Modern Linguistics 现代语言学, 2024, 12(7), 972-978
Published Online July 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.127646

Hans汉斯

基于《汉语大词典》义项统计探究上古汉语双音化现象

戴俊阳，张雨晴，包伊蕊，戴欣颖

南京师范大学文学院，江苏 南京

收稿日期：2024年5月29日；录用日期：2024年7月24日；发布日期：2024年7月31日

摘　要

本文立足于探索上古汉语词汇的双音化现象，借助基于词网(WordNet)、中文概念词典(CCD)构建的上古汉语词网，通过宏观统计分析了双音化的现象、机制与特点，意在说明汉语词汇的双音化不仅存在于中古之后，早在上古时期就有了充分的准备，打破了前人学者对于双音化现象大规模出现于中古时期的结论，通过宏观计量深度挖掘上古汉语中的文化价值。

关键词

古汉语词网，WordNet，上古汉语，双音化

To Explore the Phenomenon of Diphthongization in Upper Old Chinese Based on the Meaning Statistics of the *Hanyu Da Cidian*

Junyang Dai, Yuqing Zhang, Yirui Bao, Xinying Dai

School of Literature, Nanjing Normal University, Nanjing Jiangsu

Received: May 29th, 2024; accepted: Jul. 24th, 2024; published: Jul. 31st, 2024

Abstract

This paper is based on exploring the phenomenon of disyllabization of ancient Chinese words. With the help of the ancient Chinese Wordnet constructed based on WordNet and Chinese Concept Dictionary (CCD), this paper analyzes the phenomenon, mechanism and characteristics of

文章引用: 戴俊阳, 张雨晴, 包伊蕊, 戴欣颖. 基于《汉语大词典》义项统计探究上古汉语双音化现象[J]. 现代语言学, 2024, 12(7): 972-978. DOI: 10.12677/ml.2024.127646

disyllabization through macro statistics. It is intended to show that the disyllabization of Chinese words not only existed after the Middle Ages, but was fully prepared as early as the Ancient Times, breaking the conclusion of previous scholars that the disyabization phenomenon appeared on a large scale in the Middle Ages. It deeply explores the cultural value of ancient Chinese through macro-measurement.

Keywords

Ancient Chinese Word Network, WordNet, Ancient Chinese, Diphthongization

1. 引言

纵观汉语发展史，学界普遍认为，汉语词汇经历了漫长的复音化过程，从上古时期的单音词为主转变为现代的双音词为主[1]。然而，对双音化的研究主要集中在中古之后，而较少涉及上古阶段。同时，对上古词汇的研究也往往聚焦于少数语料和个别文献，视角不够全面。其实，双音化现象在上古已经大量存在，虽然成果不稳定，易消亡，却已经形成了成熟的双音化机制，为中古之后的汉语演变准备了充足条件。

本文将借助以词网(WordNet)、中文概念词典(CCD)为基础构建的上古汉语词网，通过对静态词型的宏观统计，探究上古汉语双音节化的情况。《汉语大词典》作为一部大型历时汉语语文辞典，大量收录古籍中的语词，尽可能给出词语每个义项最早产生的朝代和历代沿用情况，能够通过静态的词型数目反映汉语词汇的历时演变。而借助词网(WordNet)这一国际通用的词汇语义框架，可以打破“由形定位”的传统查询方式，直接深入探究上古汉语词汇系统的演变。由北京大学计算语言学研究所开发的《中文概念词典》(the Chinese Concept Dictionary，简称为 CCD)作为一个双语 WordNet，不仅提供了汉英双语的概念，还能够帮助我们完成上古汉语词表的义项映射工作，“由义索词”从而得到单音词的同义双音词集合[2]。

2. 词网简介

WordNet 是由普林斯顿大学认知科学实验室开发的词汇数据库，旨在构建一个以语义关系为基础的词汇网络，从而解决传统词典“由字母排序”、“由形排序”的局限性。WordNet 主要依据基本词性将义项分成四大类：名词、动词、形容词以及副词，并将其加以组织成同义词集合(Set of Synonyms, Synset)[3]。各个同义词集间又可通过增设指针互相联系，借此来表示各种语义关系，主要有同义关系、反义关系、上下位关系、整体部分关系、蕴涵(推演)关系。在目前的 WordNet 3.0 版本中，包含了 117,798 个名词同义词集合、11,529 个动词同义词集合、22,479 个形容词同义词集合和 4726 个副词同义词集合，从常见的词汇到专业领域的术语都有所涵盖[4]。

以 WordNet 的名词体系为例，根据概念间的上下位关系，以基本义项为根节点，所有名词分属于 25 个起始概念之下，形成彼此独立的概念树[5]。这些概念树可以从唯一根节点(顶层上位词 entity)出发，所有的名词概念都根据上位/下位关系被添加到这个根节点上；亦可以根据不同需求选择不同的“根义项”构建概念树。这样的概念树可以展示名词概念之间的层级关系，实现由词义反向检索词汇[6]。

而由北京大学语言科技研究所开发的 CCD 词表作为成熟的英汉双语 WordNet 语义资源，以 1997 年

发布的词网 1.6 版本为基础，整体可被称为词网的中文版本。CCD 的主要特点是用同义词集合表示一个概念，概念之间的上下位关系是 CCD 的关键内容，上下位关系通过增设指针从而联合其他的对立关系、部分整体关系构成整个概念网络。目前，CCD 可以应用于中英互译、信息提取以及概念检索等自然语言处理方面。故而，CCD 可以作为桥梁，帮助我们构建上古汉语词网，实现古汉语到英语的映射。

先秦时期以文献形式记录在册的词语众多，目前徐会丹(2019)的团队已完成先秦词网的搭建。由刘雪扬于 2015 年完成的对《汉语大词典》(第一版) 45 万左右义项最早出现年代信息的标注作为底本，查找初次出现于先秦时期的有记录的词语形成词库，共标注了 63,230 个先秦义项，删除先秦词库中的 5875 条熟语和反义复合词记录，构建起了涵盖 39,623 个词语、57,355 个义项、16,431 个义类的先秦古汉语 WordNet。在这之中，标记的名词义项占比 49%、动词义项占比 29%、形容词义项占比 16%、副词义项占比 6%，通过义项构筑语义树层级网络，展现了先秦汉语的完整面貌。

承袭先秦词表的标注理念，我们主要以《汉语大词典》中记载的最早出现年代为两汉的词为对象展开，依据其形成词库，得到共 54,865 个词，删除同形词后共有 43,544 个标记为 CCD 中对应的 id，由稳定的基本词汇和大量一般词汇构成。按 WordNet 的基本词类划分，其中名词占比 51%，动词占比 32%，形容词占比 15%，副词占比 2%。

我们利用中文 CCD 概念词典，将英语 WordNet 网络体系与两汉词网进行映射。利用先秦两汉构建成熟的词汇语义网络，我们可以进一步探究上古汉语内部，单音词双音化的宏观演变。

3. 上古汉语的双音化现象

前人学者有关汉语词汇双音化的研究，多从单篇书证中统计而来，往往得出上古汉语单音词多于双音词的结论[7]。譬如向熹(1980)计量出《诗经》中共出现词超 4000 个，其中包括专有名词在内，含有 1329 个双音词，占总词数比重低于 33%；针对《孟子》一书，赵克勤(1994)计量出总词数共计 2278 个，其中双音词 713 个，约占总词数的 31.3%；而万业馨(2021)指出《论语》一书中出现的总词数为 1504 个，其中包括 378 个双音词，约为总词数的 25.1% [8]。

然而，基于先秦语料库，我们统计了先秦时期首次出现并一直使用至当代的义项以及词汇数目(见表 1、表 2)，并计算出不同词长占比，得出了不同于前人描述的结果。

Table 1. The number and percentage of items that have been used since the pre-Qin period
表 1. 先秦沿用至今的义项数目及其占比

词长	先秦时期沿用至今的义项数目(个)	义项数目占比
1	929	37.13%
2	1330	53.16%
3	19	0.76%
4	200	7.99%
>4	24	0.96%
共计	2502	100.00%

Table 2. The number and percentage of words that have been used since the pre-Qin period
表 2. 先秦沿用至今的词汇数目及其占比

词长	先秦时期沿用至今的词汇数目(个)	词汇数目占比
1	762	32.93%
2	1310	56.61%

续表

3	19	0.82%
4	199	8.60%
>4	24	1.04%
共计	2314	100.00%

根据《汉语大词典》收录的词汇以及义项占比，我们发现先秦时期沿用至今的义项数目之中其实双音词占比更高，约为单音词的 1.7 倍，与学界普遍认为的“单音词占优”似乎并不相符[9] [10]。而从历时的角度来看，我们对比了两汉与先秦年均产出的义项数目(见表 3、表 4)，发现双音词的产出均占比最高，且随着时间推移，其增速也明显上升，两汉时期产出的双音节义项数目约为先秦时期单音节义项数目的两倍之多(见表 5)。

Table 3. The number and percentage of items that have been used since the two Han dynasty
表 3. 两汉沿用至今的义项数目及其占比

词长	两汉时期沿用至今的义项数目(个)	义项数目占比
1	324	17.30%
2	1357	72.45%
3	20	1.07%
4	157	8.38%
>4	15	0.80%
共计	1873	100.00%

Table 4. The number and percentage of words that have been used since the two Han dynasty
表 4. 两汉沿用至今的词汇数目及其占比

词长	两汉时期沿用至今的词汇数目(个)	词汇数目占比
1	83	6.03%
2	1103	80.16%
3	20	1.45%
4	155	11.26%
>4	15	1.09%
共计	1376	100.00%

Table 5. Average annual number of output items in the pre-Qin and Han dynasties
表 5. 先秦、两汉年均产出义项数目

词的构成单位个数	先秦义项数	先秦年均产义项数	两汉义项数	两汉年均产义项数
1	18,481	8.889	9368	21.195
2	36,625	17.616	33,035	74.740
3	251	0.121	968	2.190
4	1182	0.569	79	0.179
>4	37	0.018	2	0.005
总计	56,576	27.213	43,452	98.308

4. 双音化的内在机制

从先秦到两汉，之所以有这样日渐突出的双音化现象，与社会的高速发展分不开关系。随着人们生活的日趋丰富，单音词已经越来越难以涵盖与日俱增的语义。这种情况首先导致了大量同音同形词的出现，例如上古汉语中的“辟”，同时承担了现代汉语中“僻”“避”“辟”等多个字的含义。其次，另造新字以表新意也越来越不现实。例如“白马曰骍，黑马曰骊，马七尺以上曰骒，马高八尺曰騋，公马曰骘，母马曰騇”，非但使造字越来越困难，也使词汇繁冗，不符合语言以少驭多的经济原则[11]。

在此情形下，人们尝试用单音词的组合来扩充词汇，逐步形成了发明双音词的几条基本途径：第一，创造偏正结构，如上文由“马”延伸出“白马”“黑马”等。第二，组合近义词、同义词、反义词为联合结构[12]。反义词的组合往往成为偏义复词，如“出入”“生死”“休祲”“妖祥”等，近义词有组成偏义复词如“园圃”的，也有取二者共用义的，如“恐惧”“恭敬”[13]。值得一提的是，辞书或注疏中互训的两个单音词也经常合在一起，组成双音词。如“棘”字，在《诗经》毛传里的解释是“棘，枣也”。到了东汉的《淮南子》，“棘枣”就成为了双音词：“伐棘枣而为矜，周锥凿而为刃。”据统计，《论语》中偏正式复合词占总词数的37.2%，并列式复合词占总词数的26.7%，《周易》和《诗经》中的偏正式复合词分别占其总词数的67%和68.56%，可见创造偏正与联合结构，在上古已成为双音化的主要途径。

此外，利用双声叠韵创造连绵词，以及添加词缀，也是上古汉语词汇双音化的重要手段。连绵词多见于《诗经》和乐府，如“窈窕”一词，就有“窈窕淑女”“窈窕艳城郭”“窈窕曳罗裾”等，除此之外还有“参差”“匍匐”“绸缪”，等等。添加词缀的做法则常见于口语中，如“母”“兄”变为“阿母”“阿兄”。然而，这些双音词的产生多半是出于调和音节的需要，与词义的发展关系不大[14]。

5. 双音化成果的不稳定性

既然上古汉语已经出现了大规模的双音化现象，为什么在前人研究结论中，依旧是单音词占优呢？这是因为，前人研究选取文本基本为先秦经典之作，后世阅读、探究频次较高，也就是出现在此类研究文本中的词，本身就是“超高频经典使用词”。而本文依据的《汉语大词典》则由于其遵循“除了缺笔避讳字和已被考订为‘讹字’者不予收列外，其余无所谓规范”的原则，收录了大量低频词。由此可见，上古汉语虽然大量生产双音节词，但其成果并不稳定，往往使用过一段时间就消亡，单音节词则更具稳定性(见表 6)。

Table 6. Number and percentage of pre-Qin items that died out in the two Han dynasties
表 6. 两汉消亡的先秦义项数目及其占比

词长	两汉消亡的先秦义项数目(个)	两汉消亡的先秦义项数目占比
1	1625	38.24%
2	2459	57.87%
3	50	1.18%
4	107	2.52%
>4	8	0.19%
共计	4249	100.00%

究其原因，首先可能在于，上古时期生产的双音词多为与政治相关的专有名词，时间和空间上的局限性都很高，不适于推广沿用。我们通过统计映射至相同中文同义词集合(CSynset)下的义项出现数，统计了仅存在于先秦时期的这 19,844 个义项之中，义类最丰富的义项，前十名列举见表 7。

Table 7. The ten most frequent items in the pre-Qin lexicon
表 7. 先秦词表中收词最多的十条义项

中文同义词集合	示例义项及其释义	重复数
官僚 官吏 官员 父母官 大老爷们 政府官员	【師】：官名。太师的省称。{周}代辅佐国君的官员。	350
地名	【劉】：古地名。一在今{河南}{偃师}南。	269
国家 国度 政府	【矢】：{西周}晚期古国名。	214
祭祀	【傷】：丧祭。	205
人名 名号 名字 名称 大名 姓名 学名 本名 现名 称呼 芳名	【妊】：古人名。	135
江 河 大川 大江 江河 河川 河流 滨江 滨河	【囂】：水名。	122
山	【姖】：传说中的山名用字。	118
医学	【氣】：中医学术语。指脉气和营卫。	110
侯国 公国 封邑 领地	【囂】：古邑名。一作“廬”。在今{湖北}{襄阳}西南。本{春秋}{庐戎国}，被{楚}灭后为{楚}邑。	74
兽 禽 动物 飞禽走兽	【師】：兽名，猴属。	69
卦	【則】：《易》卦名。六十四卦之一。坎下坤上。	61

可以发现，从先秦到两汉消亡最多的词汇和官职有关。先秦时期所有与官职有关的义项共计 640 项，仅存于此时期的官职名赫然高达 54.7%，最后均未存于后世。除此之外，祭祀、侯国等义项的更替亦是如此。西周建国之初，各诸侯国的政治制度大体统一。直至春秋时期，随着周王衰微、诸侯自立，各国的制度文化开始分化，“田畴异亩，车涂异轨，律令异法，衣冠异制，言语异声，文字异形”，官名、地名、国名等也渐趋繁杂，而战国时期则被称为“古今一大变革之会”，不仅礼仪制度更迭繁多，文化思想方面亦可谓“百家争鸣”[15]。由此我们可见，先秦时期礼制相关义项的出现与消亡频繁其实间接反映了礼仪制度、思想文化的更迭[16]。

此外，先秦时期仅见文于《周礼》的官职名就高达 143 项。这是因为，《周礼》本名《周官》，据考成书于战国时期，其中精微繁复的官制其实是作者本人的政治理想蓝图，主要基于作者所处的春秋背景下的官职体系展开，并非周朝真实施行的官制，因此其中的官职名称也多系作者自创，在现实中并不存在[17]。以此类推，可见上古生产的双音词不但限于时地，还有相当的一部分出于生造，随意性、偶然性较强，消亡也就是必然结果。

还有一个重要原因，就是语言内部发展的不平衡性。这里又可分为两点讨论，一是语音的发展滞后于语义的发展。汉语词汇的双音化不仅出于语义扩充的需要，更是对语音简化的补偿。上古汉语的声母和韵母都种类繁多，且存在不少复辅音，导致单音词之间的区别度较高，发音也较为复杂。而随着语音一步步简化，浊音、复辅音和入声韵都逐渐消失了，导致大量同音词出现，一句话的音节也变得相当简短，因此需要通过双音词来区别词义，补足音节。然而，语音的变迁是一个缓慢渐进的过程，其速度在上古时期远不及语义的扩张，这就导致先秦两汉虽然生产了大量双音词，却没有适合其生存的语音环境，人们依旧习惯于用单音词表达意思。二是书面语的发展滞后于口语的发展。我们至今能见到的上古汉语语料，大多是书面语。很有可能当时的口语已经有了明显的双音化现象，但书面语仍然保留着之前的习惯，因此当我们翻阅经典文献时，才会感觉上古汉语似乎长期处于单音词为主的状态下。如果我们把视

线转向更接近口语的乐府诗，就会发现其中含有高频、大量、相对固定的双音节词，以《陌上桑》为例：

“东方千余骑，夫婿居上头。何用识夫婿？白马从骊驹，青丝系马尾，黄金络马头；腰中鹿卢剑，可值千万余。十五府小吏，二十朝大夫，三十侍中郎，四十专城居。为人洁白晰，鬑鬑颇有须。盈盈公府步，冉冉庭中趋。坐中数千人，皆言夫婿殊。”[18]

这段话模拟秦罗敷的语气与使君对话，相当口语化。其中的双音词几乎达到了每句必用的程度，而且频次上盖过了单音词。在五言一句的结构中，有两个双音词夹一个单音词如“白马从骊驹”，也有两个双音词后跟一个单音词如“盈盈公府步”。面对这一段文本，我们就很难说汉代的口语依然以单音词为主了[19]。

6. 结论

综上，汉语词汇的双音化不仅集中于中古以后，在上古阶段也大规模涌现。虽然上古汉语还不具备容纳双音词汇的成熟条件，导致双音化成果不稳定，但其过程已为后世确立了双音词生产的基本途径，为中古及其之后的汉语演化作了充分准备。

参考文献

[1] 王云路. 中古汉语词汇研究综述[J]. 古汉语研究, 2003(2): 70-76.

[2] 卢雪晖, 徐会丹, 李斌, 等. 先秦词网构建及梵汉对比研究[J]. 中文信息学报, 2023, 37(3): 36-45.

[3] Uschold, M. and Gruninger, M. (1996) Ontologies: Principles, Methods and Applications. *The Knowledge Engineering Review*, **11**, 93-136. https://doi.org/10.1017/s0269888900007797

[4] Li, Y., Bandar, Z. and Mclean, D. (2003) An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, **15**, 871-882. https://doi.org/10.1109/tkde.2003.1209005

[5] Gao, J., Zhang, B. and Chen, X. (2015) A WordNet-Based Semantic Similarity Measurement Combining Edge-Counting and Information Content Theory. *Engineering Applications of Artificial Intelligence*, **39**, 80-88. https://doi.org/10.1016/j.engappai.2014.11.009

[6] 王安节. 单音词双音节化的考查[J]. 松辽学刊(社会科学版), 1992(2): 43-47.

[7] 王浩然. 古汉语单音同义词双音化问题初探[J]. 河南大学学报(社会科学版), 1994(3): 52-55.

[8] 李振东, 张丽梅, 韩建. 古汉语双音复合词理论研究的历史与现状述评[J]. 佳木斯大学社会科学学报, 2007, 25(1): 74-76.

[9] 邱冰. 中古汉语词汇双音化研究[J]. 燕山大学学报(哲学社会科学版), 2010, 11(1): 30-33.

[10] 刘欣. 词汇语义关系研究综述[J]. 智库时代, 2019(18): 296-298.

[11] 徐时仪. 汉语词汇双音化的内在原因考探[J]. 语言教学与研究, 2005(2): 68-76.

[12] 贝罗贝, 吴福祥. 上古汉语疑问代词的发展与演变[J]. 中国语文, 2000(4): 311-326, 381-382.

[13] 陈超. 先秦汉语第一人称代词研究综述[D]: [硕士学位论文]. 长春: 东北师范大学, 2014.

[14] 王勤. 俗语的性质和范围——俗语论之一[J]. 湘潭大学学报(社会科学版), 1990(4): 107-111.

[15] 赵晓斌. 春秋官制研究[D]: [博士学位论文]. 杭州: 浙江大学, 2009.

[16] 王进锋. 先秦时期君臣观念的形成与发展[J]. 西部学刊, 2016(18): 43-49.

[17] 张雁勇. 《周礼》天子宗庙祭祀研究[D]: [博士学位论文]. 长春: 吉林大学, 2016.

[18] [宋]郭茂倩, 编. 乐府诗集[M]. 聂世美, 仓阳卿, 校点. 上海: 上海古籍出版社, 2016.

[19] 张华. 汉代文学中的神话研究[D]: [博士学位论文]. 西安: 陕西师范大学, 2013.