基于《同义词词林》的中文语体分类资源构建

黄国敬[1]，周立炜[1]，饶高琦[1,*]，臧娇娇[2]

1. 北京语言大学汉语国际教育研究院，北京,100083
2. 腾讯科技有限公司，北京，100080

ellenh1001@163.com, liweiyeahmail@163.com, raogaoqi@blcu.edu.cn, jojozang@tencent.com

摘要

语体词是指在某一语体中专用的词语，是语体的语言要素和形式标记。而语体词的资源可以服务于与现实场景息息相关的NLP应用，但目前此类资源较为稀缺。对此，本文基于《大词林》，完成了“语体词标注”“语体（词）链条标注”和“平行构式标注”三个任务，建立了以语体词为基础的语体分类资源。本资源包含55,710条词语、5,017个语体链条和433组平行构式。基于此，本文分析了中文语体词的分布概况、形态差异以及词义词性的分布情况。

关键词： 语体词；语体分类资源；同义词

Construction of Chinese register classification resources based on “Tongyici Cilin”

Huang Guojing[1], Zhou Liwei[1], Rao Gaoqi[1,*], Zang Jiaojiao[2]

1. Beijing Language and Culture University,Research Institute of International Chinese Language Education, Beijing, 100083
2. Platform & Content Group, Tencent Technology Co., Ltd

ellenh1001@163.com, liweiyeahmail@163.com, raogaoqi@blcu.edu.cn, jojozang@tencent.com

Abstract

The register (“register”is a term tentatively used here for the Chinese term yuti(语体)) words refer to words that are used exclusively in a certain register, and are the language elements and formal marks of the register. The resources of register words can serve NLP applications closely related to real-life scenarios, but such resources are relatively scarce at present. In this regard, based on the “DaCiLin”, this paper has completed three tasks of “register words tagging”, “register (words) chain tagging”and “parallel construction tagging”, and established a register categorized resources which based on register words. This resource contains 55,710 words, 5,017 register chains and 433 sets of parallel constructions. Based on this, this paper analyzes the distribution of Chinese register words, morphological differences, and the distribution of semantic parts of speech.

Keywords: register words , register categorized resources , synonym

*通讯作者corresponding author

本文系教育部人文社科基金“清末以来汉语报刊词汇使用计量研究”（20YJC740050）阶段性成果

第二十一届中国计算语言学大会论文集，第431页-第443页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

语言为了完成不同的功能、适应不同的语域而形成的相对恒定的表达模式，被称为语体。而不同的词语在不同的场合中使用，有不同的功能，表现出不同的语体色彩，形成了语体词这一概念。袁晖(2004)等学者指明，语体词是语体的语言要素和形式标记，是认识和研究语体的锁钥之一。但当前学界语体词的研究并不丰富，多是关注语体词本身的特征；且语言实际使用场景中，选取哪个词语并没有相关的参考，则是依靠使用者本人的语言习惯。所有的NLP应用都需要在现实语言中进行交互，因而其产出和输入都从属于某种语体。通过对语体词的研究可以较好地提高自动校对、语言润色等自然语言生成任务的用户体验。目前NLP研究中对语义关注较多而对语体缺乏研究。例如:

eg1:他的行为对全体职工的工作鼓舞很大。→ 他的行为极大地鼓舞了全体职工的工作热情。

eg2:这件事得跟我们头儿说。→ 这件事情需要向我们领导汇报。

以上例子中前者是来自于真实语料，后者则是修改了部分词语，可以看出通过变换词语能使得语言更为正式，表达更为得体。因此以语义为基础，发掘同一语义下对语体词的分类具有一定的可行性和必要性。目前工业界和学术界广泛使用的《哈工大信息检索研究室同义词词林扩展版》（以下简称《大词林》）较为全面地覆盖了语言生活中常用的同义词词簇。本文尝试在提供了同一语义下选取词语的范围内进行语体的分类操作，并据此提出了“语体词标注”“语体（词）链条标注”和“平行构式标注”三个任务。它们旨在通过分类建立起以词语为基础的语体分类资源，更好地服务于语体研究，从而应用于自然语言处理以及对外汉语教学等领域。

2 相关研究

2.1 语体

语体一直是学界讨论的热点话题，对于语体的定义及分类，不断有学者提出自己的看法。唐松波(1984)提出语体是言语特点的综合。李泉(2004)认为语体即语言运用过程中产生的交际功能变体。冯胜利(2010)认为语体是一种交际手段，用来拉近、拉远或保持交际过程中双方的距离。关于语体的分类，以二分法为盛，即口语体和书面语体，不少学者再次细分出下位语体，主要有符淮青(1985)、胡裕树(1995)、邵敬敏(2001)等。此外，冯胜利(2010; 2017)提出“调距”功能角度，认为“正式体、非正式体、典雅体”为语体的三大基本范畴，崔希亮(2020)借鉴此分法，将语体区分为正式语体和非正式语体。

语体语法日益引起学界关注，选择不同层面的语体特征进行语体计量的研究也不断涌现。方梅(2013)通过不同语体材料的对比分析，说明句法特征具有语体分布差异。冯胜利等(2017)通过语体标注，从“量”“质”两方面证实了“语体不同，语法不同”。邰沁清等(2021)运用语料库和统计方法对汉语语体进行特征的计量研究，进一步实现自动分类任务。

2.2 语体资源

围绕语体资源的构建工作主要有语体语料库构建、语体词的词典编纂等。北京语言大学BCC语料库包含文学、报刊、对话、古汉语等多领域。北京大学CCL语料库中也构建了口语领域。冯胜利等(2017)构建了由叙事文、新闻、说明文等6类文体类型组成的12万字左右的语体语料库，从语法、韵律、语体信息三方面进行标注。关于语体词的词典编纂，《现代汉语词典》对于常用口语词、方言词、书面上的文言词语，分别标注<口><方><书>。此外，还有闵家骥等(1991)编著的《汉语方言常用词词典》、施光亨等(2012)编著的《汉语口语词词典》等。

2.3 语体词

对于语体词，部分学者从宏观上对语体词的分类、适用范围与构成进行了研究。关于语体词的分类，目前学界以三分法为主流，将现代汉语词汇分为书面语词、口语词和通用语词，主要学者有曹炜(2003)、符淮青(2004)等。关于语体词的适用范围，谢智香(2011)认为“口语词汇在日常口头交际中所使用，一般具有通俗易懂、风趣幽默的风格；书面语词汇在正式的交际场合使用，一般具有典雅、庄重的色彩”。关于语体词的构成，刘中富(2003)指出口语词汇除日常口语用词外，还包括俗语词以及方言词语，书面语词汇包括历史词语、文言词语、行业词语、

生僻的和较典雅的成语，本文认同以上说法。还有语体词专项研究，主要有苏新春(2007)、尹惠贞(2006)、张安娜(2015)等。

近年来，基于语料库的语体词计量越来越引起学者们的重视。张文贤等(2012)计算出1343对具有显著口语、书面语语体差异的同义词，得出“口语、书面语的同义词差异主要在词性以及音节上”。宋婧婧(2013)以有声媒体与平面媒体语料库作为口语与书面语的代表，对其使用词汇进行词频、词类、音节的定量对比。张佩(2021)经过BCC 语料库及其他语体材料的测量，对汉语作为第二语言易混语体词汇的教学提出建议。

总结前人研究可知，有关语体、语体词的研究数量颇丰，语体计量的研究也层出不穷。程雨民(2004) 指出“语体建立在同义性的基础上”“语体的实质是在一些使用场合上有区别的同一变体的选择”，张文贤等(2012)也认同此观点，因而平行语体资源的构建具有重要意义。但是，目前学界在此方面的工作缺乏，仅在语体词层面存在少量平行资源，且数量和规模非常有限，对于自然语言处理等相关应用的支持不足。因此，建设多层面、大规模的平行语体资源对语体的理论研究以及对外汉语教学、自然语言处理等应用领域均具有重要价值。本研究基于《大词林》，构建了由语体分类词表、语体链条和平行构式三个层面组成的平行语体分类资源，并从资源分布、语体词形态差异、词义分布与词性分布四方面展开分析。

3 语体词林资源建设

3.1 基础资源

《大词林》在《同义词词林》(梅家驹, 1983)的基础上，参照多部电子词典资源，并按照人民日报语料库中词语的出现频度，只保留频度不低于3（小规模语料的统计结果）的部分词语，剔除14,706个罕用词和非常用词，并进行扩充，最终的词表包含77,343条词语。《大词林》按照树状的层次结构把所有收录的词条组织到一起，把词汇分成大、中、小三类，经统计，大类有12个，中类有95个，小类有1,400个。目前基于《大词林》的研究主要集中在语义层面，语体层面的工作鲜有。而词汇是语言的建筑材料，语体词反映语体、甚至在一定情况下能决定语体，《大词林》拥有的丰富的同义词簇恰能为构建平行语体资源提供重要基础。因此，本研究以《大词林》为基础，构建了由口语词表、通用语词表、书面语词表、术语词表和多义词表组成的《语体分类词表》（下称《词表》）。

3.1.1 语体词标注规范

为提高语义对应的准确性，本研究仅对《大词林》的同义词簇进行语体词的划分，即由“=”连接的词簇，筛选得出9,995组词簇，55,844条词语。

关于语体词的分类，本研究采用学界主流的三分法，即将现代汉语词汇分为口语词、通用语词和书面语词，对于书面语词中的专业术语，单独建立术语词表，四者互不重合。

本研究的语体词划分方式将《现代汉语词典》（以下简称《现汉》）、BCC语料库测量和理论研究三者相结合，主要有以下考虑：《现汉》是一部规范、权威的语文词典，其中标<书>和<口><方>的词条可以为划分语体词提供直接依据，并可对具有多个义项的多义词语体进行精细的判定，但《现汉》标记的是最典型的语体词，判定作用有限。BCC语料库作为全面反映当今社会语言生活的大规模语料库，其中的报刊、对话领域可大致代表书面语体和口语体，通过观察词语在报刊、对话等领域的数量，可以较为科学客观地对词语语体进行量化测量。但使用BCC语料库难以对多义词各义项间的语体差异进行细致分辨；且报刊及对话领域仅能大致代表两种语体，并非完全泾渭分明，且受语料来源、词语使用范围等因素影响，结果有时与平日认知有偏差。通过总结前人理论研究成果，归纳典型语体词的判定方式可一定程度上规避此问题。本研究参考宋婧婧(2015)、陈振艳(2016)及高艳(2017)的研究成果，总结出16条口语词及3条书面语词的判定方式，并配以大量举例。但理论总结无法穷尽所有可能，主观成分较多。综上，《现汉》、BCC语料库测量和理论研究各有优劣，因此本研究将三者进行结合。

由上所述，多义词的不同义项在语体划分中起到直接作用。因此，本研究根据《现汉》，筛选出《大词林》中的多义词共12,463条，标注了多义词的词义、当前义项及对应例句。

Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China.

3.1.2 语体词标注实践

根据《大词林》语体词标注规范，由相关专业的10名本科生、研究生开展标注，具体如下：

(1) 标注多义词。根据《现汉》及词簇中其他同义词的含义，标注多义词的当前义项，并配以1-3句例句。无法在《现汉》中找到当前义项的多义词，做如下处理：若百度及其他词典资源，包括汉语大词典、百度汉语等中有此义项，则进行补充并备注来源，如“无对应义，参考百度汉语补充”；若均无法找到当前义项，则备注“无对应义”，此词语将不参与语体划分。

(2) 标注术语。通过“术语在线”平台对术语进行提取与标注。“术语在线”平台由全国科学技术名词审定委员会于2016年5月创办上线，为目前较为权威的术语知识服务平台。标注术语时，仅保留审定公布库中的规范术语。

(3) 标注语体。首先据《现汉》直接进行划分，对于标记<口><方>和<书>的词语，分别划分为口语词和书面语词。其次，依据理论总结的判定方式，对典型口语词和书面语词进行划分。对于其中语体不确定以及剩余未分类词语，参考BCC语料库中的例句进行判读：在语料库各领域中均出现的词，划分为通用语词；在对话领域中出现次数为0或次数很少，基本存在于报刊、科技领域中的词，划分为书面语词，反之，划分为口语词；在各领域出现次数均小于3的词，备注为“罕用词”，不参与语体的划分。

(4) 对口语词、通用词、书面语词、术语词、多义词分别标记符号O、G、W、T、M。对于复合词，标记时语体词符号在前，多义词符号在后，如“宝贝OM”。

对标注结果进行多轮校对后，最终我们构建了基于《大词林》的《词表》，共包含9,992组词簇，55,710条词语。

3.2 语体链条标注

3.2.1 语体链条标注规范

基于上述对《大词林》的分类，可以得到某一具体语义下的口语词、通用语词以及书面语词，通过对数据的整理和挖掘，可以发现相同语义下口语词到通用语词再到书面语词的转化和替换。而这一具体语义下，从口语词至通用语词至书面语词的链条，本文称之为“语体链条”。如：“俺→我→吾”，三种语体词可以表达同样的语义。具体的标注规范有：

(1) 选择适当的词组成链条。在《大词林》的词簇完成分类后，同一语体下存在多个词语，要保证各个语体词语义要基本对等，主要参照《现汉》词义以及之前给出多义词的当前义项。词语要保持感情色彩一致；时代色彩浓厚的词语进行剔除。词语之间用“→”隔开表示。

(2) 通过给语体词配以短语搭配或者例句来保持当前词义一致。《大词林》中的词语存在大量的多义词，在不同的上下文中有不同的意义，为保持当前的词义一致，选取恰当的短语搭配和例句可以帮助限制语体词的具体语义，可参照《大词林》分类形成的多义词例句，或参考BCC语料库以及CCL语料库，也可自行造句。对应链条的位置填入“搭配/例句”一栏中。

(3) 词语的词义为固定义。一般在词典中有记录的意义为词的固定义。但是有时词语在使用中的意义在词典中找不到，使用的是临时义，如修辞义。词语的临时义变动大并不固定，需要依靠相当篇幅的上下文，有一定的使用限制，因此本研究不考虑这样的临时义。

(4) 本轮标注不要求同时存在三种语体词，允许存在同一语体的多个语体词。同时存在意义相同的口语词、通用语词和书面语词是理想的情况，但并不会大量存在，因此允许链条存在两种语体词，在并不存在的那一类语体位置上，用“？”表示。也有许多词语可以在表示相同的语义下，仍然是出现在同一种语体中，这样的情况用“/”进行隔开。

3.2.2 语体链条标注实践

根据上述语体链条的标注规范，招募了11名语言学及应用语言学专业的研究生进行实际标注，并有3名质检员进行检查修改，经多轮培训和修改后完成标注任务。

(1) 标注得到大量语义基本一致的语体链条以及语体分明且语义一致的短语搭配及例句。而以语体词为基础，为语体研究提供了一个新的角度。短语搭配和例句与语体链条相对应，形成了一批具有语体色彩的平行语料，也具有重要的价值。

(2) 标注发现存在许多的单音节语素、专有名词。在标注过程中，存在许多如“青”“紫”“木”这样的单音节语素，它们很少单独使用，大多数情况则是组成词语，因此链条中不再选取。专有名词如“江淮戏”“淮剧”，两词意义完全相同，只是不同时期叫法不同，并且“江淮戏”已经不再出现在新时期的语料之中，这样的词语也不再收入链条之中。

(3) 标注发现存在许多意义泛化的术语，可以重新进行归类。一般情况下，认为术语是存在于书面语中，是书面语词的一部分，但是有许多术语在日常生活中广泛使用，如“冰雹”等，在口语和书面语中都有许多使用，也可以根据分类规范放入通用语词之中。

3.3 平行构式标注

Goldberg(1995)提出，构式就是指这样的形式—意义对,它在形式或意义方面所具有的某些特征不能完全从其组成成分或业已建立的其他构式中推导出来。构式语法的应用非常广泛(施春宏, 2017),且在实际应用中，词语语体转化必然影响其周边成分，构式信息因而显得格外重要。因而建设以链条词为核心的构式资源是一种必然选择。此时，平行构式不再需要严格进行语体三分，由非正式语体至正式语体的转换即可达到交际需求，即正式程度有所提高即可。如“打”后接名词性短语，在名词性短语表示织物时，“打”可以转化为“织”。本研究中将“打+np[2]”作为构式，当“np”指代“织物”时，与“织+np”组成平行构式。

3.3.1 平行构式标注规范

(1) 平行构式以词语为核心，参照BCC语料库检索规则，以词类搭配和特殊符号作为限制，词类限制有名词性短语、动词性短语、名词、动词、形容词等，特殊符号则包括标点符号“w”和分句符号“sent”等。词语与词类或特殊符号之间用“+”连接。

(2) 必要情况下需要进行语义限制。符合构式限制的语义，才可以进行构式的转化。若无语义限制，则可以直接转化构式。若无法归类语义，可以尽量穷举出该类别可搭配的词语。

(3) 非正式语体构式与正式语体构式成对存在。两种语体构式需要对应存在。

(4) 若无形式化描述，可以组成平行短语。即使有语义限制，也仍存在一些词语只在某些常用搭配中才会进行语体上的替换。因此，不存在或难以产生构式时，则产出平行短语。

3.3.2 平行构式标注实践

标注得到了一批有语义限制的平行构式，当词语搭配满足语义限制时，则可以进行构式转化；同时，也形成了一批由非正式到正式的可以无条件转换的平行短语。这些词语搭配以及平行构式所配有对应的实例集合，构成了一批由非正式语体至正式语体的平行语料。

4 资源分析

4.1 语体分类资源分布

4.1.1 语体词分布

《词表》的55,710 条词语中，口语词为5,743条，通用语词为23,133条，书面语词为26,834条（其中术语词3,713条），分别占比10.3%，41.5%和48.2%。由于《同义词词林》(梅家驹, 1983)编著初衷主要服务于翻译和写作，且在进行剔除与扩充时主要参照多种词典资源和人民日报语料库，因此词语书面化程度较高。本文统计了9,992组同义词簇的语体差异情况：

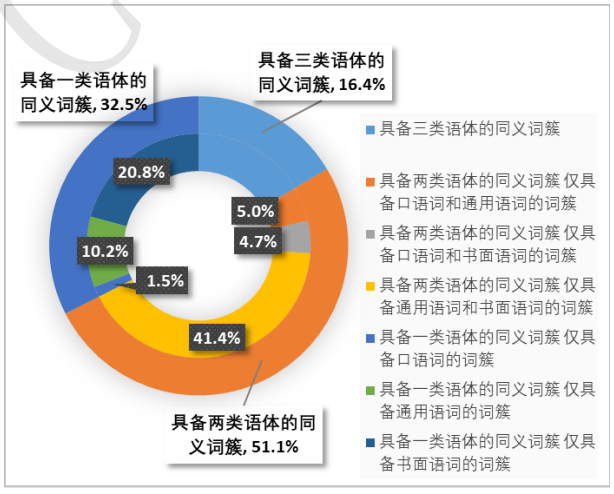

图 1: 《词表》中同义词簇语体差异情况

[2] “np”表示名词性短语。

由图1可知，有32.5%的同义词语体单一，并无其他对应语体；51.1%的同义词具备两类语体，这类同义词占比最多，以具备通用语词和对应书面语词的同义词为主，比例高达总数的41.4%；三类语体都具备的同义词仅占16.4%。可见，语体差异只是同义词差异的一个方面，部分同义词并无语体方面的明显差异。

4.1.2 语体链条和平行构式分布

本研究得到语体链条5,017条，搭配或例句有4,432条，可以得知，同义词的语体差异是存在的，在同义词辨析中值得关注；同时，这种语体差异是非常依赖上下文语境的。

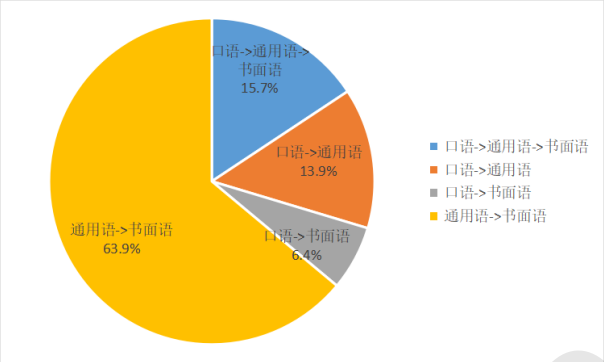

图 2: 语体链条分布

如图2所示，语体链条中，“口语词→通用语词→书面语词”链条有788条，占全部链条的15.7%，“口语词→通用语词”有699条，占全部链条的13.9%，“口语词→书面语词”链条有322条，占全部链条的6.4%，“通用语词→书面语词”链条有3,208条，占全部链条的63.9%。从中可以得知，同时出现在三种语体的语义极少，这也符合词义的概括性，词语反映的是一类事物或现象共同的特征；通用语词到书面语词的链条占比最高，一方面与《大词林》中收录文言词语相关，另一方面也反映出书面语词具有特殊的作用，有专门的使用场合。

最后，本研究得到非正式至正式的平行构式130对，是语体链条的2.6%。平行构式以语体链条为基础而得出，平行构式的数量相比语体链条是极少的，语体链条向平行构式的转化率不高，所以语体差异虽然可以用语法构式表现出来，但是效果并不理想，可以进一步改进。

4.2 语体词形态差异

4.2.1 语体词的词长

口语词、通用语词和书面语词在词长上各自具有其显著特征。《词表》中各语体词的词长占比统计如下：

	口语词	通用语词	书面语词
单音节	31.0%	14.1%	13.3%
双音节	37.6%	72.7%	52.3%
三音节	25.3%	5.7%	8.6%
多音节	6.1%	7.5%	25.3%
总计	100.0%	100.0%	100.0%

表 1: 《词表》中各类语体词词长占比

据表1可看出，在三类语体词中，双音节词均占比最高，尤其在通用语词中，双音节词占比达到了72.7%，其次是书面语词，也达到了52.3%，符合现代汉语中双音节词占优势这一基本特征。口语词中包含许多单音节实词，惯用语和俗语也多为三音节词，因此在口语词中单音节词和三音节词均占有一定比重，分别为31.0%和25.3%，这一占比远高于通用语词和书面语词。书面语词中具有大量的四字成语及类固定短语，其多音节词占比达25.3%，较口语词和通用语词高。对口语词、通用语词、书面语词的平均词长进行计算，得出三者分别为2.08、2.07、2.40，其中书面语词词长最大，口语词和通用语词二者相当。

分析以上差异产生的原因，主要是口语词较日常、随意，具有充分的语境及交际双方肢体动作、表情等辅助，允许表达的简洁、灵活而不会产生歧义，因此单音节词和三音节词占比较

高；书面语词较庄重、典雅，且语境较弱，需要更充分的表达以确保语义的准确，因此双音节词和多音节词占比较高，平均词长最大。

4.2.2 语体词构成语素

本研究调查了有多少共有语素是可以突破语体隔阂得以保存的，即在同一个链条或构式当中，有多少语素同时出现在口语词、通用语词和书面语当中。

	口语词→通用语词	通用语词→书面语词	口语词→书面语词	口语词→通用语词→书面语词
共有语素	445	1631	207	272
出现共有语素的链条	487	2545	196	301
语体链条	699	3208	322	788

表 2: 不同链条中共有语素数量

据表2可知，在不同语体中存在有许多共同出现的语素，其中，通用语词至书面语词的链条出现了最多的共有语素，有1,631个；其次是口语词至通用语词的链条，有445个；再次，口语词至通用语词至书面语词链条中共有语素，有272个；而口语词至书面语词链条中的共有语素最少，有207个，这与本身语体链条数量是相关的。与语体链条一致，横跨两种语体的语素数量较多，但是横跨三种语体的语素数量较少。另外发现，口语词至书面语词链条中，出现共有语素的链条数量少于共有语素的数量，说明该类链条中，跨语体存在的语素不止一个。

口语词→通用语词		口语词→通用语词		口语词→通用语词		口语词→通用语词	
语素	频次	语素	频次	语素	频次	语素	频次
下	7	不	28	鱼	4	老	5
手	6	人	25	小	3	年	4
不	5	风	21	前	3	实	4
后	4	心	20	家	2	不	4
子	4	信	17	子	2	大	3

表 3: 语体链条中出现最多的Top5语素及其频次

本研究继续统计了共有语素中出现频次最多的前5个语素，列举在表3中，可以看到有部分共有语素在不同链条中都会出现，如“下”“不”。

本研究同时也调查了平行构式中的共有语素，发现在130条构式中，有92条中出现了共有语素，这也说明同一语素是经常出现在不同语体中的。

4.3 语体词词义分布

4.3.1 《词表》词义分布

各语体词不仅在形态上具有差异，在词义分布上也各具特色。我们首先对《大词林》中各语体词的义项多少进行了统计。在《词表》的多义词中，实际标注语体的多义词共有12,414条，根据标签对各类多义语体词的数量及占比进行统计，结果如下：

	口语词	通用语词	书面语词	总计
数量（个）	1803	6431	4180	12414
占比	14.5%	51.8%	33.7%	100.0%

表 4: 《词表》中各类多义语体词数量及占比

如表4所示，多义语体词中通用语词最多，占51.8%，其次为书面语词、口语词。接着我们对各语体词中单义词、多义词的占比进行计算，得出如下结果：

	口语词	通用语词	书面语词
单义词	68.6%	72.2%	84.4%
多义词	31.4%	27.8%	15.6%
总计	100.0%	100.0%	100.0%

表 5: 《词表》中各类语体词单义词、多义词占比

第二十一届中国计算语言学大会论文集，第431页-第443页，南昌，中国，2022年10月14日至16日。

从表5可知，各语体词中单义情况均占绝大多数，多义情况较少，且随着词语正式程度的增加，单义词占比逐渐上升，多义词占比逐渐下降。究其原因，亦与各语体词的风格特征、语境强弱有关。从口语词到书面语词，场合逐渐庄重，语境依赖减弱，因而要求词义的表达更为细微、精准，以适应场合，避免交际障碍。

此外，本研究也对《词表》中语体词的语义类别分布状况作了进一步考察。我们以《大词林》的95个中类作为语义范畴，对各中类的语体词数量进行统计，得出口语词为三类语体词中使用数量最多的中类有7个，通用语词为三类语体词中使用数量最多的中类有37个，书面语词为使用数量最多的语体词的中类有51个，由于后两者中类数量较多，我们只取语体词使用数量前5的中类进行展示。具体数据如下：

所属中类	词性	口语/通用语/书面语词	所属中类	词性	口语/通用语/书面语词
Ah亲人眷属	名词	**225**/97/198	Ke呼叹	虚词	**47**/0/0
Kf 拟声	虚词	**168**/0/3	Kd辅助	虚词	**36**/13/22
Bc物体的部分	名词	**79**/41/65	Ac体态	名词	**27**/15/26
Bb拟状物	名词	**62**/22/30			

表 6: 《词表》中数量最多的语体词为口语词的中类（个）

由表6可知，当表达作为名词的亲人、眷属、物体的部分、拟状物、体态，比如“侄儿、大舅子、把子、耳子、疙瘩、片片、丑八怪、瘌痢头”等，以及作为虚词的拟声、呼叹、辅助的语义范畴，比如“叽里呱啦、吧唧、嘿、嗨、嘀哟、哇”等时，使用口语词居多。

所属中类	词性	口语词	通用语词	书面语词
Ed性质	形容词	469	**1623**	1280
Hj生活	动词	158	**1197**	857
Ka疏状	虚词	77	**858**	470
Gb心理活动	动词	89	**848**	394
Hc行政管理	动词	25	**623**	291

表 7: 《词表》中数量最多的语体词为通用语词的中类Top5（个）

由表7可知，当表达作为形容词的性质、境况，如“诚实、优秀、热闹、拥挤”，作为动词的生活、心理活动、行政管理，如“生活、过夜、想到、估计、安排、点名”，以及作为虚词的疏状的语义范畴，如“十分、最多”等时，使用通用语词居多。

所属中类	词性	口语词	通用语词	书面语词
Hi社交	动词	286	922	**1510**
Ee德才	形容词	250	1111	**1225**
Eb表象	形容词	250	751	**1206**
Dk文教	名词	60	658	**845**
Bh植物	名词	60	231	**806**

表 8: 《词表》中数量最多的语体词为书面语词的中类Top5（个）

由表8可知，当表达作为动词的社交，如“缔交、晤面”，作为形容词的德才、表象，如“笃实、披肝沥胆、寥若晨星、什锦”，以及作为名词的文教、植物的语义范畴，如“仰韶文化、彩陶文化、古柏、翠柏”等时，使用书面语词居多。

分析上述差异产生的原因，正如前文所述，口语词来源于日常并且较为随意、主观，因此涉及日常生活和显示出较强主观情感的人、物、拟声、呼叹等语义范畴时，多使用口语词。书面语词较为正式、严肃和客观，因此涉及社交以及表名物的文教、植物等语义范畴时，多使用书面语词。通用语词使用的范畴则更为广泛、多样。

4.3.2 语体链条和平行构式词义分布

本研究对于语体链条和平行构式在不同词义的分布方面也进行了统计，并进一步统计了语体链条数量占该词义类别词簇数量的比例，平行构式的数量以及平行构式数量占该类语体链条数量的比例与平行构式数量占该类意义《大词林》词簇数量的比例。如下图：

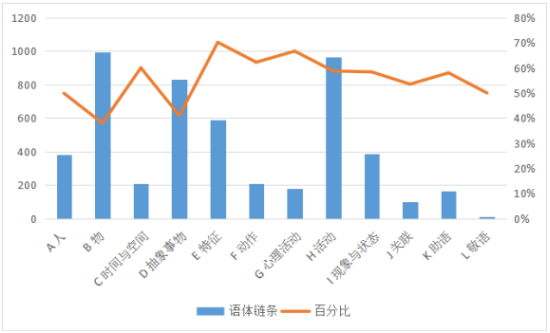

图 3: 不同词义类别中的语体链条

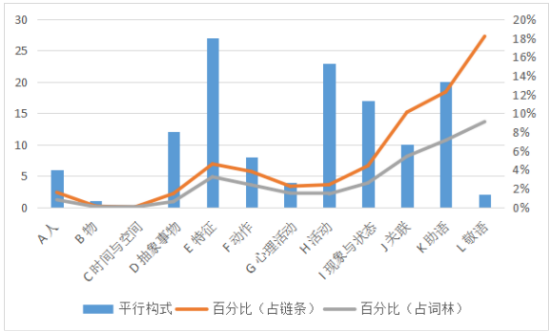

图 4: 不同词义类别中的平行构式

从图3可知，表示“物”和“活动”词义的词语中产出的语体链条数目最多，说明在这两项意义上，词语更为丰富，在语体上有所选择。但是从语体链条占《大词林》该类语义总词簇数的百分比来看，“物”的百分比最低，可以得知，虽然该词义下的词语存在语体的丰富性，但是同一意义下的可转化性不强，即尽管“物”义的词簇数量多，但是仅存在语体差异的同义词较少，该类意义的词语相对而言可以更多地同时存在于多种语体，并不完全需要因语体差异而改变词语的使用。另外，表示“特征”意义的语体链条占该意义下词簇数的比例最高，说明表示同一“特征”义的词语语体链条的能产性更强，在不同语体中更有可能有更多的选择,对于该类词簇中表示同一意义的词语与语体有极强的相关性，词语使用时要注意与所在语体相符合。

由图4可以发现：首先，平行构式占语体链条的比例走向与平行构式占该类意义《大词林》词簇数的比例走向大致是一致的，即分类的语体词中可以产出一定的语体链条，那么也可以相应地产出一定的平行构式，这也印证平行构式的出发点是可靠的。其次，各类意义下的平行构式占《大词林》词簇数和语体链条的比例均低于20%，可以看出，平行构式产出比例比较低，这与语义并不直接相关，产出只有语体差异的语法构式比较困难。最后，表示“关联”“助词”和“敬语”的平行构式百分比均超过了平均值，这说明该类意义中构式在不同语体下是非常丰富的，更容易在分类语体词和语体链条的基础上产出，更具有能产性。

4.4 语体词词性分布

4.4.1 《词表》词性分布

口语词、通用语词和书面语词在词性分布方面也不尽相同。本研究对《词表》中各语体词的词性占比进行了统计，数据如下：

	口语词	通用语词	书面语词
名词	41.2%	35.6%	46.6%
形容词	22.5%	18.9%	16.5%
动词	30.0%	40.5%	34.1%
虚词	6.2%	4.7%	2.7%
客套语	0.1%	0.3%	0.1%
总计	100.0%	100.0%	100.0%

表 9: 《词表》中各类语体词词性占比

从表9可知，在口语词和书面语词中，均为名词占比最高，分别高达41.2%和46.6%，其次分别是动词、形容词、虚词、客套语，在通用语词中，则是动词占比最高，达到40.5%，其次是名词、形容词、虚词和客套语。并且，书面语词的名词占比在三类语体词中最高，口语词中形容词和虚词的比重较另外两类更高，通用语词中动词和客套语的占比则为三类语体词中的最高。

4.4.2 语体链条和平行构式词性分布

语体链条和平行构式在词性分布方面也有其特征，本研究统计了语体链条和平行构式在不同词性下的数量，以及其占《大词林》该词性词簇数比重。

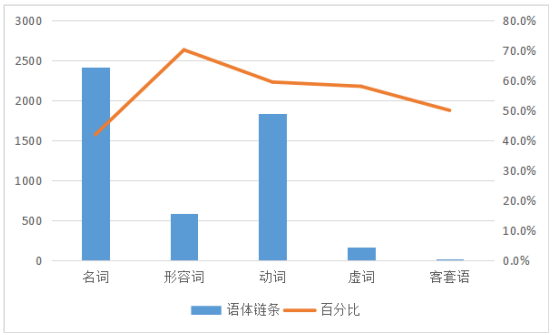

图 5: 不同词性中的语体链条

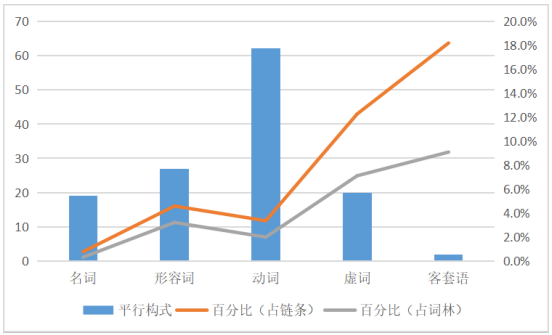

图 6: 不同词性下平行构式的分布

据图5可知，名词性的语体链条数量最多，其次是动词和形容词。形容词性的语体链条数目虽然不多，但是其占《大词林》形容词词簇数的比重最高，这也看出同义的形容词在不同语体中选择更多，表现更丰富；而名词性同义词在各语体中的选择更少，对于语体的敏感性稍弱。

如图6所示，与语体链条不同，动词性的平行构式数量最多，其次是形容词、虚词、名词和客套语，可以看出当词语范围扩展至构式时，动词仍然有比较好的表现。但动词性的平行构式占该词性的语体链条和词簇比重较低，而虚词和客套语的平行构式与之相反，说明这两类词性的平行构式更具有能产性，可在已有分类语体词和语体链条的基础上较好地进行扩充，这与平行构式词义上的分布趋势一致，亦是因为虚词和客套语的词性与助语和敬语的词义相对应。

5 结论

本研究基于《大词林》提出了语体词标注、语体（词）链条标注和平行构式标注三个任务，构建了一系列的语体分类资源，得到了《语体分类词表》、语体链条以及平行构式。《词表》中共包含9,992组词簇，55,710条词语；语体链条有5,017条，搭配或例句有4,432条；非正式至正式的平行构式130对，并且人工根据例句补充了303个平行构式。对应地制定了语体词标注规范，语体链条标注规范和平行构式标注规范。进而，本文对于语体分类资源进行了分析，描述了语体词、语体链条和平行构式在不同语体中的分布概况和形态差异：语体差异在同义词中值得关注，各语体词在词长、语义范畴与词性分布方面各具特色，不同语义与词性下，语体链条与平行构式的产出能力也不尽相同。

通过构建语体分类资源，可以为对外汉语教学和汉语作为第二语言的习得提供许多帮助。本资源也可以辅助教材编写，不同阶段和不同领域应有所侧重。其次，本资源提出的是中文语体相关的标注任务，相关的标注规范及实践的逻辑和经验也可以迁移至其他语言，其他语言可以以同义词词典为基础，依据本语言的相关语言资源及语体特点制定语体规范，得到有语体分类的词表，在此基础上进一步得到语体链条和平行构式，从而获得其他语言的语体资源。同时本资源可帮助进行语体改写、自动校对、语言润色等NLG工作，并且已经在腾讯文档中得到应用，起到了支持性的作用，取得了良好效果。

《大词林》的词语中含有许多文言文成分，各部分比重并不均衡，语义分布、词性分布都有其特点，后续可以对于这些问题进行改进。另外，《大词林》的词语资源虽然已经比较丰富，但产出的语体资源规模可能并非足够适用于庞杂的语言现实，因此可以在现有语体资源基础上进行实际的NLP测评任务，进一步体现其实际效能，同时，未来研究可以在现有资源基础上继续寻求语体词语资源，扩大语体词、语体链条和平行构式的规模。

参考文献

Wanxiang Che, Zhenghua Li, Ting Liu. 2010. Ltp: A chinese language technology platform. In *Coling 2010: Demonstrations*, pages 13–16.

Goldberg, AdeleE. 1995. *Constructions:A Construction Grammar Approach to Argument Structures*. Constructions:A Construction Grammar Approach to Argument Structures.

Qinqing Tai, Gaoqi Rao. 2021. 汉语语体特征的计量与分类研究(a study on the measurement and classification of chinese stylistic features). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 398–412.

冯胜利, 王永娜. 2017. 语体标注对语体语法和叙事、论说体的考察与发现. 汉语应用语言学研究, (1):15.

梅家驹. 1983. 同义词词林. 上海：上海辞书出版社.

张佩. 2021. 基于BCC语料库的词汇语体属性研究. 渤海大学.

冯胜利. 2010. 论语体的机制及其语法属性. 中国语文, (5):13.

刘中富. 2003. 实用汉语词汇. 安徽：安徽教育出版社.

唐松波. 1984. 文体、语体、风格、修辞的相互关系. 当代修辞学, (2):2.

宋婧婧. 2013. 汉语口语与书面语词汇使用对比分析——基于传媒语料库. 厦门理工学院学报, 21(3):88–92.

宋婧婧. 2015. 现代汉语口语词研究. 厦门：厦门大学出版社.

尹惠贞. 2006. 现代汉语口语词汇研究. 北京：北京语言大学.

崔希亮. 2020. 正式语体和非正式语体的分野. 汉语学报, (2):12.

张安娜. 2015. 现代汉语书面语词和口语词差异及其对应关系研究. 华东师范大学.

张文贤, 邱立坤, 宋作艳, 陈保亚. 2012. 基于语料库的汉语同义词语体差异定量分析. 汉语学习, (3).

方梅. 2013. 谈语体特征的句法表现. 当代修辞学, (2):8.

施光亨. 2012. 汉语口语词词典. 北京：商务印书馆.

施春宏. 2017. 构式语法的理论路径和应用空间. 汉语学报, (1):2–13.

曹炜. 2003. 现代汉语口语词和书面语词的差异初探. 语言教学与研究, 2003(6).

李泉. 2004. 面向对外汉语教学的语体研究的范围和内容. 汉语学习 ,2004(1).

程雨民. 2004. 英语语体学. 上海: 上海外语教育出版社.

符淮青. 2004. 现代汉语词汇（增订本）. 北京：北京大学出版社.

符淮青. 1985. 现代汉语词汇. 北京:北京大学出版社.

胡裕树. 1995. 现代汉语(重订本). 上海:上海教育出版社.

苏新春, 徐婷. 2007. 《现代汉语词典》标“书”词研究(下)——兼谈与古语词, 历史词, 旧词语的区别. 辞书研究, (2):38–44.

荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 2016. 大数据背景下BCC语料库的研制. 语料库语言学, 2016(1).

詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙. 2019. 北京大学CCL语料库的研制. 《语料库语言学》2019年第6卷第1期,总第11辑,pp.71-86.

谢智香. 2011. 论现代汉语口语词的特点. 西南石油大学学报*(*社会科学版*)*, 4(3):103–106.

邵敬敏. 2001. 现代汉语通论. 上海:上海教育出版社.

闵家骥. 1991. 汉语方言常用词词典. 浙江：浙江教育出版社.

陈振艳. 2016. 成语和类固定短语的语体鉴别及语体动因. 浙江树人大学学报*(*人文社会科学*)*, 6.

袁晖. 2004. 论语体词. 修辞学习.

高艳. 2017. 现代汉语口语词的主要类型及基本特征. 海外华文教育, (9):1188–1199.

A 附录.各中类语体词统计表（个）

所属中类	口语词	通用语词	书面语词	标准差
Aa泛称	62	75	98	14.9
Ab男女老少	100	66	102	16.5
Ac体态	27	15	26	5.4
Ad籍属	7	25	27	9.0
Ae职业	83	202	448	152
Af身份	27	92	323	127.0
Ag状况	50	59	152	46.1
Ah亲人眷属	225	97	198	55.1
Ai辈次	19	22	74	25.2
Aj关系	68	205	165	57.5
Ak品性	95	68	122	22.0
Al才识	85	72	133	26.2
Am信仰	3	25	29	11.4
An丑类	52	46	105	26.5
Ba统称	43	107	263	92.4
Bb拟状物	62	22	30	17.3
Bc物体的部分	79	41	65	15.7
Bd天体	5	26	92	37.1
Be地貌	16	85	263	104.1
Bf气象	3	68	113	45.2
Bg自然物	24	144	322	122.4
Bh植物	60	231	806	319.1
Bi动物	132	148	471	156.2
Bj微生物	1	1	22	9.9
Bk全身	129	150	603	218.7
Bl排泄物分泌物	10	25	54	18.3
Bm材料	36	110	245	86.5
Bn建筑物	34	260	448	169.3
Bo机具	116	185	710	265.3
Bp用品	73	390	586	211.4
Bq衣物	21	146	120	53.9
Br食品药品毒品	76	264	231	82.0
Ca时间	71	581	278	209.4
Cb空间	60	395	495	186.0
Da事情情况	54	655	682	289.9
Db事理	30	176	139	62.0
Dc外貌	19	151	144	60.6
Dd性能	36	421	427	182.9
De性格才能	5	129	121	56.7
Df意识	19	371	279	149.1
Dg比喻物	20	72	64	22.9
Dh臆想物	12	81	66	29.6
Di社会政法	28	466	576	236.7
Dj经济	13	130	257	99.6
Dk文教	60	658	845	334.8
Dl疾病	31	33	181	70.2
Dm机构	20	129	276	104.9

Dn数量单位	64	313	242	104.7
Ea外形	76	239	223	73.4
Eb表象	250	751	1206	390.4
Ec颜色味道	139	190	104	35.3
Ed性质	469	1623	1280	483.9
Ee德才	250	1111	1225	435.2
Ef境况	109	469	399	155.8
Fa上肢动作	198	525	309	135.8
Fb下肢动作	21	122	100	43.4
Fc头部动作	50	215	226	80.5
Fd全身动作	18	95	72	32.3
Ga心理状态	63	405	789	296.6
Gb心理活动	89	848	394	311.8
Gc能愿	19	109	9	45.0
Ha政治活动	4	127	76	50.5
Hb军事活动	11	317	128	126.1
Hc行政管理	25	623	291	244.6
Hd生产	8	237	178	97.1
He经济活动	11	324	173	127.8
Hf交通运输	8	166	42	67.9
Hg教卫科研	112	298	437	133.1
Hh文体活动	23	76	53	21.7
Hi社交	286	922	1510	499.8
Hj生活	158	1197	857	432.5
Hk宗教活动	6	30	13	10.1
Hl迷信活动	4	10	6	2.5
Hm公安司法	35	80	166	54.3
Hn恶行	42	75	147	43.8
Ia自然现象	49	70	156	46.3
Ib生理现象	61	242	410	142.5
Ic表情	39	116	253	88.5
Id物体状态	50	336	464	173.1
Ie事态	35	298	326	131.1
If境遇	69	276	472	164.5
Ig始末	7	158	140	67.3
Ih变化	54	252	312	110.2
Ja联系	21	86	58	26.6
Jb异同	29	117	119	42.0
Jc配合	31	43	73	17.7
Jd存在	83	244	231	73.0
Je影响	5	332	149	133.8
Ka疏状	77	858	470	318.8
Kb中介	17	102	111	42.4
Kc联接	13	112	110	46.2
Kd辅助	36	13	22	9.5
Ke呼叹	47	0	0	22.2
Kf拟声	168	0	3	78.5
La敬语	3	61	24	24.0