我们确定了包括16种构词结构类型的标签集，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。以义项区分为基础，我们为《现汉》中的全部52108个二字词依据规范标注了构词结构信息。

在构词结构基础上，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。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”系统的制约。

3.3 汉语词的字面义表示

基于上述工作，我们获得了涵盖词性、构词结构、前后语素类、前后语素义等的广义构词知识。例如，在前后语素义方面，“选材”一词中，“选”的语素义为“挑选、选拔”，语素义编码为“选 1_04_01”，“材”语素义为“有才能的人”，语素义编码为“材 1_05_04”；“铁窗”一词中，“铁”的语素义为“一种金属元素”，语素义编码为“铁 1_07_01”，“窗”的语素义为“窗户”，语素义编码为“窗 1_01_01”。它们的广义构词知识如表 1 所示，这些信息在计算上有较大价值，能够简洁、有效地形成汉语词的字面义表示[27,28]。

表 1 汉语词的广义构词知识示例

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
铁窗	名词	定中结构	名语素	名语素	铁1_07_01	窗1_01_01

4 汉语词的非字面义表示方法

对于《现汉》中收录的词，我们依据符淮青的方法[22]，判断词义属于哪一类别，确定其是字面义还是非字面义，并做人工校验和筛查。对于非字面义，其释义文本也往往携带“比喻”、“借指”等标识，借此可进一步标注其隐喻、转喻类别，涉及的源域、目标域以及非字面义的字面义词承担者，以此来表达并计算词的非字面义。

4.1 非字面义的隐喻、转喻类型表示

4.1.1 非字面义的隐喻类型表示

对隐喻，可以从不同角度进行分类。从发生隐喻的单位看，可以分为词汇级隐喻、语句级隐喻、篇章级隐喻等[4]；从语言使用中的情况看，可以分为死喻、活喻等[27]。我们描述的从字面义到非字面义的隐喻现象，基本上属于词汇级的死喻，即已经约定、固定下来的那些隐喻。

在隐喻定性方面，Lakoff & Johnson[8]指出，隐喻是有方向性的，人们一般会通过一个具体的概念去理解一个相对抽象的概念。此外，Sweetser[28]指出人们会用外部世界的事物去理解人内在的心理、情感。针对汉语词的隐喻现象及特点，我们认为，通过隐喻方向来给隐喻分类是较好的选择。Stern[29]将隐喻分为基于相似性的隐喻和基于其它关系的隐喻，前者又细分为：外观相似，质量、功能相似，以及感官、情感相似等三类。针对隐喻产生的词义引申，Xu 也提出了可能的隐喻方向[30]：具体→抽象、涉身→非涉身、外在→内在、有生→无生、情感效度低→情感效度高、交互主观性强→交互主观性弱。借鉴以上观点和成果并做适应性调整，在语言工程上，我们将汉语词的隐喻方式分为如下 5 种类别：

① 具体→抽象：如果词的字面义表示的是具体的事物，而词义表示的事物相对更抽象一些，那么隐喻方式标为“具体→抽象”。例如，“主流”的一个义项为“比喻事情发展的主要方面”，这是非字面义，而它的字面义是“河流的主要部分”，表示具体的事物，非字面义相对来说更抽象；

② 涉身→非涉身：如果词的字面义表示的是和人的身体相关的动作或事件，而词义表示的动作与人自身不相关，那么隐喻方式标为“涉身→非涉身”。例如，“上马”的词义为“比喻开始某项较大的工作或工程”，是一个非涉身的事件，这是它的非字面义，它的字面义为“骑上马”，是一个涉身的动作；

③ 外在→内在：如果词的字面义表示的是外部世界的事物或者它们的行为，非字面义表示的是与人的身体、心理相关的事物或者行为，那么隐喻方式标为“外在→内在”。例如，“沉浸”的字面义为“浸入水中”，非字面义为“比喻处于某种境界或思想活动中”，它的字面义是外部世界的实体的状态，非字面义是与人的思想、心理相关的内在的状态；

④ 非人生物→具象物：如果词的字面义表示的是除了人以外的生物或其性质、行为，而非字面义表示的是另一种生物（包括人）或非生物，或其性质、行为，那么隐喻方式标为“非人生物→具象物”。例如，“复苏”的字面义是“生命体在机能减缓后恢复正常的活动”，非字面义是“经济再生产周期中继萧条之后的一个阶段”，它的字面义原本是生物体活动，而非字面义则代表了经济活动；

⑤ 客观→主观：如果词的字面义更客观、中性，而非字面义偏主观、褒贬，那么隐喻方式标为“客观→主观”。例如，“伸手”的字面义是“伸出手”，非字面义是“插手”，字面义是中性的，非字面义是贬义的。

需要注意的是，以上类别选项并不完全互斥，有时字面义和非字面义之间可以有不止一种隐喻方式，这时应该标注所有的隐喻方式。例如，“下游”的字面义为“河流靠近入海口的部分”，非字面义为“比喻落后的地位”，非字面义相对于字面义不仅更抽象，而且主观性更强。“①具体→抽象”和“⑤客观→主观”都是适合的隐喻方式，这时，两种方式都要标注。

4.1.2 非字面义的转喻类型表示

和隐喻表示一样，对于转喻，我们也标注其转喻方向。

Lakoff 和 Johnson[8]对转喻的类型进行了一些初步的分类，提出了部分→整体、生产者→产品等 7 种常

见转喻方向。Meta5 系统中区分了 5 种转喻类型：容器→内容、艺术家→艺术品、部分→整体、属性→实体和同事→活动[10]。这些分类在句子层面上可以覆盖一些转喻现象，但有些不会出现在词层面上，甚至极少出现在汉语中。借鉴以上观点和成果，并考虑分类完备性和一致性的要求，我们强调转喻词的字面义和非字面义在语法、语义上的对立性质，将汉语词的转喻方式分为如下 3 种类别：

① 词性改变因素：从语法的角度观察，汉语有大批的词的字面义和非字面义分属于不同的词类，这些转喻在语义上包括“事件→主体”、“事件→客体”、“物体→属性”等，我们将其归为“词性改变因素”。例如，动词的“编辑”是字面义，与字面义的词性不同，名词的“编辑”是非字面义。表 2 列举了一些典型的“词性改变因素”的例子，表中“ABi”代表“AB”一词在词典中的第 i 个义项，如果没有用数字区分，则为第一个义项。以下表格同理；

表 2 “词性改变因素”情况示例

词	释义	词性改变	转喻方向
领导 $_2$	担任领导工作的人	v→n	事件→主体
收入 $_2$	收进来的钱	v→n	事件→客体
摆渡 $_3$	摆渡的船	v→n	事件→工具
撰述 $_2$	撰述的作品	v→n	事件→结果
整齐 $_2$	使整齐	a→v	属性→致使
无辜 $_3$	没有罪的人	a→n	属性→主体
新潮 $_2$	符合新潮的；时髦	n→a	事物→属性
补益 $_2$	产生益处；使获得益处	n→v	客体→事件

② 部分与整体：“部分和整体”也是比较常见和典型的转喻类别，如果词的字面义表示的是非字面义的一部分，包括事物的成分、动作的分解等，那么转喻方式标为“部分与整体”。例如，“铁窗”的非字面义为“监狱”，字面义为“铁质的窗户”，字面义是非字面义的一部分；

③ 其它类别：如果词的字面义和词义的关系不存在以上情况，往往体现为“典型特点”，则将其归为“其它类别”。例如，“人烟”的字面义为“人和炊烟”，非字面义是“人家、住户”，“人和炊烟”是“人家、住户”的典型特点。从语义的角度看，该类别下的源域和目标域的映射状况较多，难以穷尽，而且每一种映射下的词的数量不多。表 3 列举了一些典型的“其它类别”的例子。

表 3 “其它类别”情况示例

词	词义	转喻方向
花卉 $_2$	以花草为题材的中国画	内容→艺术品
文书 $_2$	从事公文、书信工作的人员	工作内容→工作者
官府 $_2$	封建官吏	工作处所→工作者
割席	跟朋友绝交	故事中的行为→行为反映的典故事件

4.2 非字面义的源域、目标域表示

Shutova 指出[31]，设立源域和目标域是很难的，既要保证它们的选取能够覆盖所有的隐喻现象，又不能设得太泛，要兼顾覆盖面和具体性。考虑这些要求并借鉴现有资源，避免人为因素带来的主观性，我们采用哈工大《同义词词林扩展版》（以下简称《词林》）来绑定源域、目标域信息。《词林》分为 12 个大类、97 个中类、1400 个小类，每个小类中的词，依据词义的远近和相关性分为若干词群，每个词群中的词，又进一步分为若干行。在这样的层次结构下，源域、目标域的映射可以通过不同语义类来表达，满足不同的规约要求。

我们标注词的字面义和非字面义在《词林》中的位置，分别作为它们的源域和目标域。非字面义词的目标域，也就是其义项，实际上在《词林》中已经有体现，只标出其在《词林》中的位置即可。而该词的源域，也就是字面义，如果在《词林》中也存在，那么也只标出其位置即可，否则，依据构词知识在《词

林》中选择意义最接近的语义类，将它标为源域。例如，在《词林》中，“铁窗”和“监狱”作为同义词位于同一词群的同一行，该行就是“铁窗”词义的目标域。而“铁窗”的字面义为“铁质的窗户”，在“百叶窗、吊窗、钢窗……”这一小类，则把它标为源域。

4.3 非字面义的字面义词承担者表示

依据基于语义构词的词义计算的思想，对于非字面义词，需要寻找对应的字面义词承担者，作为统一计算框架下的替换来表达词义。“汉语概念词典”资源覆盖全部《现汉》词汇，对于每个非字面义词，需要寻找一个无转义的同义或近义词作为它的字面义词承担者，而同义或近义的差别暂不做区别性描述。

对于一个非字面义词，在《词林》中，往往会出现它的同义词，这时只需选取《现汉》中收录的一个无转义词作字面义词承担者即可。如果《词林》中并没有它的同义词，或者它的无转义的同义词没有被《现汉》收录，此时根据对词义的理解，挑选《词林》中意义最接近的无转义的近义词作字面义词承担者。例如，在《词林》中，非字面义的“铁窗”和《现汉》中的“监狱”是同义词，“监狱”即是它的字面义词承担者。通过该信息的标注，我们在语义计算中可以简单地将“铁窗”替换为“监狱”，并用“监狱”的字面义表示来表达“铁窗”的非字面义。表 4 列举了部分非字面义的字面义词承担者的情况，其中，关于字面义表示的约定详见此前论文[25]。

表 4 汉语词的非字面义的字面义词承担者示例

词例	字面义表示	词义描述	承担者	承担者的字面义表示
铁窗	<窗 1_01_01，铁 1_07_01 >	铁质窗户，借指监狱	监狱	<狱 1_02_01，监 2_02_02>
全豹	<豹 1_02_01，全 1_05_03>	比喻事物的全部	全貌	<貌 1_03_02，全 1_05_03 >

5 资源评估与应用

5.1 汉语词的非字面义数据分析

5.1.1 关于非字面义类型的数据分析

依据上述非字面义的判定标准及标注规范，我们发掘并标注了《现汉》中有非字面义现象的所有二字词，共计 3524 个，占《现汉》全部 52108 个二字词的 6.76%。其中，发生隐喻、转喻的分别为 1997、1527 个，分别占 3.83%、2.93%，以隐喻、转喻方式产生的词大体相当，前者稍多于后者。首次实证地给出了汉语词的非字面义的基本情况和占比分布，揭示了汉语以隐喻、转喻方式成词的状况。

在发生隐喻现象的汉语词中，1674 个词只涉及了一种隐喻类别，321 个词涉及了两种隐喻类别，另有 2 个词涉及了三种隐喻类别。各种隐喻类别的数量、占比和对应词例见表 7，在记录上，表中隐喻类别为①代表当且仅当涉及①这种单一类别，隐喻类型为①④代表当且仅当涉及①④这两种类别，以下标签含义同理约定。数据表明，在汉语的非字面义词中，“①具体→抽象”的隐喻类别占了多数，“②涉身→非涉身”以及“①具体→抽象+⑤具体→抽象”这两个类别的隐喻的占比也都超过了 10%，以上三种隐喻类别的占比超过了 80%。汉语词的隐喻类别的分布情况见表 5，这些信息有助于对汉语新词的隐喻方向做出预测。

表 5 汉语词的隐喻类别的分布

隐喻类别	数量	占比（%）	词例	词的字面义表示	词的非字面义表示
①	1160	58.1	主将 $_2$	<将 1_02_01，主 1_13_06 >	<力 1_05_02，主 1_13-06>
②	243	12.2	失足 $_2$	<失 1_07_02，足 1_04_01>	<堕 1_01_01，落 1_12_04>
③	73	3.7	沉浸	<沉 1_06_01，浸 1_03_01>	<处 2_06_03，于 1_02_01>
④	90	4.5	复活	<活 1_06_01，复 3_02_01>	<兴 2_07_01，复 3_02_01>
⑤	109	5.5	保守 $_2$	<保 1_07_02，守 1_05_01>	<守 1_05_03，旧 1_05_01>
①④	6	0.3	葛藤	<藤 1_02_01，葛 2_02_01>	<纠 1_02_01，纷 1_02_02>
①⑤	234	11.7	上供 $_2$	<上 2_14_03，供 1_02_02>	<行 4_13_07，贿 1_02_02>
②⑤	39	2.0	插手 $_2$	<插 1_02_02，手 1_07_01>	<介 1_04_01，入 1_05_01>
③⑤	11	0.6	气焰	<气 1_14_06，焰 1_01_01>	<势 1_06_01，焰 1_01_01>

④⑤	30	1.5	羽翼	<羽 1_04_02，翼 1_07_01>	<手 1_07_07，助 1_01_01>
①④⑤	2	0.1	鸡肋	<肋 1_01_01，鸡 1_02_01>	<事 1_06_01，琐 1_02_01>

而在发生转喻现象的汉语词中，“①词性改变因素”转喻类别占比为 49.1%，“②部分与整体”类别占比为 12.4%，“③其它类别”类别占比为 38.4%，明显没有交叉、重叠。其分布情况如表 6 所示。数据表明，汉语的转喻词中几乎一半涉及了“①词性改变因素”，可见该类别的转喻现象比较常见和典型。

表 6 汉语词的转喻类别的分布

转喻类别	数量	占比（%）	词例	词的字面义表示	词的非字面义表示
①	750	49.1	编辑 $_2$	<编 1_09_03，辑 1_02_01>	<者 1_06_01，编 1_09_03>
②	190	12.4	铁窗	<窗 1_01_01，铁 1_07_01>	<狱 1_02_01，监 2_02_02>
③	587	38.4	官府	<府 1_06_01，官 1_04_01>	<官 1_04_01，吏 1_03_02>

5.1.2 关于源域、目标域映射的数据分析

我们采用《词林》的语义类来表达源域和目标域，《词林》共有 97 个中类和 1400 个小类。隐喻词的源域覆盖了其中 88 个中类、622 个小类，目标域覆盖了 84 个中类、658 个小类；转喻词的源域覆盖了其中 85 中类、665 个小类，目标域覆盖了 89 个中类、591 个小类。数据表明，源域和目标域涉及了《词林》中的大部分中类，只涉及了不到一半的小类。用中类来考察源域和目标域，可以保证覆盖面，用小类来考察，可以满足具体性的要求，这两级语义类是比较合适的考察粒度。表 7、8 列出了在小类这一层面上，分布占优的源域以及分布占优的目标域，可以对转义映射的情况做出基本判断。从表中可以看出，隐喻的源域往往更为具象、与人的关系密切，而目标域相对抽象，但在转喻中，源域和目标域并没有类似的性质。

表 7 分布占优的隐喻源域、目标域示例

占优的隐喻源域	个数	词例	占优的隐喻目标域	个数	词例
Bk08（四肢）	26	巨擘、手心 $_2$、肘腋	Da21（情况、境地）	29	地狱 $_2$、平台 $_4$、底牌 $_2$
Dh01（妖魔鬼怪）	24	凶神、天仙、夜叉	Da20（形势、潮流）	26	春潮、暗流 $_2$、死棋
Hb02（行军作战）	16	挑战 $_2$、收兵 $_2$、攻关	Ig01 （开始、结束）	21	萌发 $_2$、起步 $_2$、上马

表 8 分布占优的转喻源域、目标域示例

占优的转喻源域	个数	词例	占优的转喻目标域	个数	词例
Bk02（头、脸）	13	头脸、头面、秃头 $_4$	Dj08（款项、费用）	27	积蓄 $_2$、茶钱 $_2$、开销 $_2$
Hc09（主持、统率）	13	指挥 $_2$、领队 $_2$、主席	Ae01（职工、职员）	26	剧务 $_2$、总务 $_2$、文书 $_2$
Da01（事情）	12	剧务 $_2$、常务、机密	Af10（领袖、领导）	20	总裁、主席、领导 $_2$

图 1 到图 4 以热力图的形式，展示了隐喻、转喻语义域映射的总体情况。图中的横坐标、纵坐标分别为非字面义词的源域、目标域在《词林》中的语义类编码，其中，图 1、图 2 从《词林》中类的角度做考察，图 3、图 4 从大类的角度做考察，格子的颜色越深表明发生映射的个数越多。

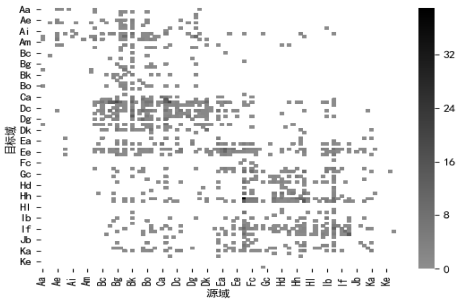

图 1 《词林》中类上的隐喻映射分布

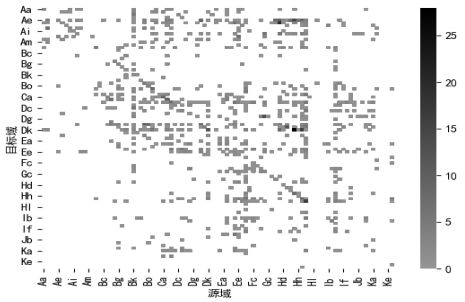

图 2 《词林》中类上的转喻映射分布

从《词林》中类单位看，最多的一个隐喻映射包含了 39 个词，源域为 Fa（手部动作），目标域为 Hi（人际交往），属于“②涉身→非涉身”类型；最多的一个转喻映射包含了 23 个词，源域为 Hg（教育），目标

域为 Dk（文化），属于“①词性改变因素”类型。有意思的是，图 1、图 2 中都存在一条对角线，即不少隐喻、转喻的源域和目标域属于同一中类。这说明，映射距离有时并不远，源域和目标域之间的语义相差不大。这可能是由于词的字面义和非字面义存在一些共性导致的。以对角线为参照，隐喻映射相对集中在两侧，而转喻映射相对更分散、更均匀，但它们关于对角线都是不对称的，这表明隐喻和转喻具有方向性。

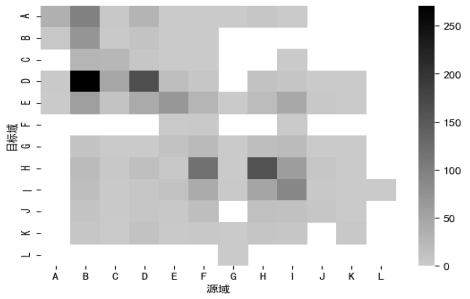

图 3 《词林》大类上的隐喻映射分布

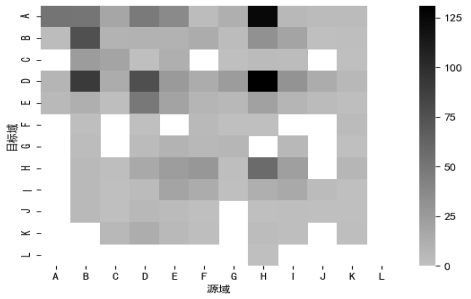

图 4 《词林》大类上的转喻映射分布

从《词林》大类单位看，在隐喻方面，B（物）、D（事）中充当源域较多，D（事）、H（抽象行为）中充当目标域较多；B→D、D→D、H→H、F→H 是较常见的隐喻，可见隐喻源域往往是具体的事物和行为，而目标域往往是抽象的。在转喻方面，B（物）、H（抽象行为）中充当源域较多，而 A（人）、D（事）中充当目标域较多；H→A、H→D 是较常见的转喻，表明汉语中存在较多用特定行为来指代人和事的情况。

5.1.3 关于字面义词承担者的数据分析

考虑基于语义构词的词义计算的需求，我们试图为有非字面义现象的词在《现汉》中寻找字面义词承担者。在全部 3524 个非字面义词中，3488 个词可以找到同义词或近义词作为字面义词承担者，占 99.0%。有 36 个词无法找到合适的近义词，只能用语义相关的词来作为字面义词承担者，占 1.0%。

5.2 对隐喻、转喻认知研究的实证

在隐喻、转喻等语言认知研究中，Lakoff 和 Turner[32]、Deignan[33]等学者认为，隐喻涉及两个语义域之间的映射，而转喻映射发生在同一个语义域中。我们的资源建设覆盖了《现汉》中的全部非字面义二字词，其源域、目标域信息也实证地支持了这一观点，并给出了详细的定量数据。

《词林》依据词义的远近来给词归类，同一语义域的词往往会被归入同一词群或者同一小类中的相邻几个词群中，而不同语义域的词倾向于出现在不同中类或大类中。不过，由于《词林》在大类的设置上考虑了词性因素，因而同一语义域中的不同词性的词会分别出现在不同大类中。转喻类别为“①词性改变因素”的 750 个词属于这种情况，这些词的字面义和非字面义实际上是属于同一个语义域的，如“事件→主体”、“属性→主体”等转喻方向。如果 Lakoff 和 Turner 及 Deignan 的观点成立，那么隐喻词的源域和目标域距离相对更远，转喻词的源域和目标域距离相对更近。表 9 按照距离由近到远的原则，列出了隐喻词和转喻词的源域和目标域在《词林》中的距离分布。表中最后一行的转喻词个数统计，涵盖了包含“①词性改变因素”的情形（涉及 750 个转喻词）及不包含的情形。

表 9 隐喻和转喻的源域、目标域的距离分布

源域和目标域的位置关系	隐喻词个数	比例（%）	转喻词个数	比例（%）
属于同一词群同一行	0	0	0	0
属于同一词群不同行	90	4.5	84	5.5
属于同一小类不同词群	14	0.7	13	0.9
属于同一中类不同小类	108	5.4	66	4.3
属于同一大类不同中类	420	21.0	170	11.1
属于不同大类	1365	68.4	444（1194）	29.1（78.3）

从表中可以看出，转喻的源域和目标域属于同一词群和同一小类的词的比例都显著高于隐喻，而属于

同一中类或同一大类的词的比例则低于隐喻。如果排除“①词性改变因素”类别的转喻词，那么源域和目标域属于不同大类的转喻的比例也显著低于隐喻。由此可以断言，大致上，隐喻的源域和目标域距离更远，一般分属不同语义域，而转喻倾向于发生在同一语义域中，但也不排除部分例外。这也系统地印证了语言中关于隐喻、转喻映射规律的认知假说。

5.3 在词义相似度计算上的应用

康司辰实验表明[26]，基于汉语的语义构词知识的词义相似度计算表现突出，并有合理的分布规律。但当遇到非字面义词的时候，相似度计算会遇到一些困难和干扰，因为在原方法中，汉语词的语义构词知识仅仅表达了字面义。在本资源构建后，对于非字面义词，本文方法采用它的字面义词承担者作为计算的替代，能在算法框架不变的情况下克服这一困难，提升计算结果的合理性。

表 10 原方法和本文方法的词义相似度计算结果对比

词 1	词 1 的非字面义表示	词 2	词 2 的字面义表示	原方法	本文方法
司机	<手 1_07_07，车 1_07_01>	开车	<开 2_18_06，车 1_07_01>	0.527	0.083
司机	<手 1_07_07，车 1_07_01>	乘客	<客 1_10_05，乘 4_04_01>	0.304	0.642
铁窗	<狱 1_02_01，监 2_02_02 >	纱窗	<窗 1_01_01，纱 1_04_03>	0.967	0.780
铁窗	<狱 1_02_01，监 2_02_02>	牢房	<房 1_07_02，牢 1_05_03>	0.374	0.910
脾胃	<志 1_03_01，趣 1_04_03>	肠胃	<肠 1_03_01，胃 1_02_01>	1.000	0.282
脾胃	<志 1_03_01，趣 1_04_03>	兴趣	<兴 1_01_01，趣 1_04_03>	0.542	0.917
全豹	<貌 1_03_02，全 1_05_01>	老虎	<虎 1_04_01，老 1_17_16 >	0.895	0.336
全豹	<貌 1_03_02，全 1_05_01>	全局	<局 1_10_05，全 1_05_03>	0.426	0.648

选取涉及非字面义词的若干词对进行实验，如表 10 所示，词 1 是非字面义，词 2 是字面义。在原方法中，非字面义词往往与和其字面义相近的词有更高的相似度，而与真正近义的词的相似度反而偏低。本文方法通过非字面义表示的引入，对该类共性问题有针对性地进行了校正，在涉及非字面义词的词对上的计算结果明显优于原方法。由此可见，增加了词的非字面义知识后，能够显著改进词义相似度计算效果，也使得基于语义构词的一种通用的相似度计算模型成为可能。

6 结束语

作为一种意合型语言，汉语由字组词的特性十分明显，字面义词的词义大体可由其构词结构和“语素概念”来表达。在语义构词分析方面，此前的语言资源建设和计算应用取得了积极的进展，但对非字面义词的处理则存在偏差与空白，这也是当前语言深度理解中期待解决的一个棘手问题。

我们从语言认知的角度出发，提出了一套适用于汉语词的非字面义的知识表示方式：全面发掘汉语中的非字面义二字词，判定它们作为隐喻或转喻现象的非字面义类型，标注其在《词林》中的源域、目标域，并选取面向计算的适合的字面义词承担者。依据该表示方式和规范，我们开展了扎实的语言知识工程，对《现汉》中涉及非字面义的 3524 个二字词进行了严格的认定分析和多种信息的标注。该工作首次在词汇级别上，系统地揭示了汉语隐喻和转喻现象的数量、类型及语义域映射分布状况，并且在算法框架不变的情况下，显著改进了词义相似度计算效果。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深入开展。

在未来，我们将对汉语的三字及以上词开展工作，推动基于语素构词的“汉语概念词典”资源的研究和发布，并且基于已经取得的成果，深入探索汉语新词的字面义与非字面义的识别、预测及应用。

参考文献

[1]袁毓林.自然语言理解的语言学假设[J].中国社会科学，1993,1:189-206.

[2]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.

[3]刘扬，林子，康司辰. 汉语的语素概念提取与语义构词分析[J]. 中文信息学报，2018，32(2)：12-21.

[4]俞士汶，王治敏、朱学锋. 文学语言与自然语言理解研究[C]//中文信息处理前沿进展——中国中文信息学会二十五周年学

术会议.北京:清华大学出版社，2006, 72-79.
[5]Bloomfield, L. Language[M]. New York City: Henry Holt,1933.
[6]Ullmann, S. The principles of semantics[M]. Glasgow: Jackson, Son & Co., 1957.
[7]Waldron, R.A. Sense and sense development[M]. London: Andre Deutsch, 1979.
[8]Lakoff, G., Johnson, M. Metaphor we live by[M]. Chicago: University of Chicago Press, 1980.
[9]Lakoff, G., Espenson, J., Schwartz, A. Master metaphor list. Draft 2nd Edition. Cognitive Linguistics Group, University of California at Berkeley, CA. 1991.
[10]Fass, D. met*: a method for discriminating metonymy and metaphor by computer[J]. Computational Linguistics,1991,17(1):49-90.
[11]Martin, J. Metabank: a knowledge base of metaphoric language conventions[J]. Computational Intelligence,1994,10(2):134-149.
[12]Goatly, Metalude metaphor at Lingnan University department of English, http://www.ln.edu.hk/lle/cwd03/lnproject_chi/introduction.html[EB/OL]
[13]Krennmayr, T., Steen, G. VU Amsterdam Metaphor Corpus[A]. Ide, N., Pustejovsky, J., Handbook of Linguistic Annotation[M]. Springer Netherlands, 2017: 1053-1071.
[14]朱德熙.语法讲义[M].北京：商务印书馆，1982.
[15]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.
[16]亢士勇，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论文集.上海辞书学会，2004.
[17]李剑锋，杨芸，周昌乐.面向隐喻计算的语料库研究和建设[J].心智与计算，2007,01：142-146.
[18]王治敏.汉语名词短语隐喻识别研究[D].北京：北京大学，2007.
[19]Lu Xiaofei, Wang Ben Pin-Yun. Towards a metaphor-annotated corpus of Mandarin Chinese[J]. Language Resources and Evaluation, 2017, 51(3):663-694.
[20]赵元任.汉语口语语法[M].吕叔湘译.北京：商务印书馆，1979.
[21]李晋霞、李宇明.论词义的透明度[J].语言研究，2008,28(3)：60-65.
[22]符淮青.词义和构成词的语素义的关系[J].辞书研究，1981，01：98-110.
[23]徐通锵.核心字和汉语的语义构辞法研究[J].语文研究，1997，03：2-16.
[24]刘叔新. 汉语描写词汇学[M]. 北京：商务印书馆，1990.
[25]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.
[26]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，2017，31(1)：94-101.
[27]Goatly, A., The language of metaphors[M]. Abingdon: Routledge,2011
[28]Sweetser E. From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure[M]. Cambridge: Cambridge University Press, 1991.
[29]Stern, G. Meaning and change of meaning: with reference to the English language[M]. Bloomington: Indiana University Press,1931.
[30] Xu Y, Malt B C, Srinivasan M. Evolution of polysemous word senses from metaphorical mappings[C]//Proceedings of the 38th annual meeting of the Cognitive Science Society,2016.
[31]Shutova, E. Annotation of linguistic and conceptual metaphor[A]. Ide, N., Pustejovsky, J., Handbook of Linguistic Annotation[M]. Springer Netherlands,2017: 1073-1100.
[32]Lakoff, G. More than cool reason: a field guide to poetic metaphor[M]. Chicago: University of Chicago Press,1989.
[33]Deignan, A. Metaphor and corpus linguistics[M]. Amsterdam: John Benjamins Publishing,2005.
陈龙（1995-），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。Email: 1400014101@pku.edu.cn
饶琪（1982-），博士后，副教授，主要研究领域为汉语语法、中文信息处理。E-mail: 337245309@qq.com
刘扬（1971-），通讯作者，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