基于《汉语大词典》的汉语词汇历时演变计量研究

李　斌　刘雪扬*

［摘　要］ 汉语词汇演变是汉语史的重要研究课题，然而由于带标注历时语料库的缺乏，词汇史的研究多为定性研究，宏观的、整体的定量研究还很难实现。本文运用数据库技术和计量方法，在人工标注历史性语文辞典《汉语大词典》的30多万个词条的80多万条书证的时代信息后，对词典中的词汇、义项数量和词长在历代的分布进行了统计学描绘，分析词汇的宏观演变，使用回归分析方法获得了当代词汇的词汇留存度和时代的对数曲线方程，为汉语史研究提供了重要的基础资源和公式。

［关键词］ 汉语大词典；词汇演变；汉语史；语言年代学

一、引言

汉语词汇的动态发展规律是汉语词汇史研究的重要命题。从历时的角度来观察汉语词汇是如何产生、消亡和意义变化，追溯当代词汇来源显得尤为重要。然而受限于大规模历时词典的缺失，学界的研究多限于定性研究或专书的定量分析，缺乏从先秦到当代汉语的大词汇量的整体定量研究。本文通过构建大规模汉语历时词库，尝试用语言计量方法分析汉语的历时演变情况，特别是当代汉语词汇的历时累积过程，拟合词汇历时演变的公式和曲线。

汉语词汇史研究历来强调利用真实的历史语料来细致描写词汇事实。潘允中[①]、王力[②]、向熹[③]、史存直[④]等都曾著书讨论汉语词汇两千多年的发展面貌和规律，论据来源于研究者对历史语料的长期积累和钻研，但大多是定性的考察，容易产生观点的差异，如汉语史分期问题，历来众说纷纭[⑤]。断代研究和专书研究也是汉语史研究的重要组成部分。前者描述了汉语词汇在某汉语史分期中的面貌，如

*李斌，计算语言学博士，南京师范大学文学院副教授，210097；刘雪扬，南京卫生高等职业技术学校讲师，210038。本文是教育部社科青年项目（16YJC740034）、江苏高校优势学科建设工程、江苏高校哲社优秀创新团队建设项目的资助成果。

①潘允中：《汉语词汇史概要》，上海：上海古籍出版社，1989年。
②王力：《汉语词汇史》，北京：商务印书馆，1993年。
③向熹：《简明汉语史（上）》，北京：高等教育出版社，1998年。
④史存直：《汉语史纲要》，北京：中华书局，2008年。
⑤董志翘：《汉语史的分期与20世纪前的中古汉语词汇研究》，《合肥师范学院学报》2011年第29卷第1期。

古代汉语词汇[①②③]、中古近代汉语词汇[④⑤]等，建立在定性分析的基础上；后者探讨汉语史各时代中有代表性的文献中的词汇，如《吕氏春秋》[⑥]、《国语》[⑦]等，大多运用计量方法进行穷尽性考察，但缺乏词汇的动态演变研究。词汇的历时面貌尚难以量化，当代汉语词汇的历代来源更无从获得。学界或是提出了研究设想[⑧]，或是探讨了某个或某类当代词汇的历代发展[⑨]，却尚未获得对当代词汇来源的整体的、定量的分析。

近年来，计量的方法被广泛应用，试图在探索语言的数学特性中探讨语言本身的内在规律[⑩]。但是，若没有大规模的、规范的电子语料库，计量方法则难以实施。随着汉语史研究的不断深入，对于汉语史电子语料库的需求更加迫切。汉语史研究不再是随机举例，而是要运用统计方法、比较方法，这样才能达到研究结论的可靠性[⑪]。尉迟治平[⑫]、王建军[⑬]等也都倡导利用计算机技术辅助汉语史研究。语料库的建立便利了计量方法的使用，而计量方法一方面能够开拓词汇研究的新领域，另一方面能够重新考察、审视、印证现有的词汇理论与观点[⑭]。由于标注义项的大规模历时语料库的缺失，收词数量丰富的历时语文词典《汉语大词典》则成为一些学者研究的语料来源。闫从发[⑮]、李娜[⑯]等分别研究了明代、民国等断代词汇，但仍以个案分析为主。因此，本文基于《汉语大词典》，建立汉语史词汇演变研究的基础资源——汉语词汇历时数据库，运用计量的方法获得对汉语词汇历时分布的定量描述分析，并拟测汉语词汇在历代发展的统计学规律，探讨当代汉语词汇的来源问题。

二、汉语词汇历时数据库的构建

从词汇演化的研究角度来说，历代自然口语和真实文本最能真实而直观地反映词汇的使用情况。古代没有录音技术，口语语音材料自无可寻，而历代真实文本虽有大量的电子化文本库，但只能进行字符串的全文检索，如中华经典古籍库[⑰]等。这些文本没有经过词语切分和义项标注，无法进行词汇级别的演化研究。由于没有大规模、高质量的、深加工的、反映词汇演变的语料库，我们只能暂且退而求其次，利用现有的大规模历时语文词典来统计不同时代的词汇的使用信息。

《汉语大词典》是一部大型历时汉语语文辞典，按照“古今兼收，源流并重”[⑱]的原则，收录了30多万个词语的详细释义和80多万条书证材料。词典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，在很大程度上反映了词语的历时演变。表1给出了词典对“玉環”第1个义项的书证。

---

①徐朝华：《上古汉语词汇史》，北京：商务印书馆，2003年。
②赵克勤：《古代汉语词汇学》，北京：商务印书馆，1994年。
③蒋绍愚：《古汉语词汇纲要》，北京：商务印书馆，2005年。
④蒋冀骋：《近代汉语词汇研究》，湖南：湖南教育出版社，1991年。
⑤方一新：《中古近代汉语词汇学》，北京：商务印书馆，2010年。
⑥张双棣：《吕氏春秋词汇研究（修订本）》，北京：商务印书馆，2008年。
⑦陈长书：《〈国语〉词汇研究》，北京：中国社会科学出版社，2014年。
⑧张能甫：《关于现代汉语词汇历史层次研究的一些思考——以现代汉语词语中的W字头的词或词组为例》，《西南科技大学学报》（哲学社会科学版）2012年第29卷第6期。
⑨蒋绍愚：《汉语词义和词汇系统的历史演变初探——以“投”为例》，《北京大学学报》（哲社版）2006年第43卷第4期。
⑩冯志伟：《用计量方法研究语言》，《外语教学与研究：外国语文》2012年第44卷第2期。
⑪董志翘：《加强汉语史语料库建设，促进汉语史研究的现代化》，《当代语言科学创新与发展国际学术研讨会暨〈语言科学〉创刊十周年庆典》，江苏师范大学2012年。
⑫尉迟治平：《计算机技术和汉语史研究》，《古汉语研究》2000年第3期。
⑬王建军：《语言科技新思维与汉语史研究的现代化》，《南京师范大学文学院学报》2011年第3卷第1期。
⑭苏新春：《计量方法在词汇研究中的作用及频级统计法》，《长江学术》2007年第2期。
⑮闫从发：《基于〈汉语大词典〉语料库的明代汉语词汇研究》，山东大学博士论文2009年。
⑯李娜：《基于〈汉语大词典〉的民国词汇研究》，山东大学博士论文2011年。
⑰中华经典古籍库：http://old.gujilianhe.com/。
⑱汉语大词典编纂处：《汉语大词典缩印本（上）·前言》，上海：上海辞书出版社，2007年。

表1 “玉環”的释义和书证

义　项	书　　证
玉制的环	《韩非子·说林下》：“吾好珮，此人遺我玉環。”此指佩环。唐张籍《蛮中》诗：“玉環穿耳誰家女，自抱琵琶迎海神。”此指耳环。金元好问《以玉连环为吕仲贤寿》诗：“玉環何意兩相連，環取無窮玉取堅。”此指玉连环。

《汉语大词典》的释义和书证信息，可以用来构建汉语历时词汇数据库。不过，《汉语大词典》[①]虽有书证材料，但一般只列出书名，在很多时候并没有给出书证材料的具体时代信息，例如“玉環”的书证中，《韩非子·说林下》只有一个书名，需要人工补充其具体时代。为了获取汉语词汇[②]使用的完整时代信息，我们利用传统的文献考证法，辅以读秀知识库[③]、百度百科、维基百科等其他网络资源，历经3年时间，人工考证出《汉语大词典》中书证材料中缺失的朝代信息。由于精确的年代信息很难全面获取，往往只有大致的朝代。因此，我们根据中国历史的分期，将书证的时代分为“先秦”（~ –221）、“秦”（–221 ~ –206）、“汉”（–206 ~ 220）、“三国”（220 ~ 280）、“晋”（265 ~ 420）、“南北朝”（420 ~ 589）、“隋”（581 ~ 618）、“唐”（618 ~ 907）、“五代十国”（907 ~ 979）、“宋”（960 ~ 1279）、“元”（1206 ~ 1368）、“明”（1368 ~ 1644）、“清”（1644 ~ 1911）、“民国”（1912 ~ 1949）和“当代”（1949 ~ 1986）共十五个时代。

经校对整理，我们构建了汉语的词汇历时数据库，包含了《汉语大词典》中的词汇信息，如词形、义项、书证等。该库共收词条336164个，义项493772个，其中39972个义项缺失书证材料。去除没有书证的义项后，数据库包含具有朝代信息的词条315747个、义项451493个、书证881182条。

三、词汇及义项的历代分布

基于《汉语大词典》的词汇历时数据库，我们可以统计出汉语词汇在历代演变的基本情况，例如每个时代有多少词语和义项，每个时代新产生和消亡的词语有多少，词语长度的历时分布如何等等。

（一）历代分布的基本数据

首先来观察全部词条和义项在十五个时代上的分布情况。由图1可以看出，先秦、汉、唐、宋、明、清六个时代的词汇量大，很有代表性。而秦、三国、隋、五代十国由于历史时间短，《汉语大词典》所引用书证少，其词汇量较少并不能说明真实情况。同样地，民国和当代的词汇量也显得偏少。如果仅根据这个图，很难看出这部词典对于汉语史研究的价值，反而容易招致收词不平衡、不全面的批评。

同时，受限于词典性质和词典编纂技术，《汉语大词典》没有给出一个词语和义项在所有朝代的文献中的出现情况，只给出某个义项的最早用例和少量后代用例，并未交代各词汇（义项）例证的朝代是孤立的还是连续的，所以需要对朝代的连续性问题进行两种不同假设，再进行计量研究。

假设1（孤立假设）：凡是《汉语大词典》列举了书证的，就认为该词（义项）在该书证的朝代中曾使用过；凡是《汉语大词典》未列举书证的，就认为该词（义项）在该书证的朝代中未曾使用过。

①本文使用的《汉语大词典》版本为：汉语大词典2.0（CD），香港：商务印书馆，2005年。
②本文不区分《汉语大词典》中的同形词，字形一样的词均作为一个词型（词条），词汇数量也就是词型（词条）数。
③读秀知识库：http://www.duxiu.com/。

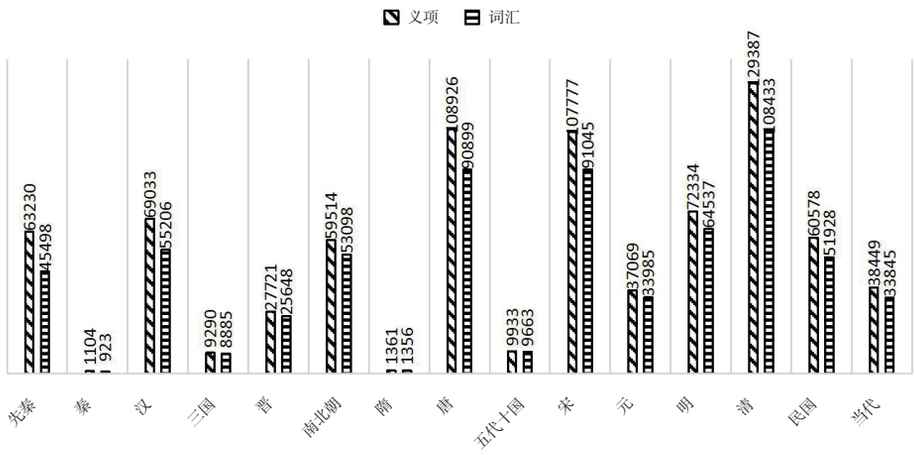

图1　词汇和义项的历代分布（孤立假设）

假设2（连续假设）：凡是《汉语大词典》列举了唯一时代书证的，就认为该词汇（义项）是该时代的特有词汇（义项）；凡是《汉语大词典》列举了多个朝代书证的，则认为该词汇（义项）在最早和最晚朝代之间一直沿用。

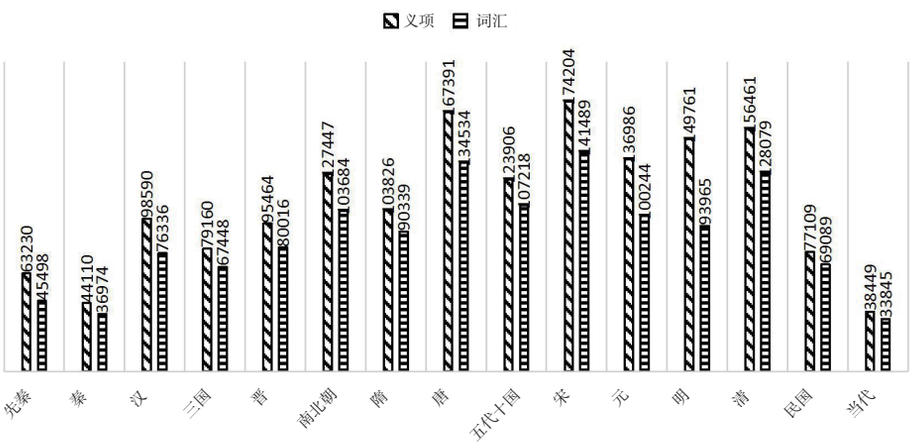

图2　词汇和义项的历代分布（连续假设）

从图2可以看出，词汇和义项从先秦到宋代基本上处于递增状态，元代到当代则处于下滑状态。考虑到宋元之后的传世古籍数量比前代丰富，词汇量也应该处于不断累积提升的趋势。所以，从这个结果可以看出，《汉语大词典》比较注重早期传世文献的书证材料，民国和当代的书证材料相对偏少。

（二）历代词汇的新生与消亡

虽然《汉语大词典》的收词在宋元之后略有下降，但出于其尽可能收录词语的最早用例，所以仍然能够窥见历代新增的词汇情况。图3给出了15个时代新出现和消亡的词语和义项数量。可以清晰地看到，先秦、汉、南北朝、唐、宋、清是产生新词最多的时代，这和图1的情况基本相似。

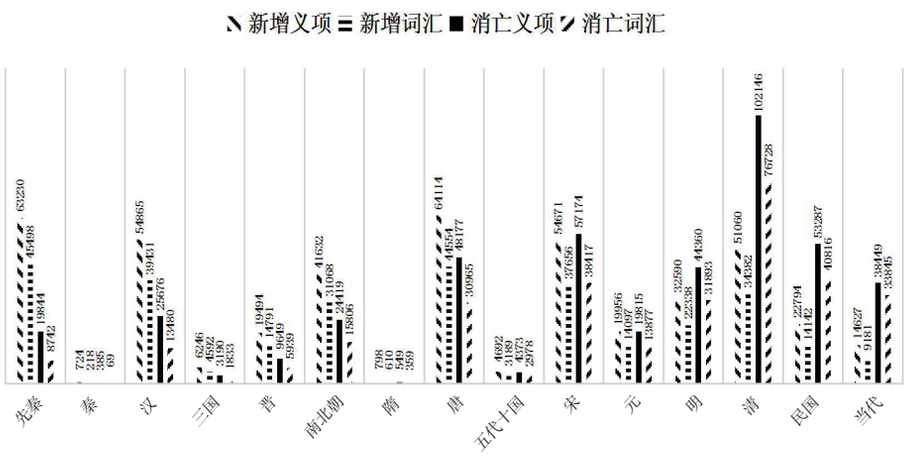

图3　词汇和义项的历代新生与消亡（连续假设）

汉语史研究一般认为：汉代西域词汇的输入、南北朝唐时期佛教词汇的输入、戊戌政变前后（清）西洋词汇的输入、解放后（当代）社会主义社会词汇的输入都曾产生大量新词，促进汉语词汇的发展[①]。图3中不同时代间词汇量的差距，表明清代新增词汇量明显大于唐代和汉代，该图也说明当代新增词汇量远甚于民国，而民国时期词汇也有增长。

词汇和义项的消亡也是词汇发展变化的常态。每个时代在新增词语和义项的同时，也会有一部分消亡的词汇。我们尝试利用历时词汇数据库来观察词汇的消亡问题。依据《汉语大词典》的编写原则，可以把词典中每一个词语和义项的最晚书证所对应的时代认作该词语和义项消亡的时代，从而进行词汇消亡的历时分析。

图3反映了义项和词汇历代消亡数量的时代分布。可见，在清代以前产生和沿用的词有很大一部分在清代之后就不再使用了，所以清代（清代最后使用，到民国便消失）是词汇消亡最剧烈的时代。除此之外，唐代、宋代、明代、民国和当代消亡的词汇也较多。由此，我们可以得出不同时代词汇和义项的词汇演化面貌。尽管汉代、唐代、宋代词汇新增量远多于消亡量，但是在发生大量词语新增的时代之后，随即发生着大量词语的消亡，这与古代盛世文化发展、乱世文化衰退的趋势基本一致。

（三）词长分布

在汉语词汇史的研究中，词长的变化一直是学术界关心的问题。《汉语大词典》的数据正可以用来计量分析。需要注意的是，词典收录了不少“语”，比严格定义的“词”要宽松许多。表2按照孤立假设给出了历代的词长分布。由于五字以上的词语很少，所以将词长分为一、二、三、四字和四字以上五类。

表2　时代所属词汇词长分布

词长 / 年代	一字	二字	三字	四字	>四字	平均词长	单字词比例
先秦	6209	35847	672	2629	141	2.01	13.65%
秦	725	183	5	10	0	1.24	78.55%
汉	6359	44709	1281	2719	138	2.02	11.52%
三国	1230	7084	161	385	25	1.98	13.84%

①王力：《汉语史稿》，北京：中华书局，1980年，第516—537页。

续表

晋	2407	21533	682	982	44	2.02	9.38%
南北朝	3971	45591	1350	2109	77	2.04	7.48%
隋	92	1155	53	54	2	2.06	6.78%
唐	5532	76778	4336	4071	182	2.09	6.09%
五代十国	836	7835	451	516	25	2.08	8.65%
宋	5793	73943	5440	5559	310	2.13	6.36%
元	2223	25162	3181	3062	357	2.26	6.54%
明	4003	50612	4006	5538	378	2.20	6.20%
清	6287	84653	6734	10155	604	2.22	5.80%
民国	3299	39007	3859	5361	401	2.25	6.35%
当代	2258	23296	3606	4261	424	2.35	6.67%

先秦时期单字词占有比例为13.65%，这表明先秦时期单字词数量并不占优势，这与单字词在古汉语中占优势的一般认识相悖。一方面这是《汉语大词典》收录的接近短语的“语”较多；另一方面，也可以看出单字词并不像人们想象的那么多。Li通过统计先秦25本经过人工分词的文献，得出了二字词在词条数量上超过单字词，但是在使用频率上远不及单字词[①]。

纵观各时代单字词占有率，从先秦开始，该指标基本上是逐步减小的，这从词型角度证明了汉语词汇由单字词向多字词发展的趋势。而时代的平均词长却表现出了变化的不规律性，尤其是先秦时代词汇的平均词长竟略高于2。相对而言，先秦词汇中四字及以上词汇数量较多，而汉语中的四字及以上词汇大多是成语或熟语。对于这类词语，《汉语大词典》通常会标示出这种反映成语雏形的书证，并使用“语出”标记，如成语“一日之長”（语出《論語·先進》：“子路、曾晳、冉有、公西華侍坐，子曰：‘以吾一日長乎爾，毋吾以也。’”）。

四、当代词汇的历代分布情况

基于这个数据库，也可以观察当代汉语词汇的历时演变情况。在数据库中，当代词汇共有33845个词条、38449个义项、43863条书证。其中，只具有当代书证的词语共9181个，如“魅力”“泥石流”“做假”等，属于当代特有词。其余的24664个词则是历代发展至今的。本文以这3万多个当代词条为对象，追寻当代词汇的源头，梳理汉语当代词汇的发展脉络，验证并补充汉语史的研究成果。

（一）当代词汇和义项的历代沿用

当代词汇是在先秦到当代的某个时代产生并沿用至今的，当代词汇（义项）在历代沿用的各时代总量反映了每个时代对当代词汇（义项）的贡献。我们在连续假设下，统计了数据库中所有当代词汇（义项）在各时代使用的总量。

---

①B. Li, M. Feng and X. Chen. “Corpus Based Lexical Statistics of Pre-Qin Chinese”, *Lecture Notes in Computer Science*, vol. 7717, 2013, pp.145—153.

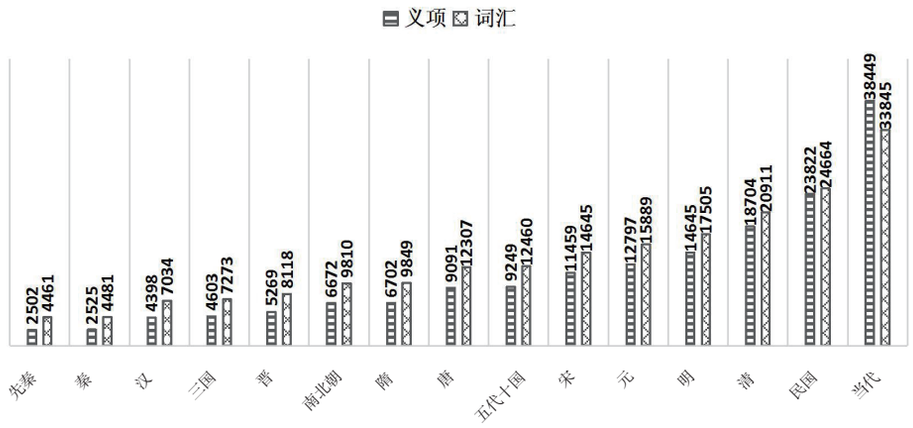

图4　连续假设下当代义项和词汇的历代沿用总量

图4给出了连续假设下，词汇（义项）时代的累积量，能够清楚地看出当代汉语词汇不断地增长过程，每个时代都有一定的贡献率，当代新增的数量最多。当然，词汇（义项）并不是简单地一直沿用，也会经历消亡到重新使用的过程，这一假设可能夸大了时代对词汇的贡献。所以，连续假设虽不能完全真实反映词汇演变的真实面貌，但可以对当代词汇（义项）的历代沿用面貌有概貌上的大致把握。

（二）当代词汇和义项的历代新增情况

上文揭示了词汇（义项）在各时代使用的概貌，为了进一步厘清每个时代新增的词语和义项，我们把每个当代词汇（义项）列出的书证的最早时代作为当代词汇（义项）的产生时代，并对当代词汇（义项）历代增长总量进行分析，反映在图5中。能够看出，先秦、汉、唐、宋、清、民国的贡献很大。

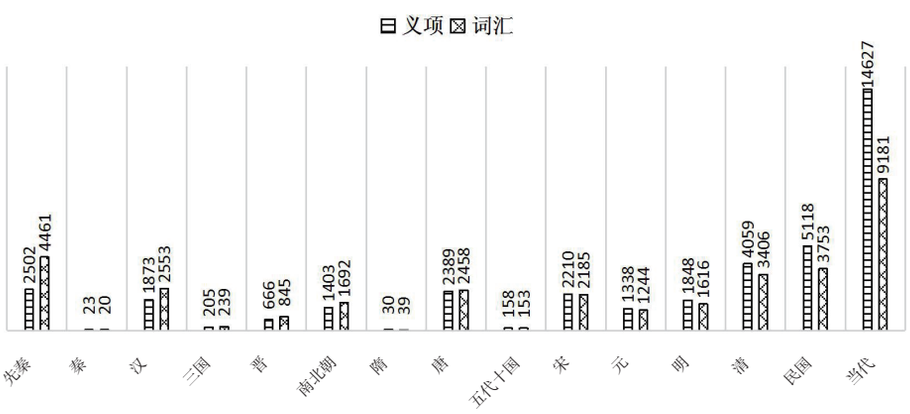

图5　当代义项和词汇历代增量的分布

（三）当代词汇历时留存度的回归分析

对《汉语大词典》词汇资源的计量研究不仅能够揭示词汇在不同时代的发展面貌，验证和补充汉语史研究的成果，也可以通过词汇面貌的统计来推测汉语词汇的时代，即建立汉语词汇数量与词汇时代的关系，挖掘出词汇演变的规律。Liu拟测了当代汉语的历时曲线[①]，但存在一些问题，论述不足。本

①X. Liu, B. Li and Y. Zhang, et al., "Quantitative Research on the Origins of Contemporary Chinese Vocabulary Based on The Great Chinese Dictionary", *Lecture Notes in Computer Science*, vol.8922, 2014, pp.112—123.

文则进一步展开研究。20世纪50年代，美国语言学家斯瓦迪士（Swadesh）通过语言的词汇统计，测定了古代语言与其发展而成的现代语言间的时间间距，这种方法也称为语言年代学（glottochronology）。斯瓦迪士以各语言共有的200个核心词汇作为研究对象，统计了不同语言与1000年前该语言的相似程度，即保留词汇的比例，发现该比例维持在0.72至0.85之间，由此得出了词汇年代的公式[①]：

$$t = -\frac{\ln(c)}{2\ln(r)} \qquad (1)$$

其中，r为常数，以英语为例，r等于0.85时，c表示现代英语与1000年前英语的保留词汇比例为0.72，t代表时代语言与现代语言的时间距离1000年（t=1）。这个公式是语言年代学中非常重要的贡献。根据这个公式，可以进行3种计算。（1）通过一组已知数据t和c，就可以进行r的估算。（2）得到常数r之后，可以得出语言年代t和词汇留存比例c的关系，即由c求t，根据留存比例c，可以大致明确这类古代语言究竟有多古老。（3）由t求c，根据这种语言的大致年代，推算词汇留存比例。除此之外，这个公式还可以根据词汇留存比例，计算两种有亲属关系的语言在历史上的分化时间等。

200个核心词建立的回归分析公式已经有了诸多用途，那么数万词条的历时演化数据应该可以得出一条更有价值的曲线或公式。我们用上文假设2中各时代词汇总量来计算时代语言在现代语言保存下来的词汇比例，前代语言与现代语言的时间间距，则取朝代的中间值与当代最大值（《汉语大词典》的出版时间1994年）差值的绝对值，根据15个时代的数据形成15个数据点，然后进行数据拟合。

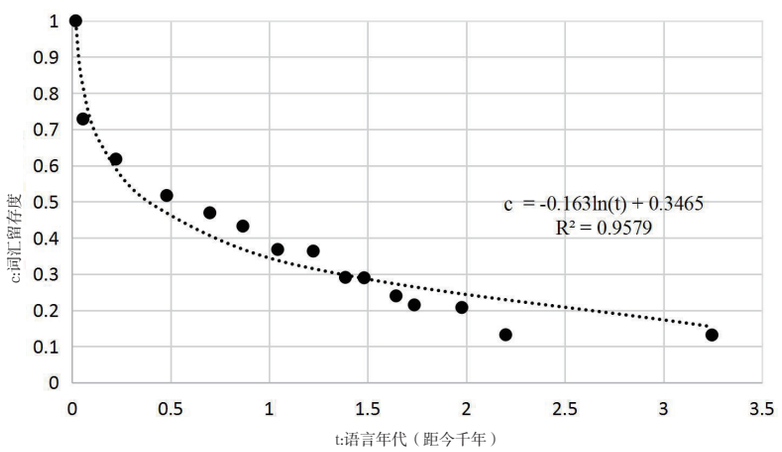

图6　词汇的时代回归分析

曲线拟合的结果，形成了类似于200核心词的对数函数曲线，在大数据量的历时词库上验证了斯瓦迪士的公式。图6直观地反映了词汇年代t和词汇留存度c之间的关系，t反映的是汉语时代词汇存在的距今年代，c反映汉语某一时代的词汇在当代语言中保留下来的词汇比例。在此基础上使用回归分析拟合出了对数回归方程，可用来计算汉语时代词汇存在的时间。该方程为：

$$c = -0.163\ln(t) + 0.3465 \qquad (2)$$

$$t = -\frac{\ln(c)}{0.163} \qquad (3)$$

拟合出的对数回归方程能较好反映两变量间的关系，拟合判定系数R^2为0.9579，接近于1，说明拟

①M. Swadesh, "Lexico-statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos", *Proceedings of the American Philosophical Society*, vol.96, no.4, 1952, pp.452—463.

合程度很高。可见，词汇距今年代与词汇留存度呈反比例对数关系。公式（2）在形式上与斯瓦迪士的公式（1）有较大差别，利用公式变形后，两者非常相似。舍弃常数项0.3465后，可以写成公式（3），进而根据公式（1）算得常数r为1.378。r的值相比斯瓦迪士的取值区间［0.72，0.85］较大，但考虑到本文是以3万多当代词汇作为计算依据，远超斯瓦迪士的200核心词，有一些差异仍算正常。

本文得到的拟合曲线和公式，对于汉语词汇史研究来说有重要参考价值。根据公式，可以算得4537年前（按1994年为当前时间），词汇留存度为0.1，即公元前2543年还有3000多个词和当代汉语相同。对比斯瓦迪士的公式（1），当词汇留存度为0.1时，约为7084年前。汉语的大词表留存度曲线明显要下降得快一些，这也比较容易解释。首先，核心词一般都较为稳定，数量也少，衰减速度慢。其次，大词表收词多达数万，词汇更替情况更容易发生。虽然两者在计算语言年代时存在一些差异，但本文的大词表方法仍提供了重要参考，将来可以应用于更多的年代计算，例如和各种方言和民族语言的大词表进行计算。可以利用已经出版的诸多方言词典的词表进行计算，与核心词得到的语言年代进行对比分析，更好地测定语言年代和分化年代。当然，《汉语大词典》收词不可能十分完备，但在目前很难找到更全的汉语历史词典的条件下，拟合出来的词汇留存度曲线已经显现出明确的对数曲线，与斯瓦迪士的公式极为相似，能够让我们更好地抓住汉语词汇演化的数学规律。

六、结语

历时真实文本收集加工的困难使汉语词汇研究集中于专书或断代的词汇研究和对词汇发展面貌的定性描写上，缺乏对词汇在各时代整体发展面貌的定量分析。本文依据反映词汇历时演变的历时语文词典《汉语大词典》，考证补全了88万多条书证材料的时代信息，建立了包含30多万词条的大规模汉语历时词库。其次，采用计量方法刻画了汉语词汇和义项在十五个时代发展的整体面貌。然后根据当代词汇的历代使用数据，拟测出了汉语词汇演变的统计学规律。当然，词典数据还不能完全代替历时的真实文本，这只是在现阶段条件不允许下的一种委曲求全的方式，我们计划进一步完善汉语历时词库的基础资源，利用各时期的真实语料补充词汇演变词条和用例，从而获得更为真实、精确的汉语词汇发展全貌。

（**责任编辑**：高　峰）

A Quantitative Analysis of Chinese Vocabulary Evolution Based on *The Great Chinese Dictionary*

LI Bin, LIU Xue-yang

Abstract: The evolution of Chinese vocabulary is the key research field of the Chinese language history. For the lack of the tagged diachronic corpus, the overall quantitative analysis of the evolution of the Chinese vocabulary is hard to achieve. *The Great Chinese Dictionary* recording senses of both ancient and contemporary words as historical resources was used. By manually labelling the periods of over 800,000 example sentences concerning over 300,000 entries in the dictionary, a diachronic Chinese lexical database was constructed. Then the number and word length of contemporary vocabulary in each historical period were presented and the correlation between the number of words and their age was estimated by regression analysis. This study provides basic resources and facilitate the use of quantitative analysis in the study of Chinese language history.

Key words: *The Great Chinese Dictionary;* vocabulary evolution; Chinese language history; glottochronology