《二十四史》古代汉语语义依存图库构建

黄恬　　邵艳秋　　李炜*
北京语言大学，信息科学学院，
国家语言资源监测与研究平面媒体中心，
北京市海淀区学院路15号，100083
huangtian_blcu@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

摘要

语义依存图是NLP处理语义的深层分析方法，能够对句子中词与词之间的语义进行分析。该文针对古代汉语特点，在制定古代汉语语义依存图标注规范的基础上，以《二十四史》为语料来源，完成标注了规模为3000句的古代汉语语义依存图库，标注一致性的kappa值为78.83%。通过与现代汉语语义依存图库的对比，对依存图库基本情况进行统计，分析古代汉语的语义特色和规律。统计显示，古代汉语语义分布宏观上符合齐普夫定律，在语义事件描述上具有强烈的历史性叙事和正式文体特征，如以人物纪传为中心，时间、地点等周边角色描述细致，叙事语言冷静客观，缺少描述情态、语气、程度、时间状态等的修饰词语等。

关键词： 古代汉语；语义依存图；二十四史

Construction of Semantic Dependency Graph Bank of Ancient Chinese in twenty four histories

Tian Huang　　Yanqiu Shao　　Wei Li*
Information Science School, Beijing Language and Culture University,
National Language Resources Monitoring and Research Center,
15 Xueyuan Road, HaiDian District, Beijing, 100083
huangtian_blcu@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

Abstract

Semantic dependency graph is a deep analysis method of computer processing semantics, which can analyze the semantic relationship of sentences. In view of the characteristics of ancient Chinese, this paper formulates the annotation guidelines of Ancient Chinese semantic dependency graph and constructs the semantic dependency corpus of ancient Chinese which contains 3000 sentences from the twenty four histories, achieving the kappa value of annotation consistency is 78.83%, Finally, through the comparison with the semantic dependency Gragh Bank of modern Chinese, anlyzing the basic situation of the dependency library, the semantic characteristics and laws of ancient Chinese. Statistics show that the semantic distribution of ancient Chinese conforms to Zipf's law macroscopically, and has strong historical narrative and formal stylistic features in the description of semantic events, such as taking biographies of characters as the center, detailed description of surrounding roles such as time and place, calm and objective narrative language, less modal particles, etc.

Keywords: Ancient Chinese , semantic dependency graph , Twenty-four Histories

* 通讯作者　Corresponding Author

第二十一届中国计算语言学大会论文集，第444页-第454页，南昌，中国，2022年10月14日至16日。

1 引言

语义分析是自然语言处理的核心问题，现代汉语领域进行了多种方法的研究，如基于词(刘琦, 2012)、基于短语(Xue, 2008)、基于句法树(Hacioglu, 2004)和基于依存图(Ding et al., 2014)的分析方法，取得了丰硕的成果。但古代汉语的语义研究相对匮乏，主要集中词汇层面上，如邹璐对《战国策》的副词进行意义归类(赵娟, 2005)，张丽丽对先秦帝王、诸侯谥号词汇进行语义分析归类并描绘出语义网(张丽丽, 2014)，舒蕾建设古汉语包含多义词的单音节词词义标注语料库(Shu et al., 2021)。在句子层级上，对句子的标注停留在句法关系，还未涉及深层次的语义关系标注，如京都大学建立了由《四书》构建的古代汉语依存树库(Yasuoka, 2019)。

随着时代发展，传统文化的专家研究和学习辅助亟待要求古代汉语信息化，对古代汉语语义分析提出了新的要求。文章以语义依存图理论为指导，结合古代汉语语法和语义特点，制定古代汉语语义依存图标注规范，并以《二十四史》为语料来源，对语料进行语义依存图标注，初步建立了3000句规模大小的古代汉语语义依存图库，对依存图库基本情况进行统计描述，并通过与现代汉语语义对比分析其语义特色。统计结果显示，《二十四史》语义标签频次分布总体上符合长尾分布，语义事件通过谓语动词紧密连接；叙事以人物事迹为中心，时间、地点等周边角色描述细致；使令句和目的句丰富，反映皇权制度下的上下等级制度和政治话语权；语言具有强烈的纪传体和历史题材风格特征。

2 现代汉语语义依存图和古代汉语语义依存图标注规范

2.1 现代汉语语义依存图

自然语言处理中传统的语义分析多采用语义依存树，依存树为句子中的每个词语（除核心词）找到它的依存词（父节点），并指出该词语与依存词之间的语义关系，传统的语义依存树结构规定每个句子成分只能有一个父节点与其存在依存关系，且不同的依存弧之间不能存在交叉现象。但汉语语序灵活、词类功能多样化，语言变式繁多，在真实语言现象中经常会出现某个词语同时依存于多个词语(王跃龙and 姬东鸿, 2007)，同一句子成分可能被多个成分支配，树结构不能满足汉语的真实语言现象表达，汉语的语义分析研究也从依存树走向依存图。如王悦龙提出构建一种全新的汉语依存图库(王跃龙and 姬东鸿, 2007)，哈尔滨工业大学(丁宇, 2014)结合鲁川定义的汉语意合网络语义关系体系和依存语法的特点构建了一套语义依存图关系体系，郑丽娟建立了包含30000句子的兼语句语义依存图库(郑丽娟and 邵艳秋, 2015)。图结构突破原有的依存树表达的限制，放宽了依存树的限制条件，主要表现在两个方面：（1）允许多父亲节点的出现；（2）允许非投射现象出现，即允许依存弧之间存在交叉(郑丽娟et al., 2014)。

如句子“我扶老奶奶上车”，用依存树分析的结构如图所示，句子的核心词为“扶”，“扶”直接支配“我”“老奶奶”和“上车”，他们之间的关系分别是施事、受事和后继关系，而“上车”还有另一动作参与者“老奶奶”则没有在树结构中直接指出。而用图结构分析，句子中的“我”有两个父节点“扶”和“上车”，都承担施事角色；“老奶奶”有两个父亲节点“扶”和“上车”，分别承担受事和施事角色。针对这个句子，机器自动问答想知道是“谁”“上车”，依存树不能准确分析出正确结果，依存图则能得出“我”和“老奶奶”两个动作参与者。

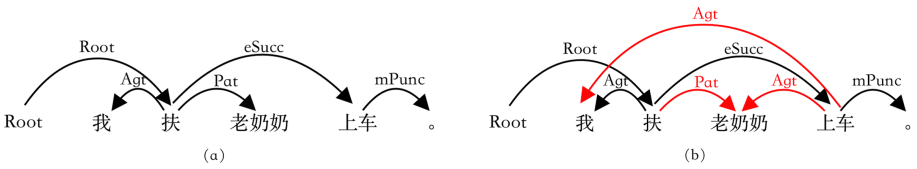


Figure 1: （a)为语义依存树；（b)为语义依存图

基金支持：本成果受国家自然科学基金项目（61872402）,教育部人文社科规划基金项目(17YJAZH068)，北京语言大学校级项目（中央高校基本科研业务费专项资金）（18ZDJ03）资助

2.2 古代汉语语义依存图标注规范

语义现象和规律具有继承性及相似性，古代汉语语义依存标注借鉴现代汉语语义依存的标注规范，以程荣辉等中文语义依存图标注体系[0]为参考，针对古代汉语的语法和语义特点，对其中的句子切分规则、语义关系、语义标签等进行调整，如下：

（1）分句采用字切分方式。古代汉语以单音词为主，在历史中逐渐演变为复音词，词汇在不同阶段有不同的形式。通过标注过程中的人工检验，我们发现语料历史时间跨度长，不论使用何种古汉语分词工具[1][2]，部分句子成分都难以界定语法单位为词或词组，如“兄弟”既可以作为一个词泛指同辈男子，也可以作为一个词组指哥哥和弟弟。为了提高标注效率，同时也便于进一步研究合成词的历时演变规律，后续预训练模型的输入，在切分古汉语句子时以单个汉字作为单位，但在依存标注时仍以词为单位做依存分析。词与词之间通过核心字进行连接，汉语词的核心语素通常是该词的最后一个汉字或前面的实义语素。如下图词语“天子”最后一个字“子”为该词的核心节点，“子”指向“天”，标合词“mHc”标签表示为一个词。命名实体“太尉”内部成分用“mHc-NR”连接，与其他词通过“尉”进行连接。

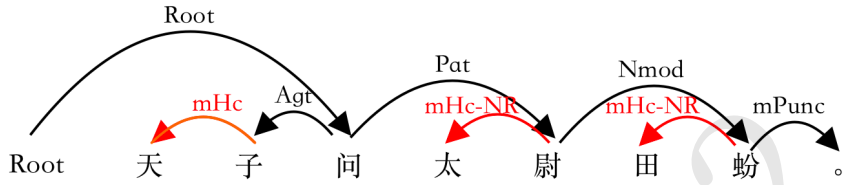


Figure 2: 合词和命名实体标注方式

（2）制定复音词标注规范。对标注过程中难以界定的复音词，本文制定了复音词标注规范。复音词由两个或两个以上音节构成，但在语法单位上都属于词。包括由一个语素构成的多音节单纯词和多个语素构成的合成词，复音词内部各个汉字通过合词标签mHc连接。判断复音词主要有意义和语法两个标准。单个或多个语素表示一个完整的意义单位时，这几个语素为一个分词单位，如单个语素的多音节词“参差”为一个词，表示“长短、高低不齐的样子”；两个语素“四海”结合构成了新义指代天下，两个相近的语素“亲戚”凝结成更具概括性的意义指内外亲属，重叠形式的复音词“世世”不是原义的简单重复，表示“后代子孙”。组合前后语法性质发生改变的复音词也划分为一个词单位，如“学”和“问”都为动词，结合后的“学问”为名词表示学问。

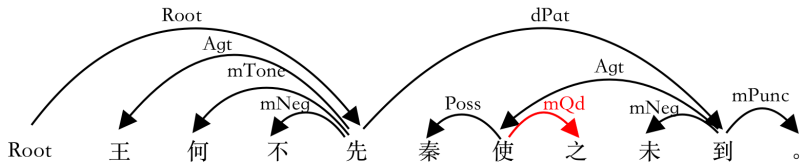
（3）制定命名实体标注规范。为了后期文本理解和知识图谱等应用，本文把命名实体划分为一个分词单位，并对难以界定的命名实体划分进行规范。本文的命名实体除普通人名、地名、组织名、机构名、时间、日期等，还包括特殊的人物名，如尊号加身份的称谓“孝武皇帝”；表示地形地貌的普通名词如“呼蚕水”“羌谷”；书名《蜀鉴》；带有类名的机构名“吏部”；历史朝代名称“唐朝”；古代刑法、学说等专用属于名“佐学”；历史上的重要事件、运动，如“渭水之盟”等。

（4）增加特殊的语义标签。除了增加以上提到的mHc和mHc-NR语义标签，我们还增加了特殊的语义标签“mQd”取独标签，用来表示取消句子独立性作用的“之”与其依存词之间的语义关系。“之”在古代汉语中可作代词、动词、助词和语气词，使用频率高，其中表示取消句子独立性的助词用法为古代汉语“之”的特殊用法。“取独用法”用在主谓结构的主语和谓语之间，使这一主谓结构不能单独成句而成为句子里的一个成分。在语义依存分析中，“之”使该主谓结构表示的事件成为降级嵌套事件。如3“王何不先秦使之未到”中，“秦使之未到”是一个主谓结构，整个主谓结构整体作根结点“先”的宾语，成为“先”的降级嵌套受事。

3 语义依存图库建设

本文构建的古代汉语语义依存图库建设经历语料采集、语料预处理、标注工具完善、标注

[0]https://csdp-doc.readthedocs.io/
[1]甲言https://github.com/jiaeyan/Jiayan
[2]HanLP https://github.com/hankcs/HanLP/



（大王为什么不先于秦国使者没到达之前。）

Figure 3: “取独标签”

规范完善、标注人员培训、语料标注、标注规范再完善等流程，初步建立了3000句规模大小的的古代汉语语义依存图库。

3.1 语料采集和预处理

古代汉语语义依存图库语料选取自《二十四史》，二十四史历史跨越度长，基本由达官或著名文人兼官员负责编修(赵志伟, 2018)，以二十四史作为语料，具有可取体量大、覆盖不同朝代、收录全、用语规范等特点。标注前进行以下工作：

（1）语料采集。语料通过汉程网[3]和丁佳鹏[4]等整理的古文—现代文平行语料库以篇章为单位自动获取后进行筛选，整体筛选方法如图4，在原文语料上使用doc2vec(Le and Mikolov, 2014)工具获得对应的嵌入表征，再通过kmeans聚类算法(MacQueen and others, 1967)，对语料进行聚类，这样可以较好地保证语料的平衡性。通过肘部法则，本文将语料篇章分为了20个类，再从中随机选取3000个句子，通过进一步的人工审核，获得了最终用于构建语料库的古代汉语语义依存语料。相应的，标注平台页面添加了“显示原文”“隐藏原文”“显示译文”“隐藏译文”四个按钮，用于显示/隐藏语料所对应的原文段落和显示/隐藏语料所对应的段落译文。

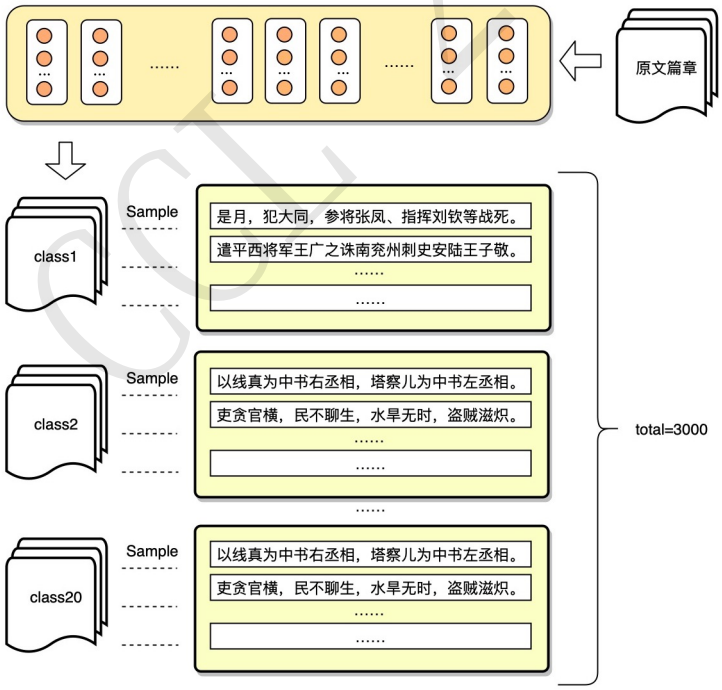


Figure 4: 语料筛选流程

（2）语料对齐。古代汉语相对于现代汉语语义理解的难度增加，为了帮助标注人员理解原

[3]http://www.httpcn.com
[4]https://github.com/NiuTrans/Classical-Modern

文语义，本文标注语料在采集时选取原文和对比翻译两部分，统一使用简体字。语料采集后对语料进行对齐工作。对齐时自动和人工校对相结合，以段对段即句子所在原文段落和所在翻译段落对齐的方式进行，部分对齐结果如图5。

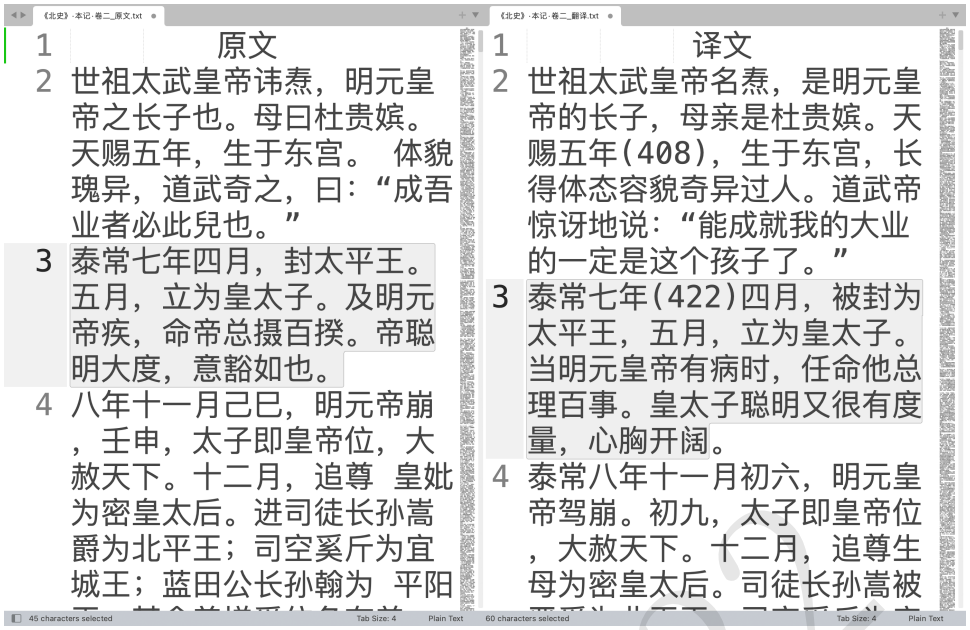


Figure 5: 部分对齐结果

(3)分句和词性标注。分句使用句末标点符号，即句号、问号、感叹号、省略号等进行分句，句长限制为15-30汉字。使用哈工大LTP语言云平台提供的词性标注模型对字（词）进行自动词性标注。通过以上的步骤，待标语料的处理已经基本完成。

3.2 语料标注

图库标注采取自动标注和人工标注相结合的方式，语料通过语义分析器自动进行语义标注后导入标注平台，标注人员在自动标注的基础上进行人工标注。共有5名语言学的硕士研究生参与标注工作，标注人员在经过培训后对语料进行标注，标注页面如图6。其中200条语料为5名标注人员共同标注，用于检查语料标注的一致性；另外2800条语料5名标注人员分别标注，历经五月余，五位标注员完成了3000句古汉语句子的语义依存的标注工作。

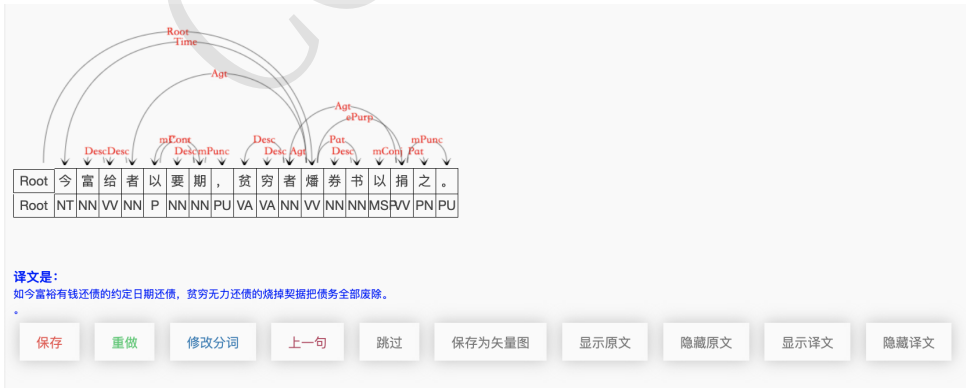


Figure 6: 标注页面

4 标注一致率

图库依存标注的一致率通过kappa值进行检验，公式如下：

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

其中P_0为总体分类精度，即标注一致的标签数量总和除以标注标签总数量；P_e为标注一致的标签种类的乘积与标签总数的平方的比例，经过计算kappa值为78.83%，整体一致率较高。句子标注不一致的原因主要有，（1）标注人员的古代汉语知识储备不能覆盖到所有知识，如对类似人名“曹旦”，单纯词“密迩”词义不了解；（2）古汉语中一些易混淆的语法结构，如兼语结构、连动结构和主谓短语作宾语的结构混淆导致标注语义标签也出现错误；（3）标注人员对标注规范熟悉程度不够，如细粒度结局标签（Cons）标注到粗粒度标签（CONS）上；（4）语义现象本身的模糊性，例如“九族”“亲疏”，后者做名词时，两者为并列关系，后者为形容词时，两者为修饰关系。针对以上问题，本文在后续标注过程中采取可改进措施，如加强标注人员的标注前培训，对准确率较低的标注人员取消标注资格，对易混淆的结构增订统一标注细则，对标注问题及时讨论，以进一步提高标注的准确性和统一性。

4.1 基于语义依存图库的现汉古汉对比分析

语义差异反映语言的特殊性，程荣辉[5]等已经建立了规模为2万句的现代汉语语义依存图库，本文通过现代汉语与古代汉语的对比，对古代汉语图库有个全面基本的认识，研究其特殊语言现象和规律。图库的语义标签反映词与词之间的语义关系，本文统计了古代汉语和现代汉语语义依存图库语义标签频次，图7结果直观地表明两者在语义标签分布上存在相关性和差异性。

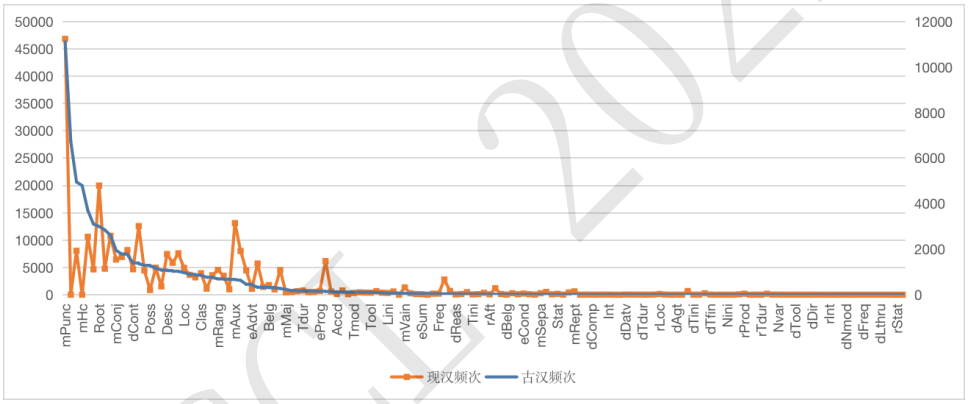


Figure 7: 现汉古汉语义标签频次分布

对标签数量进行皮尔逊卡方检验[6]，具体统计结果如表1。结果表明，卡方值和线性关联渐进显著性值都小于0.05，古代汉语和现代汉语语义标签分布具有统计学意义上的差异且有相关性。

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	11395.626a	9324	0.000
似然比	1073.584	9324	1.000
线性关联	87.925	1	0.000
有效个案数	149		
a 9520 个单元格(100.0%) 的期望计数小于5。最小期望计数为.01。			

Table 1: 现汉古汉语义标签数量卡方和相关性检验

[5]https://csdp-doc.readthedocs.io/
[6]统计检验使用SPSS Statistics 26.0.0.0得出

4.2 图库基本情况对比

对现代汉语和古代汉语语义依存图库基本情况进行统计，结果如表2。总体上看，古代汉语平均句长大于现代汉语，平均句语义标签数量比现代汉语更多，单句表示的语义内容比现代汉语更丰富。去除语法标签后将现古今汉语义标签频率分别由大到小排序，如图8。

语言类型	句数	字数	语义标签总数	平均句标签
现代汉语	20000	380839	275489	13.77
古代汉语	3000	72888	61736 （去除语法标签）	20.58

Table 2: 现汉古汉语义依存图库基本情况

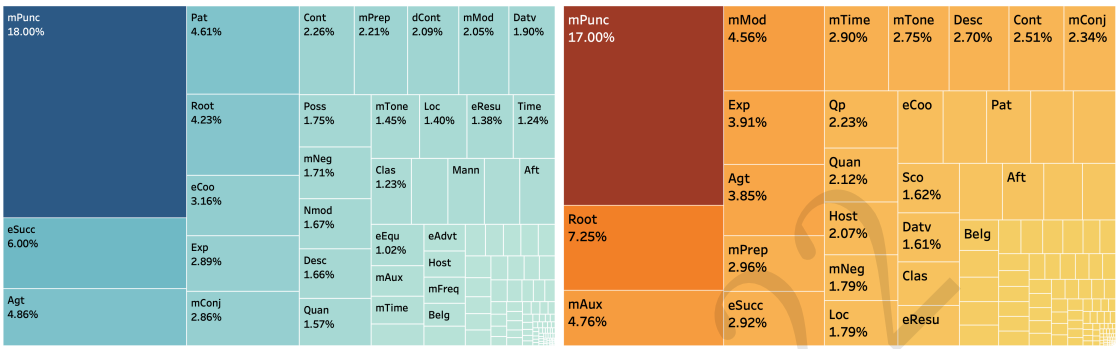


Figure 8: (a)古代汉语语义标签占比 (b)现代汉语语义标签占比

树状图显示，现汉和古汉语义标签总体上高频标签数量少占比大，低频标签数量多占比小，根据齐普夫定律(Wyllys, 1981)，语义标签出现的频次计作Pr，该语义标签的频次排位（即频级）计为r，在双对数坐标系下绘制出现代汉语和古代汉语语义标签频率分布曲线图9。可以看出古今汉语语义标签分布都符合齐普夫定律，说明现代汉语和古代汉语中在实际语言现象使用中都符合人类的省力和记忆原则，体现了语义现象的共性和规律。

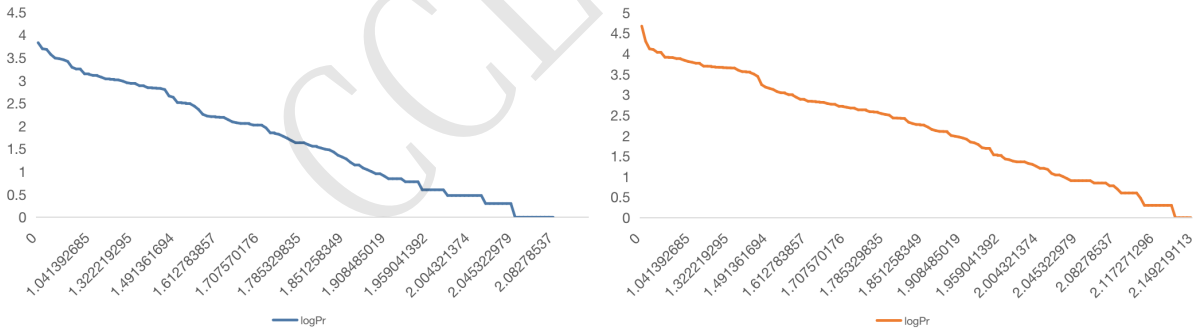


Figure 9: (a)古代汉语语义标签频次分布； (b)现代汉语语义标签频次分布

4.3 语义标签分布情况

受限于时间和语料规模，现汉图库和古汉图库规模大小不一致，以下统计分析采取将语义标签转化成相应所占比例方式进行对比。受限于篇幅，图中只展示重要靠前的语义标签，下面进行分类讨论：

4.3.1 周边角色标签

周边语义角色，即语义事件的参与者角色，包括主客体角色和情境角色，差异较大的标签如图10。语义事件由谓语支配，而主体角色作为动作行为的主要参与者，一般伴随谓语出现，

在语义事件中具有重要的地位。主体角色（施事Agt、当事Exp、感事Aft、领事Poss等）为动作的主体，客体角色（受事Pat、客事Cont、属事Belg等）为动作的第二参与者，多数为动作作用的对象。总体上，古今汉语主体角色的比例与客体角色不平衡，说明汉语中存在省略现象较多。

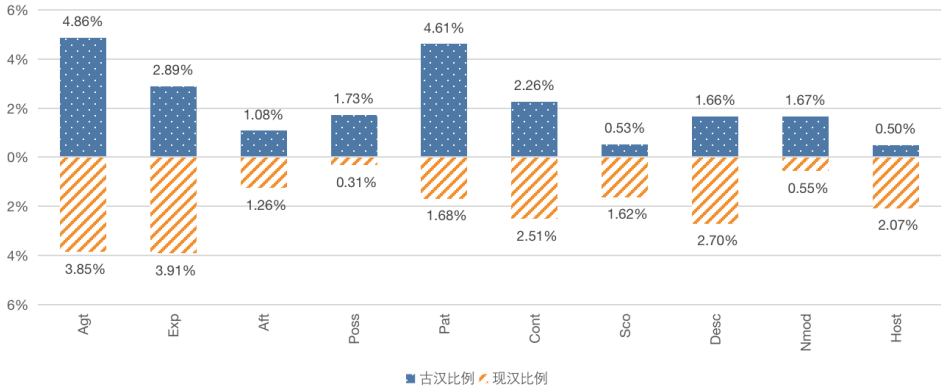


Figure 10: 现汉古汉周边角色标签占比

（1）施事（Agt）和受事（Pat）分别是自主性动作行为的发出者和承受者，当事（Exp）是非自主性动作行为的发出者，领事（Poss）表示领属关系的主体或整体部分关系的整体，客事（Cont）是事件所涉及但是并未改变的客体以及动作行为产生的新事物或结果，领事和当事常由非生命主体承担。统计显示，古代汉语中自主性动词多于非自主性动作行为，现汉中则反之，古汉领事占比明显高于古代汉语，都说明现汉中非生命主体作主语的句子更加丰富，对除人以外的事物描写更多。

（2）感事（Aft）表现心理活动的有意识的主体，二十四史属于历史传记，表示情感、心理活动的词少于现代汉语，叙事较客观冷静。

情境角色是事件涉及到的外围角色，主体使用的工具、材料，事件发生的时间、空间，引起事件发生的原因、目的等。

（1）范围（Sco）指的是事件中所关涉的方面、限定的界限、被审视的角度、发生作用的范围，如“诸贼”的“诸”，“其人”的“其”，现代汉语中描写事物时对议论的对象主体限制，事物的数量范围等限定较多，古代汉语集中于指称代词指称人物事物。现代汉语中描写事物时对数量范围、谈话对象限定较多,古代汉语限于指称代词指称人物事物。

（2）描写角色（Desc）表达的是一种特征，常作修饰成分，现代汉语中修饰成分更加复杂，多种性质的词和短语都可以做定语和状语，而古代汉语句子长度短，修饰成分数量远少于现代汉语，一般只用单一的修饰语修饰一个词，如“谩词”“寒隽之士”“杀父之仇”都用单个的述谓性质的词修饰一个名词。

（3）名称修饰语（Nmod）古代汉语占比例高主要是因为二十四史纪传体介绍人物年份等实体时描述详细，常与所属地、官职、年号等连用。宿主角色（Host）是属性的主体，或带有意义、功能、作用、价值的主体，通常出现的名词短语中，如“其奸恶”的“其”作为“奸恶”的主体，“宫庙大小”的“宫庙”，现代汉语中不管对人属性的描述还是对事物的功能作用描述都多于古代汉语。

4.3.2 依附标记标签

语义依附标记是对语义事件中依附性成分的标注，实际意义较虚，少单纯出现，但对句子语义有着一定作用，其中差异较大的标签如图11所示。

古代汉语词汇数量总体上少于现代汉语，语料涉及的事物范围更窄，在对事件情态、程度表达时也常常省略依附成分，现代汉语整体上限制语和修饰语更加丰富，表达更细致，表达主观情感的词数量更多。总体上，古代汉语依附标签占比低于现汉汉语，

（1）介词标记（mPrep）是对语法事件中介词和起到介词作用的词的标注。古代汉语此类标签主要用于表示时间、处所、方向、对象等的介词短语，表示目的句的“以”，所字结构的“所”字，及比较句中，如“于边境”“自数世以来”“请斩以谢天下”“所遇”“难于登天”等，而

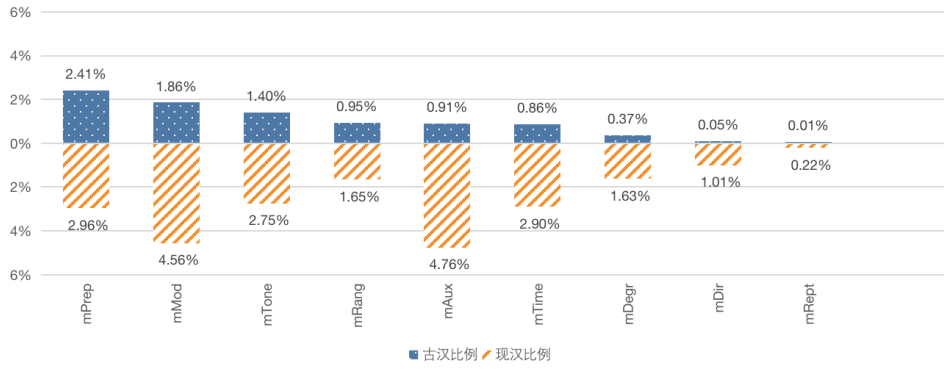


Figure 11: 现汉古汉依附标签占比

表示方式、依据、态度等则可以不使用介词，如“剑斩之”“法当斩”“兄事之”。除此，现代汉语“把”字句，“被”字句使用频率极高，介词一般不省略，分句当中起转折、并列等作用的词，如“还”“就”“才”等丰富，所以古代汉语介词使用比现代汉语更少。

（2）情态标记（mMod）是对句中表示情态的词进行的标注，这样的词一般表达的是主体的一种情形状态，比如惊讶、疑问、感叹或是能力、猜测等。古代汉语限于表示反问语气“岂”，必要性“须”“当”，可能性“可”，勇气“敢”，现汉比古汉更多表示推测、可能性、情感的词，如“只得”“总是”“真的”“就”等词汇，现代汉语情态词更丰富。

（3）语气标记（mTone）指对句中语气词的标注，二十四史语料的严肃性决定了语料的语气词较少，一般用于表示陈述、疑问、感叹、祈使等的少数虚词“也”“矣”“乎”“耳”“焉”等中，而现代汉语中多语气词“的”“了”“呢”“吧”“吗”“啊”“呀”等种类丰富得多。

（4）范围标记（mRang）是对句中表示范围的词进行的标注，可以是空间、时间或所指对象的范围，古代汉语一般出现在表示范围的具体位置，如“南羌中”“五品以下”，数量限制词“皆”“各”中。现代社会人们物质生活更加多丰富多彩，语句更加白话化，范围限制词使用频率更高，还会用在非生命主体作话话题限制谈话的内容当中。

（5）的字标记（mAux）是对汉语中出现的结构助词“的、地、得、之”进行的标注。古代汉语多为单音节词，修饰事物时修饰语可直接加被修饰语，如“锐兵”，少部分双音节词作修饰语时才用“之”进行连接，如“杀父之仇”，在近明清时期语言才更加白话化，使用“的”连接定语修饰语和被修饰语，如“他的官府”，而古代汉语补语数量更少，一般用“之”字，如“思之深”。现代汉语句子词汇语法形式发生变化，多音节词成主要优势，做定语、补语、状语等修饰成分常用结构助词连接。

（6）时间标记（mTime）是对句中一些时间副词以及动态助词的标注，如“正”“从此”等。程度标记（mDegr）是对句中表示程度的词进行的标注，如“广延百里”“大喜”。趋向标记（mDir）是对句中表示趋向的词进行的标注，如“蹈海去”“东征”等。二十四史记录的是已经发生的历史事件，作为官方历史材料叙述时冷静客观，较少描述时间动态的词语。重复标记(mRept)，如“唯唯”。这些标签都与语料范围有关，二十四史为正式语体，缺乏口语材料，在描述事件动作时对动作的状态，情态的程度上描述较少。口语化表达的“去”和语句重复也出现较少。

4.3.3 语义结构关系标签

语义结构关系的标注对象是语义事件，描述的是不同述谓概念之间形成的各种结构关系，差异较大的标签有后继标签（eSucc）、并列标签（eCoo）、目的标签（ePurp）、等同标签（eEqu）、目的标签（ePurp），如图12。

（1）后继事件（eSucc）指接着先行事件发生的事件，在时间上或逻辑上或空间上发生在后的后续性事件。古代汉语叙述事件表达简略紧凑，数词名词都能做谓语核心，一个句子中的几个分句往往有多个谓语，后继关系标签频词多。如“有司承旨奏戏，免为庶人”，“承旨”“奏戏”“免庶人”几个先后连续发生的事件共用一个主语，连接紧密，两个分句就包含三个动作，后继事件描述更丰富。

（2）并列（eCoo）表示的是前后两个事件或多个事件，其除了表示事件关系之外还能表

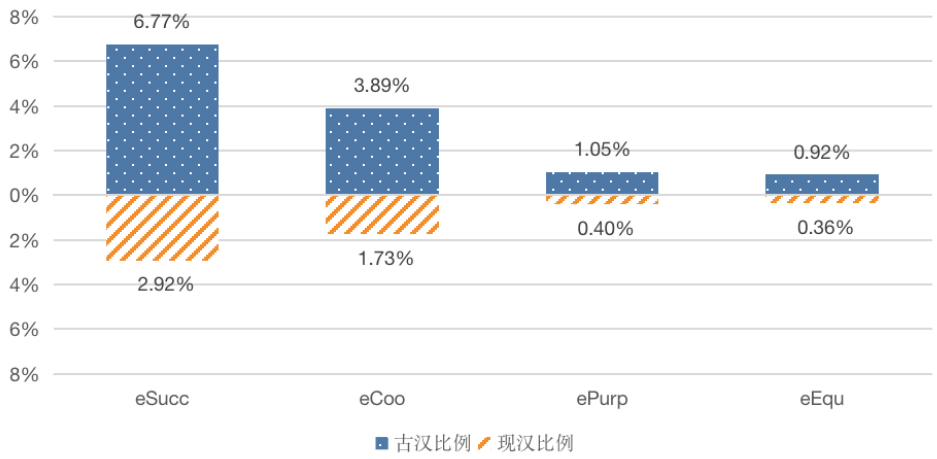


Figure 12: 现汉古汉语义结构关系标签占比

示平行的语义关系，在古代汉语中单音节词占多数，相近语义的词常并列使用，如“攻击”为“进攻击打”，“改易”为“改变更换”，“田庄”为田地庄园。分句结构中也常用相似结构对举并列，如“既而改元天历，郊庙，建后，立储”，“改”“郊”“建”“立”四个动并叙排列，使用并列结构多于现代汉语。

（3）目的标签（ePurp）是通过某些手段而要达到的目的性事件，统计发现，二十四史历史事件中使令句和目的句丰富，一定程度上反映了二十四史的官方政治历史题材和人物的等级关系，如“使人函封汉使者节赛上”，“使”与“函封“两事件为目的关系，反映上级对下级的命令。等同标签（eEqu）是对于同一个事物的复指或注释，和纪传体以人物为中心叙事有关，如“丞相方进“，“丞相”“方进”之间为等同关系。标签一定程度上说明了二十四史强烈的历史性叙事特征。

4.3.4 缺失语义标签对比

对现代汉语和古代汉语未出现的标签进行统计，去除语法标签合词标签（mHc）和命名实体标签（mHc-NR），结果表明，古代汉语特有的语义标签有4种，按频次高低排序分别为取独标签（mQd）、反描写（rDesc）、嵌套结局（dCons）、嵌套时距（dTrang）、反数量（rQuan）。mQd为古代汉语“之”做取消句子独立性作用时使用的标签，这是古代汉语的特殊语法现象。降级事件dCons、dTrang，出现古代汉语话语句中，二十四史多人物言论，同样的语义角色出现在话语中便成为降级事件。反关系rDesc和rQuan出现在“形容词”+“者”和“数量词”+“者”的结构中，如“贤者”“近者”“降者”“一辈大者”，古代汉语中用此结构代称一类人，而现代汉语中指称此类用助词结构“的”把修饰成分介绍给核心词。

现代汉语特有的语义标签有26种，按频次高低排序分别是先行关系（ePrec）、插入语（mPars）、反方式（rMann）、变化量（Nvar）、嵌套宿主（dHost）、嵌套终止状态（dSfin）、嵌套工具（dTool）、反结局（rCons）、嵌套源事（dOrig）、嵌套趋向(dDir)、反终处所(rLfin)、嵌套起始状态(dSini)、反意图(rInt)、反通过处所(rLthru)、嵌套成事(dProd)、嵌套名词修饰语(dNmod)、嵌套终处所（dLfin）、反比较（rComp）、嵌套频率（dFreq）、反数量短语（rQp）、嵌套数量短语（dQp）、嵌套通过处所（dLthru）、反源事（rOrig）、反趋向（rDir）、反状态（rStat）、反时间起点（rTini）。除了插入语（mPars）为现代汉语特有现象外，其他标签多为降级事件标签和反关系标签，其本身在现代汉语中出现数量少，说明这几类语义现象在语言环境中为不常见现象，而现代汉语语料规模更大，涵盖的更多的语言现象和语义关系。另外，随着时间发展，现代社会新事物不断增多，语言上词汇和句法形式更加复杂，语义现象种类也更加繁多。

4.4 结论及未来工作

本文在已有的标注规范及语料资源基础上，以《二十四史》作为古代汉语语义依存语料库的语料，建立古代汉语语义依存图库，并对语料库统计分析，发现古代汉语的语义特色和规

律。统计显示，二十四史语义关系总体上高频词数量少，低频词数量多，整体符合齐普夫定律，语义事件谓词连接紧密，叙述事件以人物为中心，常引用人物话语，地点、时间等周边角色描述细致，目的句和使动句较多，具有更强的政治色彩和书面语叙事特征。

该工作的不足之处是受限制于时间和人力成本，语料库规模较小，语料范围局限于历史题材和正式语体，不能完全反映古代汉语的语言规律，对古代汉语语言规律刻画描述还不够细致，标注受专业知识限制大，标注规范还不能囊括古代汉语所有的语言现象。在未来工作中，我们将进一步提高语料质量，扩大图库的语料规模和覆盖范围，对语义现象和原因作更加细致的分析。将来，依存图库还能应用于机器翻译，及古代汉语教学等更广阔的人文文化场景中。

参考文献

Yu Ding, Yanqiu Shao, Wanxiang Che, and Ting Liu. 2014. Dependency graph based chinese semantic parsing. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 58–69. Springer.

K. Hacioglu. 2004. Semantic role labeling using dependency trees. *Association for Computational Linguistics*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, number 14, pages 281–297. Oakland, CA, USA.

Lei Shu, Yiluan Guo, Huiping Wang, Xuetao Zhang, and Renfen Hu. 2021. 古汉语词义标注语料库的构建及应用研究(the construction and application of Ancient Chinese corpus with word sense annotation). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 549–563, Huhhot, China, August. Chinese Information Processing Society of China.

Ronald E Wyllys. 1981. Empirical and theoretical bases of zipf's law.

N. Xue. 2008. Labeling chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255.

Koichi Yasuoka. 2019. Universal dependencies treebank of the four books in classical chinese. In *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28. Digital Archives and Digital Humanities.

丁宇. 2014. 基于依存图的中文语义分析. Ph.D. thesis, 哈尔滨: 哈尔滨工业大学.

刘琦. 2012. 一种基于*WordNet*上下文的词义消歧算法. Ph.D. thesis, 吉林大学.

张丽丽. 2014. 先秦时期帝王、诸侯谥号词汇—语义系统研究. Ph.D. thesis, 山西师范大学.

王跃龙and 姬东鸿. 2007. 汉语依存图库建设研究. In 中国计算技术与语言问题研究——第七届中文信息处理国际会议论文集.

赵娟. 2005. 《战国策》副词研究. 山东师范大学.

赵志伟. 2018. 从"前四史" 到"二十四史". 中学语文教学, 9.

郑丽娟and 邵艳秋. 2015. 基于语义依存图库的兼语句句模研究. 中文信息学报, 29(6):8.

郑丽娟, 邵艳秋, and 杨尔弘. 2014. 中文非投射语义依存现象分析研究. 中文信息学报, 28(6):41–47.