·临床研究·

汉语名词语义特征建模与分析*

向伟华[1] 林 枫[1,2] 江钟立[1,3]

摘要

目的：构建汉语名词语义特征模型，为临床言语治疗提供量化和可视化语义特征数据库。

方法：以60个常见名词概念为刺激词，对健康受试者实施语义特征提名，采集各概念的语义特征条目，继而根据汉语语义特征分类方案对条目进行分类。最后以R软件实施分类数据可视化、聚类和统计检验。

结果：①根据语义特征进行的概念聚类与经验性分类结果基本一致，语义特征可以反映概念间语义关系。②不同概念领域的语义认知处理具有类型偏向性，生物概念具有更多的感觉信息特征，而非生物概念则具有更多的功能用途特征。③秩次居于首位的语义特征，在分类范畴型显著高。

结论：根据汉语语义特征数据建立的模型可以有效反映概念语义结构，有助于根据量化指标提取语义训练素材。

关键词 名词概念；语义特征；聚类分析；秩次；康复

中图分类号：R493 **文献标识码**：A **文章编号**：1001-1242(2015)-11-1118-07

The modeling and analysis of semantic features for the Chinese nouns/XIANG Weihua,LIN Feng,JIANG Zhongli//Chinese Journal of Rehabilitation Medicine, 2015, 30(11): 1118—1123

Abstract

Objective: To explore the model of semantic features for the Chinese nouns in order to provide quantitative and visual semantic database for clinical speech therapy.

Method: The semantic feature entries were collected from a total of 60 stimulating nouns in volunteers by feature nomination. These entries were classified into different feature types according to Semantic Feature Classification Scheme for Chinese. With the R statistical computing environment, the distribution of feature types was visualized and a clustering analysis based on properties was conducted.

Result: Firstly, the clusters based on properties were basically in line with the categories based on the experience. Secondly, cognitive processing of different domain had a certain bias. The living domain had significantly more perception properties, and nonliving domain had significantly more usage properties. Thirdly, the frequency of first rank features were significantly higher than expectation in the taxonomic categories.

Conclusion: The model of semantic features for the Chinese nouns by feature nomination can effectively reflect the semantic structures of concepts and help to chose semantic training materials according to quantitative indicators.

Author's address Dept. of Rehabilitation Medicine, The First Affiliated Hospital of Nanjing Medical University, Nanjing, 210029

Key word feature norms;semantic feature;clustering analysis ;order of production;rehabilitation

正常人在交流过程中可以迅速理解他人意思，并及时做出合理反应，这需要迅速地提取相关概念的语义知识[1]。了解语义知识如何在大脑中组织、储存对于理解正常人的语言理解和产生有重要意

DOI:10.3969/j.issn.1001-1242.2015.11.006

*基金项目：江苏省科技支撑计划(BE2012675)；国家自然科学基金资助项目(81171854)

1 南京医科大学第一附属医院康复医学科，210029；2 南京师范大学语言科技研究所，江苏省哲学社会科学重点研究基地；3 通讯作者

作者简介：向伟华，女，硕士研究生； 收稿日期：2015-01-04

义,更重要的是可以指导临床言语治疗。近年来在欧美得到快速发展的语义特征分析技术(semantic feature analysis,SFA),就是利用患者残存的语义知识系统改善患者命名功能的言语治疗技术[2]。这种技术认为概念可以解析为多个特征。例如,"筷子"的特征包括<餐具>、<用于进食>、<竹制的>、<夹取>、<细长的>、<成对出现>和<在桌上>等。但是,不同语义特征携带的信息量有所不同,<用于进食>比<竹制的>被提及的频率高,并且常是首位被提及的特征,对"筷子"具有重要的作用。这提示,在把语义特征分析技术应用于临床言语治疗前,需要建立正常人的语义特征库,对语义特征的性质进行定性和定量研究。

目前欧美国家多采用语义特征提名法,在正常人群中采集概念的语义特征,对特征进行分类,分析不同类型语义特征在概念中分布及其关系[3—4]。汉语中使用人群调查来研究概念语义特征,多用于心理学研究,如刘烨等[5]研究了三类范畴(动物、自然事物、人造物)的语义特征提取时间差异。此类研究通常单纯以频次统计作为分析指标,而没有对语义特征进行分型,也就没有在定性的基础上获取定量研究结果,从而难以直接移植到言语治疗实践中。

本研究旨在通过采集正常人群的名词概念语义特征,对其加以分类,从而建立汉语名词语义特征库。研究不同类型的语义特征在各类概念中的分布特点,为失语症及脑高级功能障碍提供评估和治疗素材。

1 对象与方法

1.1 受试者

受试者为66例自愿者,母语为汉语。其中男性33例,年龄为(20.30±2.01)岁,受教育年限(12.64±1.55)年,女性33例,年龄(20.45±0.89)岁,受教育年限(13.15±0.74)年。男性和女性年龄(t=-0.47,P=0.64)和受教育年限(t=-1.69,P=0.097)无显著性差异。所有受试者均为右利手。本实验通过南京医科大学伦理委员会批准,所有受试者自愿参加本次实验。

1.2 研究方法

1.2.1 刺激词:选用60个常用的汉语名词作为刺激词。刺激词分为生物和非生物两大领域。包括十个范畴:鸟类、身体部位、水果、蔬菜、哺乳动物、交通工具、建筑物、服装、家具和器物类。每个范畴含有6个概念。所选名词词频参考北京语言大学汉语国际教育技术研发中心汉语常用词频表,不同领域间词频差异t检验无显著性差异(P=0.588)。

1.2.2 语义特征采集方式:使用语义特征提名完成概念的语义特征的采集。60个概念词随机制成3个手册,每个手册包括20个概念。手册由两个部分组成,第一部分是填表说明,包括指导语和两个示例,第二个部分是20个刺激概念。受试者先对该概念熟悉度进行评价,之后写出概念的特征。不限制受试者填写的时间,但一般每个概念在1min内完成。每一册各由15例男性,15例女性完成,部分受试者完成了2—3册,但同一名受试者不会重复做同一册。从而确保每一册都有男女各15例完成。

1.2.3 分类方案:对语义特征的分类采用汉语语义特征分型方案。该方案主要参照Cree和McRae提出的脑区分类法[6],以及Lebani和Pianta提出的康复分类法[7]。该方案包括2个级别:初级型(6个)和一级型(23个)。①分类范畴:是指概念所处的范畴,包括种类、并列、特例、子类、同义、反义6个一级型,例如"苹果"<水果>属于种类,而"桥梁"<卢沟桥>则为特例;②组织构成:指概念的组成部分和其所属的总体,包括部分和总体2个一级型,例如"苹果"<有果皮>、"鼻子"<人体的一部分>属于组织构成型特征;③内省特征:指各种涉及主观情绪、评价的特征,包括情绪、客观特质、内省量和评判4个一级型,如"老鼠"<讨厌的>、<胆小的>分别内省特征中的情绪和客观特质;④感觉信息,包括视觉、触觉、听觉、味觉、嗅觉、实体数和共现7个一级型,例如"猫"<有四条腿>,"老虎"<动物园里>分别为实体数和共现;⑤功能用途:指有关用途和使用者信息的特征,一级型包括用途和用者,例如"围巾"<用来保暖>、<女人用>分别是用途和用者;⑥杂类包括百科知识和其他联系,例如"苹果"<在树上>属于百科知识,其他联系是难以归类的特征。

1.2.4 语义特征标注:实验结束后,由1名专业人员对所收集的原始反应条目进行标注,另外2名专业人员参与核对。例如"卡车的原始条目【可以运送货物】、【可运输货物】和【用来运输货物】统一转写成<

用来运输货物>。“鸽子”的原始条目【白色或灰色】转写为<白色的>和<灰色的>。“板凳”【有四条腿】包含了组成部分和数量的信息，转写为<有腿>和<有四条腿>。为减少人为误差，参考Lebani和Pianta的康复分类法[7]，在标注时加入特征提示符。例如，茄子【吃起来软软的】、沙发【软的】的中心词统一采用<软的>，但是它们的特征类型是不同的，前者是<品尝可知→软的>[味觉]，后者是<触摸可知→软的>[触觉]。

1.2.5 基本指标：衡量语义特征的指标很多，考虑到本特征库主要用于临床康复评估和治疗，仅选取了秩次。秩次是指受试者产生原始条目的顺序，它能反映受试者对特征通达难易程度及特征对概念的重要程度，秩次越靠前则越易通达，对概念越重要。

1.3 统计学分析

采用R软件3.1.2版实现分类数据可视化和聚类分析[8]。

2 结果

受试者对于生物和非生物领域的熟悉度分别为(7.35±1.03)和(7.54±1.25)，无显著差异(t=-0.67，P=0.50)。男性和女性对于概念的平均熟悉度分别为(7.25±1.11)和(7.66±1.21)，无显著差异(t=-1.95，P=0.053)。对受试者的原始条目进行标注，最终得到了13782条语义特征。

2.1 基于语义特征的范畴聚类

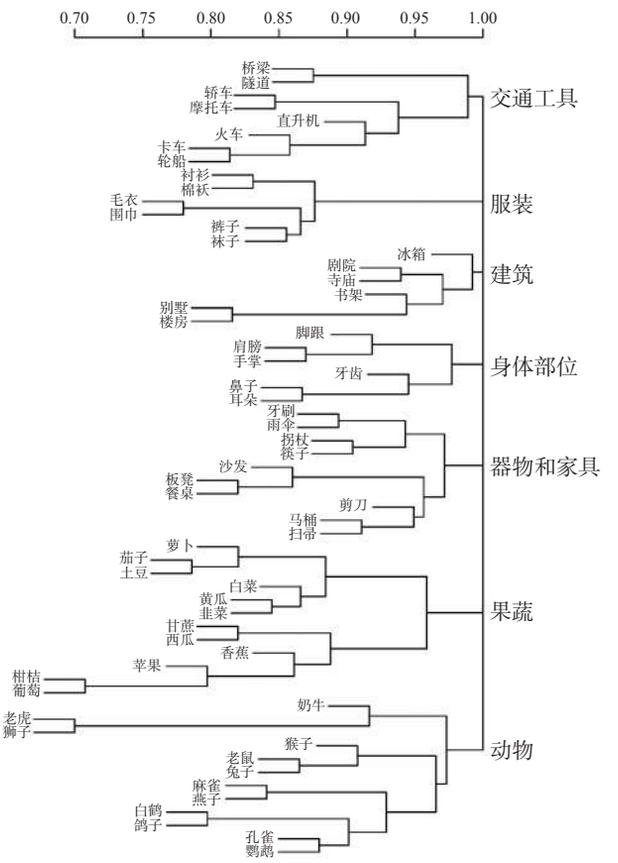
图1示60个概念的聚类树状图。该结果与经验性范畴划分非常接近，提示根据语义特征进行概念范畴分类符合常人经验性语义加工结果。

2.2 不同概念范畴的语义特征分布

各个初级型语义特征在所属范畴的频数分布情况使用马赛克图展示。马赛克图是近年来发展起来的分类数据可视化图形，已经在欧美语义特征的分型分析中得到使用[9—10]。

图2显示了不同概念范畴初级型语义特征分布。图中的各个维度对应各因素所代表水平，各个镶嵌矩形的面积越大，相应类型的语义特征被提名的频数越多。图右下角的P值表示对图中每个矩形进行卡方检验的结果。图2中P=2.22×10^{-16}(P<0.05)，表示语义特征类型与概念范畴相关。马赛克

图1 概念聚类树状图



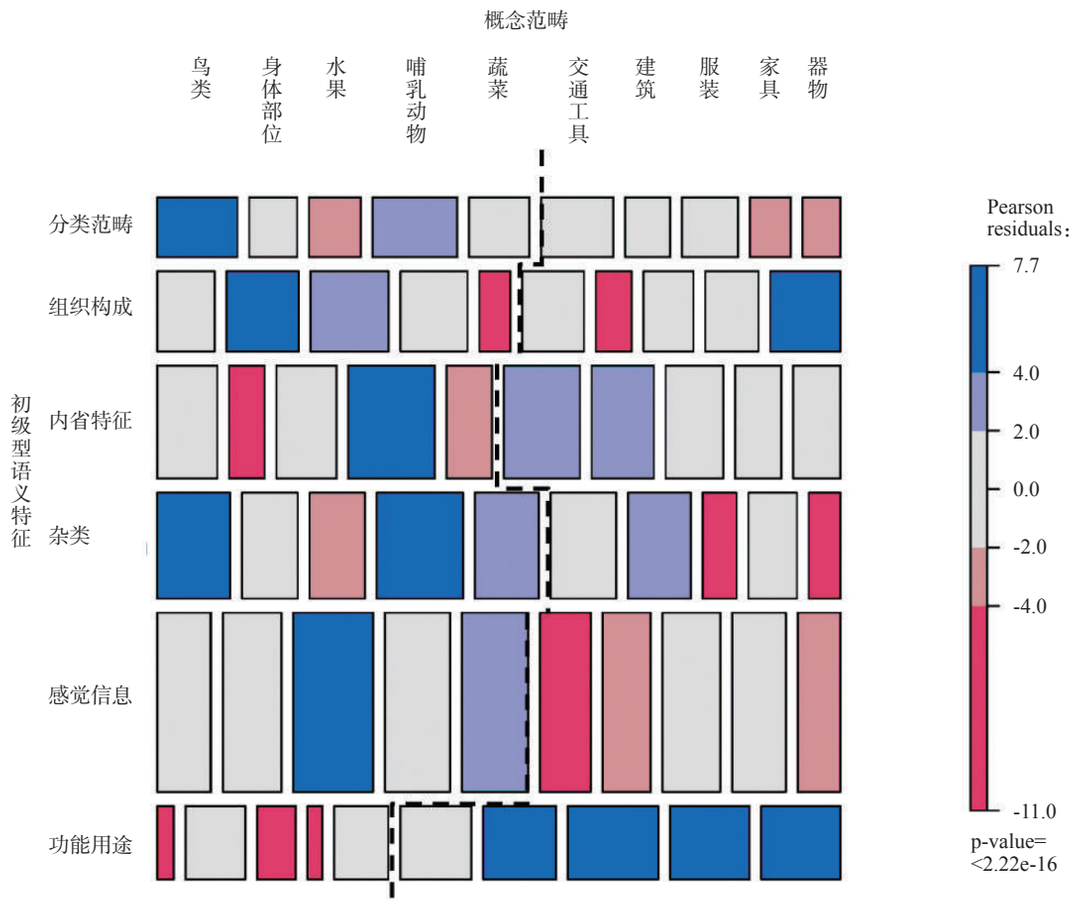
图上方显示余弦距离值，当值为1时，两个概念相似。当值为0时，概念语义无关。右侧标签为基于语义特征重叠的聚类范畴。

图中的颜色表示各个矩形进行皮尔逊残差检验的结果，即实际频数与预期频数之间比较的结果。蓝色矩形对应的是显著高于预期值，而红色对应显著低于预期值。颜色的明亮度代表皮尔逊残差的值的大小，深色表示大的皮尔逊残差(>4)，对应P<0.0001，而浅色代表小的残差(2<r<4)，对应P<0.05。灰色则表示分布没有显著性差异。

对概念各初级型特征的实际的频数比较可以发现，感觉信息所占实际频数最多，其次是内省特征和杂类，这表明对具体的事物的认识主要还是依靠感知觉信息。

对概念范畴的初级型特征分布比较可以发现，第一，鸟类和哺乳动物在分类范畴型、杂类型特征中所占频数显著高于预期，在功能用途型特征所占频数显著低于预期。第二，水果和蔬菜都在感觉

图2 初级型语义特征在所属范畴的马赛克图



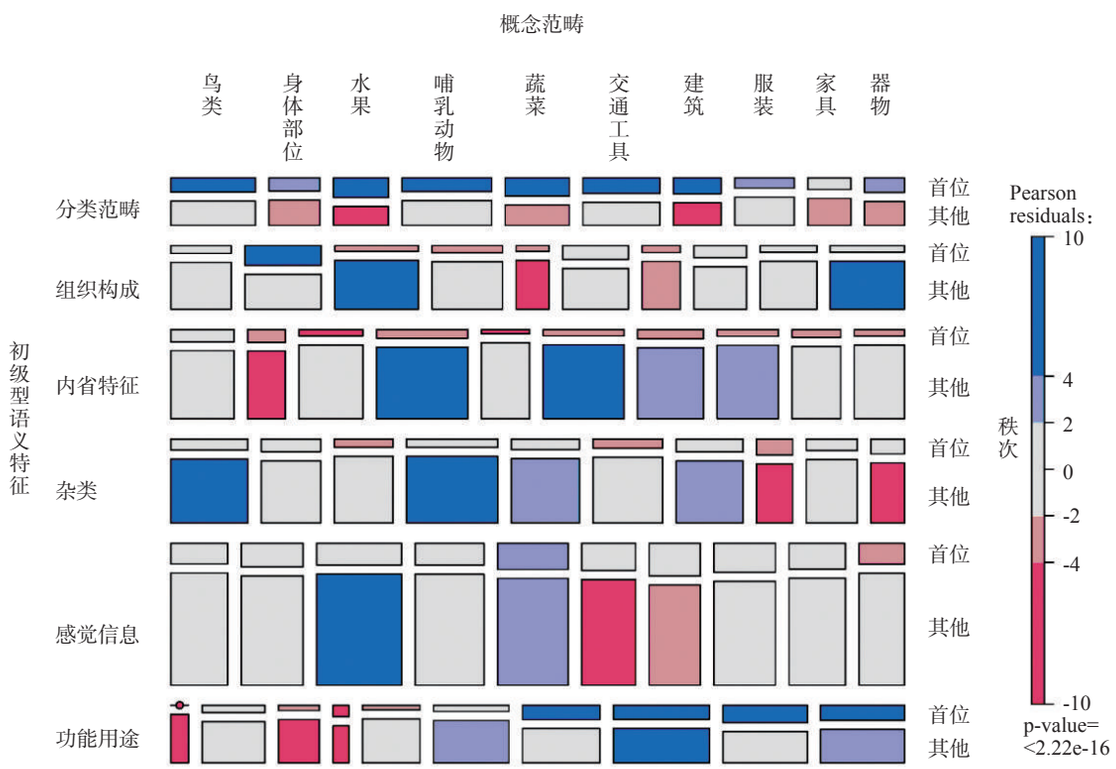
图左侧标签对应6个初级型,上方标签对应10个概念范畴。各个镶嵌矩形的面积越大,则该类型语义特征被提名的频数越多。图右下角的P值表示对图中每个矩形进行卡方检验的结果。图中$P=2.22\times10^{-16}$($P<0.05$),表示语义特征类型与概念范畴相关。马赛克图中的颜色表示各个矩形进行皮尔逊残差检验的结果,即实际频数与预期频数之间比较的结果。蓝色矩形对应显著高于预期值,而红色对应显著低于预期值。颜色的明亮度代表皮尔逊残差值的大小,深色表示大的皮尔逊残差(>4),对应$P<0.0001$,而浅色代表小的残差($2<r<4$),对应$P<0.05$。灰色则表示分布没有显著性差异。

信息显著高;而水果在分类范畴、杂类和功能用途型显著低,在组织构成型显著高;蔬菜在组织构成和内省特征中显著低,在杂类中显著高。第三,身体部位仅在组织构成显著高,在内省特征显著低;服装在功能用途显著高,在杂类显著低。第四,交通工具、建筑类在内省特征显著高,在感觉信息显著低;建筑类在功能用途显著高,组织构成显著低;交通工具在杂类中显著高。第五,家具和器物在功能用途显著高,在分类范畴显著低。综合以上结果可以看出,不同范畴初级型特征具有不同分布模式,但有一定的偏向性。大部分生物领域范畴的功能用途型特征显著低,而感觉信息范畴显著高,大部分非生物领域的范畴功能用途型特征显著高,而感觉信息显著低。

2.3 不同概念范畴首位秩次语义特征的分布

首位秩次的语义特征是受试者对概念最先产出的特征,它比其他位秩次的特征更易通达,在语义特征治疗中是要优先选取的素材。图3显示了首位秩次的语义特征分布情况。对不同范畴首位秩次语义特征比较发现,第一,除家具外,所有范畴的首位秩次的特征在分类范畴型特征的频数都显著高于预期。第二,大部分非生物领域范畴首位秩次的特征在功能用途型显著高。第三,身体部位的首位秩次的特征在结构组织型显著高,而蔬菜类的首位秩次的特征在感觉信息显著高。

图3　不同范畴首位秩次语义特征分布的马赛克图



图左侧标签对应6个初级型，上方对应10个范畴，右侧标签对应首位秩次和其他位秩次的特征。各个镶嵌矩形的面积越大，则该类型语义特征被提名的频数越多。图右下角的P值表示对图中每个矩形进行卡方检验的结果。图中$P=2.22\times10^{-16}$（$P<0.05$），表示语义特征类型、概念范畴和秩次三因素相关。马赛克图中的颜色表示各个矩形进行皮尔逊残差检验的结果，即实际频数与预期频数之间比较的结果。蓝色矩形对应显著高于预期值，而红色对应显著低于预期值。颜色的明亮度代表皮尔逊残差值的大小，深色表示大的皮尔逊残差（>4），对应$P<0.0001$，而浅色代表小的残差（2<r<4），对应P<0.05。灰色则表示分布没有统计差别。

3　讨论

3.1　语义特征的范畴聚类

本研究结果表明：第一，语义特征可以有效指导概念聚类。在图1中，从右往左看，图中最初分为交通工具、服装、建筑、身体部位、家具器物类、果蔬和动物7个范畴，接近于经验性划分的概念范畴。其中，家具和器物聚成一类，蔬菜和水果聚成一类，哺乳动物和鸟类聚合成一类，这符合经验对这些类别更高级别的归类。第二，生物领域概念比非生物更加紧密的聚集在一起，但是身体部位类概念聚集趋势和非生物类更为相似。Kremer等[9]对德语和意大利语名词的语义特征研究发现，身体部位类和工具类倾向于聚成一类。这提示语义特征可以很好的反应语义记忆的组织方式，为利用语义特征进行相关的评估和治疗提供前提和基础。

3.2　不同概念范畴的语义特征分布

对概念各初级型特征的实际的频数比较可以发现，感觉信息（鸽子<白色的>、甘蔗<甜的>、楼房<高>等）所占实际频数最多，这与信息加工心理学的观点是一致的。Gainotti等[11]人研究发现，视觉信息对动物、植物和人造物的识别都是重要的。

不同范畴的初级型语义特征分布模式具有自身特点。鸟类和哺乳动物类具有较多分类范畴和杂类特征，这和人们对于动物具有明确分类和较多的百科知识（如狮子<在动物园里>、<食肉><会奔跑>等）有关。蔬菜和水果的感觉信息型特征（如茄子<紫色的>、香蕉<黄色的>）显著高，感觉信息对于识别蔬菜和水果可能至关重要。身体部位仅在组织构成

型特征(鼻子<身体部位><人体的一部分><两个孔>)显著高,其特征分布更加接近非生物领域范畴。Masullo对视觉失认症患者语义知识的研究中发现,身体部位的语义知识受损程度和非生物范畴接近,显著低于其他生物范畴[12]。之前的研究也表明在聚类分析中,身体部位和工具类常聚合为一类[9]。这可能是由于身体部位作为人的外延,各个部位有特定的结构和功能,而且涉及对各类工具的操纵。交通工具和建筑在内省特征显著高,而在感觉信息显著低,因为其具有比较多的主观判断特征(如摩托车<速度快><很危险>,别墅<很贵><很大>)。Gainotti等[11]在研究不同类型的特征对动物、植物和人造物作用时发现,视觉信息对于识别三类范畴都是重要的。视觉以外的感觉特征对于生物是重要的,功能特征对于人造物是重要的。听觉信息和百科知识对于动物是重要的,行为动作(如剥皮、切,在本研究中属于感觉信息型中的共现)、味觉和嗅觉对于植物是重要的。

虽然不同范畴的语义特征在初级型分布模式具有自身特点,但是也具有一定偏向性。生命领域感觉信息型特征显著高,而非生物领域功能用途型特征显著高,这与Farah[13]及Cree[6]研究结果是一致的。根据感觉功能理论,感觉信息尤其是视觉型特征对于识别生物有重要作用,而功能型特征对于识别非生物范畴具有更重要的作用[14]。Hoffman等[15]让受试者直接评价感觉-运动模块对于概念的重要性,也获得类似的结果。

这些研究结果提示不同概念范畴的语义特征分布模型有可能发展成为脑损伤患者知觉信息加工受损的筛查工具。

3.3 不同范畴概念首位秩次语义特征的分布

Montefinese等[16]认为首位秩次的特征是概念最具有代表性的特征。例如,“围巾”的特征<用来保暖>、<长的>产生频次接近,但是受试者会更早的产出<用来保暖>。因此,分析首位秩次特征的分布可以进一步了解不同特征类型在概念中重要性,从而指导临床的言语治疗。

不同范畴概念首位秩次语义特征的分布表明,几乎所有范畴首位秩次在分类范畴型显著高,这暗示正常人在名词概念的特征提名时倾向于首先检索概念范畴有关的特征。范畴特征,尤其是分类对于概念是重要的。Clarke等[17]对物品识别的研究表明,正常人在图片呈现最初120ms即激活物品的共享性特征,可以对物品的所属的范畴进行区分。Martin[18]发现新习得的分类可以在视觉识别早期调节事物的感知。大部分非生物概念的首位秩次特征在功能用途型显著高,表明对于非生物功能用途型特征很容易产出,这从另一个方面证明功能用途型特征对于非生物概念的重要性。在失语症治疗中,语义特征分析技术并未提出让患者以何种顺序产出或者是识别特征,首位秩次语义特征分布模型为临床言语治疗提供了训练线索,Boyle在语义特征分析技术的综述中报道,尚不能确定识认、产出和混合这三种形式的语义特征分析方式对于失语症治疗疗效最好[19],因而对于大部分概念,可以让患者先尝试对概念所属范畴进行分类,而对于非生物概念,可以尝试说出其功能用途,不能完成则按以上顺序进行提示,对于“筷子”可以让患者先尝试说出或是提示患者<餐具>、<用来进食>,而对于“鼻子”可以让患者先说出或提示<身体部位>、<人体的一部分>、<两个孔>等。

3.4 研究局限性和展望

本研究中采用了语义特征提名测试来探究语义记忆的组织方式,这种方式本身存在一定的局限性。第一,受试者仅描述他们认为重要的特征,对于显而易见的特征可能避而不谈。例如,虽然大部分动物都有耳朵,但是耳朵这个特征常出现在“兔子”<长耳朵>这样的描述中,而较少的出现在“老虎”或者“狮子”这些概念中。第二,特征描述的难易程度不同。视觉信息对于生物和非生物的识别都具有重要作用,但是有些视觉信息难以用简短的汉语词来描述。例如,受试者很容易说出“剪刀”<用来剪东西>,但是说出“剪刀”的形状就比较困难。第三,本研究招募的自愿者为青年人,青年人产生的语义特征可能与中老年人存在一定的差异。在意大利语的研究中,Lenci等[10]将其研究中正常中老年受试者与Kremer和Baroni[9]研究中的青年受试者产生的语义特征的比较发现,两者对于11个相同概念产生的概念特征配对有66%是重叠的,进一步对概念特征配对产生频率的比较发现,有73%是重叠的,而产

生这些配对的受试者的人数的相关性高达0.84。青年受试者产生的语义特征大致可以反映出人群语义记忆的特点。当然，对于汉语名词概念，不同年龄受试者产出的语义特征是否存在差异尚需要进一步的研究。虽然存在这些问题，但特征提名提供事物语义组织的知识，有助于了解不同范畴概念的特点，以指导临床康复的评估和治疗。

综上所述，本研究初步尝试了建立一个小型正常人汉语名词语义特征分布的模型，分析不同种类概念的语义特征分型和定量指标的分布特点。该特征库的构建和分析方法，可供进一步应用于认知功能评定和治疗的临床实践，有利于为这些实践提供量化、可视化和可扩增的素材选取策略。

参考文献

[1] 林枫, 贺丹军, 江钟立. 适应性存储和快速提取的记忆结构模式分析[J].复杂系统与复杂性科学, 2009, 6(2): 40—49.

[2] 祁冬晴, 江钟立. 语义特征分析在失语症治疗中的应用进展[J].中国康复医学杂志, 2014, 29(3):282—285.

[3] Vinson DP, Vigliocco G. A semantic analysis of grammatical class impairments: semantic representations of object nouns, action nouns and action verbs[J]. Journal of Neurolinguistics, 2002, 15(3): 317—351.

[4] McRae K, Cree GS, Seidenberg MS, et al. Semantic feature production norms for a large set of living and nonliving things[J]. Behavior Research Methods, 2005, 37(4): 547—559.

[5] 刘烨, 傅小兰. 自然概念语义特征提取的范畴效应[J]. 心理科学, 2006, 29(2): 286—289.

[6] Cree GS, McRae K. Analyzing the factors underlying the structure and computation of the meaning of chipmunk,chey, chisel,cheese,and cello(and many other such concrete nouns) [J]. Journal of Experimental Psychology: General, 2003, 132 (2): 163.

[7] Lebani GE. STaRS. sys: designing and building a commonsense- knowledge enriched wordnet for therapeutic purposes [D]. University of Trento, 2012.

[8] R-Core-Team.. R: A language and environment for statistical computing. Vienna, Austria. : R Foundation for Statistical Computing[J].2012. Retrieved from http://www.R-project.org/.

[9] Kremer G, Baroni M. A set of semantic norms for German and Italian[J]. Behavior Research Methods, 2011, 43(1): 97—109.

[10] Lenci A, Baroni M, Cazzolli G, et al. BLIND: a set of semantic feature norms from the congenitally blind[J]. Behavior research methods, 2013, 45(4): 1218—1233.

[11] Gainotti G, Spinelli P, Scaricamazza E, et al. The evaluation of sources of knowledge underlying different conceptual categories[J]. Frontiers in Human Neuroscience, 2013, 7: 40.

[12] Masullo C, Piccininni C, Quaranta D, et al. Selective impairment of living things and musical instruments on a verbal 'Semantic Knowledge Questionnaire' in a case of apperceptive visual agnosia[J]. Brain and Cognition, 2012, 80(1): 155—159.

[13] Farah MJ, McClelland JL. A computational model of semantic memory impairment: modality specificity and emergent category specificity[J]. Journal of Experimental Psychology: General, 1991, 120(4): 339.

[14] 俞建梁, 陈先梅. 语义范畴特异性损伤研究三十年[J]. 外语研究, 2013 (4): 49—53.

[15] Hoffman P, Lambon Ralph MA. Shapes, scents and sounds: quantifying the full multi-sensory basis of conceptual knowledge[J]. Neuropsychologia, 2013, 51(1): 14—25.

[16] Montefinese M, Ambrosini E, Fairfield B, et al. Semantic significance: a new measure of feature salience[J]. Memory & Cognition, 2014, 42(3): 355—369.

[17] Clarke A, Taylor KI, Devereux B, et al. From perception to conception: how meaningful objects are processed over time[J]. Cerebral Cortex, 2013. 23(1): p. 187—197.

[18] Maier M, Glage P, Hohlfeld A, et al. Does the semantic content of verbal categories influence categorical perception? An ERP study[J]. Brain and Cognition, 2014, 91: 1—10.

[19] Boyle M. Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name?[J]. Topics in Stroke Rehabilitation, 2010, 17(6): 411—422.