2017 年 5 月　　云南师范大学学报　　May，2017
第 49 卷第 3 期　　(哲学社会科学版)　　Vol.49 No.3

基于语料库的汉语字母词自动抽取与分析*

黄居仁，　刘洪超
(香港理工大学 中文与双语学系，香港 999077)

[摘　要]　本文为汉语字母词研究提出新视角。首先，我们实现了基于语料库的汉语字母词自动抽取和计量分析。其次，我们在理论上澄清了汉语字母词中的字母并非拼音字母(letter)，而是书写字母(alphabet)。同时从音韵、构词与书写系统理论出发，进一步分析了汉语字母词对现有语言学理论的挑战及字母词分析对语言本体研究可能具有的重大贡献。基于对组字字母 K 的分析，我们主张组字字母的语言行为与汉字类似，而与拼音字母有很大的差距。最后，本文强调字母词的研究是语言科技对语言生活影响的一个重大课题。

[关键词]　汉语字母词；语料库语言学；计量语言学；语言科技
[中图分类号] H0-0　　[文献标识码] A　　[文章编号] 1000-5110(2017)03-0010-12

一、研究背景

字母词是指汉语书写时带有字母的词。字母词主要包括两种类型，一类全部由字母组合而成，如：

1) 日立和 JVC 可能退出中国低价彩电市场。

2) 当第二次国际电讯会议时，曾有人建议“SOS”作为统一的求救信号，还是没有得到一致同意。

3) 贸易通短短两年间依据市场和用户需求，不断进行人性化设计，在即时通讯软件中，其下载量仅次于 QQ 和 MSN。

另外一类则是与汉字组合在一起，汉字可以出现在字母的前、后和中间：

4) 章明基说，不管赢不赢全国男篮甲 A 冠军我都能去 NBA。

5) 虽然是物理学博士，但一样照穿帆布球鞋，T 恤的图案也总有些特别的意味。

6) 举个现代的例子，俄亥俄可以支持三 K 党，纽约却不行。

字母词中的字母不仅可以是英文字母 A～Z 和 a～z，还可以是其他西文字母如：

7) 20 世纪 60 年代以后，科学家通过用紫外线或 γ 射线照射甲醛的水溶液，还获得了核糖和脱氧核糖。

但是由于字母词中最常用的是英文字母，希腊字母等其他字母比较少见，所以本文的研究主要集中于含有 26 个英文字母的字母词。

根据张铁文①研究，20 世纪初的文献记载显示，汉语中最早出现的字母词很有可能是对英文 x-ray 的翻译，即“x 射线”“x 光线”或“x 光”，其更明确指出，1903 年出版的《新尔雅》收录了“X 光线”一词，很可能是第一个收入中文辞典的字母词。在之后 60 多年中，汉语字母词的使用范围主要局限在科技术语领域，只有极个别几个字母词的词汇地位得以确认并被编入汉语词典中，如“K 金”等。

20 世纪末，字母词的地位发生了翻天覆地的变化。《现代汉语词典》(中国社会科学院语言研究所

---

* [作者简介] 黄居仁，男，台湾人，香港理工大学讲座教授，博士，研究方向为词汇语义学、计算语言学、语料库语言学、知识本体等。

[基金项目] 香港研资局 PolyU 5435/12H 及北京大学-香港理工大学汉语语言学联合研究中心项目。

① 张铁文.词源研究与术语规范——X 射线词族的词源研究[J].产品安全与召回，2005，(1).

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

词典室，1996.以下简称《现汉》）在 1996 年为字母词专设一节，共收录了 39 个条目。《现汉》是官方词典，字母词的收录表明了官方权威对字母词地位的认可。该列表数目增长迅速，截止到 2012 年第 6 版，《现汉》中收录的以西文字母开始的词语达到了 239 个，引发了巨大的争论，也催生了一大批相关的语言学研究。①

值得注意的是，我们使用“字母词”（Mandarin Alphabetical Word，MAW）来称呼这些词语。使用“alphabet”而不用“letter”，主要基于以下几个原因：

首先，使用字母词能够突出这些词的汉语地位。它们是整合了字母字形系统的汉语词，而不是外文词，也不是两者的混合体。

其次，使用 alphabet 指字母词能够避免引起一些混淆和误解。使用 letter 可能被解释成“由字母组成的词”（letter words）或“以字母书写的词”（lettered words）。容易让人误以为这些词只包含拼音字母（letters）。甚至会让人认为这些词能够像西文一样，依照拼音字母（letters）组合的方式“拼读”出来。比如说，fun，not，mad 这 3 个词，都是英文的三字母词，每个词由 3 个字母（letters）组成。“fun”这个词组成的 3 个拼音字母 f，u，n，主要的功能是表达这个词的语音形式。这个定义与用法，事实上也可实现在以汉语拼音表达的中文文本上。比如说拼音文本中，门（mén）是三字母词（three-letter word），妈（mā）是二字母词（two-letter word），掉（diào）是四字母词（four-letter word）等。换句话说，用拼音书写时，每个汉语的词都是“letter word”。当文献中讨论字母词时，是完全不考虑汉语拼音文本的。因此我们采用“alphabet”，强调这个字母是中文汉字书写系统里作为书写单位，而非拼音单位的特性，以避免用“letter”时与汉语拼音书写系统可能会出现的混淆。

汉语字母词的语言学表现比较特殊，虽然是汉语词语，但是却表现出了一些非汉语的特征，充满了矛盾。

第一，汉语词和非汉语音系特征的矛盾。字母词缺乏普通汉语词汇所具有的典型声调，并且允许非汉语音节甚至是音位进入。一般说来，字母词中的字母都不带有汉语声调，如 MTV；一些汉语中不存在的音节或音位也可以出现在字母词中，前者如 K（[kʰeɪ]）、Q（[kʰjy]），后者如 V（/v/）。换句话说，字母词中的一些音节甚至是音位，并不符合汉语音系系统，也违反了音系的自主性，因为字母词的字形系统决定了其音系性质。

第二，借词和本土新词的矛盾。尽管字母词在很多地方都表现出了借词的音系调整特征（phonological adaptation），但是许多字母词事实上是本土新造词（indigenous neologisms），而不是借词。例如“AA 制”，根据亓华的研究，“AA 制”是 20 世纪 80 年代在香港开始流行的新词语，意思为“聚餐后各付各账的方法”，AA 是英语“each each”简化，后来由移民澳大利亚的香港人传播开来。② 由此看来，“AA 制”并非属于借词，而是汉语使用者根据英语和汉语的语言学规则自己新造的词语。

第三，在字形相关层，汉语字母词同时兼具英语词语的音系相关特征（phonologically relevant features）和汉字的语义相关特征（semantically relevant features）。根据 Sproat ③提出的字形相关层（ORL，orthographically relevant level），汉语字母词同时表现出了英语的音系相关特征和汉字的语义相关④特征。比如对于“K 他命、拉 K、K 仔”等词语，K 并不仅仅是一个简单的音节 [kʰeɪ]，它还表示 ket-

① 苏新春，吴晓芳.字母词的生命力与局限性——兼论《现代汉语词典》处理字母词的慎重做法［J］. 北华大学学报（社会科学版），2013，（2）；张铁文.字母词使用是语言接触的正常现象［J］. 北华大学学报（社会科学版），2013，（2）.

② 亓华.“AA 制”的“AA”源自何方［J］. 语文建设，2001，（8）.

③ Sproat，R.W..A computational theory of writing systems［M］. Cambridge：Cambridge University Presss，2000.

④ Huang，C.R.Semantics as an Orthography-Relevant Level for Mandarin Chinese［A］.Paper presented at the The 17th Annual Conference of the International Association of Chinese Linguistics，Paris.2009

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

amine,字符 K 兼具了英文的表音功能和汉字的表意功能。对于字母词的解释,显然这两方面缺一不可。

此外,字母词还有其他一些矛盾的语言学表现,后文将一一讨论。

二、相关文献

近期关于汉语字母词的研究,尤其是中国大陆进行的相关研究,主要从词汇学和语言政策角度切入。① 同时,从社会文化角度出发研究汉语字母词对全球化的影响也是一个非常热门的话题。② 但是,关于汉语字母词的形态-词汇地位(morpho-lexical status)的讨论却非常少,只有 Riha and Baker、Riha 以及 Lun 等几篇。③ 这些研究多半集中在个别实例,缺乏全面视野与理论解释。同时,我们没有发现关于汉语字母词音系调整(phonological adaptation)这个对某一语言中的单位借入其他语言系统时,其音系地位所发生的变化的理论探讨。

汉语中关于音系调整的研究非常丰富,但更多的是集中在音位层面上,比如 Lin 关于元音调整的研究,④Yip 关于语法系统和感知之间关系的研究。⑤ 与音系调整研究相似,汉语中关于声调调整的研究主要是集中在考察某一个或某几个声调如何调整入或调整出某一语言声调系统,例如 Hsieh and Kenstowicz 以及 Kenstowicz and Suchato 的相关研究,⑥而关于同一语言中声调和非声调系统并存的研究则明显不足。因此,Sproat 所提出的文字系统相关层面(ORL,orthographically relevant level)的概念⑦对于本文来说则显得尤其关键,可以用来对比字母词中的汉字和字母两种拼写系统。

正如前文所说,汉语字母词融合了汉字和字母两种字形系统,所以需要对这两种完全不同的系统进行感知,而音系调整研究中与之高度相关的一个话题是感知在音系调整中所起的作用。Peperkamp 等人进行了一系列的研究,⑧他们认为音系调整是由感知驱动的,Jong and Cho 的观点与之相似。⑨ Peperkam 团队后来所进行的研究⑩进一步表明,拼写和一些社会因素在音系调整中同样发挥了重要作

---

① 苏新春,吴晓芳.字母词的生命力与局限性——兼论《现代汉语词典》处理字母词的慎重做法[J].北华大学学报(社会科学版),2013,(2);张铁文.字母词使用是语言接触的正常现象[J].北华大学学报(社会科学版),2013,(2).

② Kozha,K.Chinese Via English: A case study of "lettered-words" as a way of integration into global communication[A].J.Liu & H.Tao.Chinese under globalization: emerging trends in language use in China[M].WORLD SCIENTIFIC.2001: 105~125; Miao,R.Loanword adaptation in Mandarin Chinese: perceptual,phonolgical and sociolinguistic[Z].Stony Brook University.2005.

③ Riha,H.,& Baker,K.Lettered Words: Using Roman Letters to Create Words in Chinese[J].ResearchGate,2010; Lun,K.Y.Morphological Structure of the Chinese Lettered Words[A].University of Washington Working Papers in Linguistics 30[C].2013; Riha,H.Lettered Words In Chinese: Roman Letters As Morpheme-Syllables[A].M.L. B.J.S.K.Campbell-Kibler & P.W.Culicover.The Ohio State University Linguistics Working Papers[C].Columbus: The Ohio State University,2010.

④ Lin,Y.H.Variable vowel adaptation in Standard Mandarin loanwords[J].Journal of East Asian Linguistics,2008,(4).

⑤ Yip,M.The symbiosis between perception and grammar in loanword phonology[J].Lingua,2006,(7).

⑥ Hsieh,F.F.,& Kenstowicz,M.J.Phonetic knowledge in tonal adaptation: Mandarin and English loanwords in Lhasa Tibetan[J].Journal of East Asian Linguistics,2008,(4); Kenstowicz,M.,& Suchato,A.Issues in loanword adaptation: A case study from Thai[J].Lingua,2006,(7).

⑦ Sproat,R.W.A computational theory of writing systems[M].Cambridge: Cambridge University Presss,2000.

⑧ Peperkamp, S.A Psycholinguistic Theory of Loanword Adaptations[M].Paper presented at the Annual Meeting of the Berkeley Linguistics Society,2004; Peperkamp,S.,& Emmanuel,D.Reinterpreting Loanword Adaptations: The Role of Perception[A].Paper presented at the Proceedings of International Congress of Phonetic Sciences.2003.

⑨ Jong,K.D.,& Cho,M.H.Loanword phonology and perceptual mapping: Comparing two corpora of Korean contact with English[J].Language,2012,(2).

⑩ Vendelin,I.,& Peperkamp,S.The influence of orthography on loanword adaptations[J].Lingua,2006,(7); Lev-Ari,Peperkamp,S.,& Sharon.An experimental study of the role of social factors in language change: The case of loanword adaptations[J].Laboratory Phonology,2014,(3).

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

用。因此，Peperkam 等人的观点实际上发生了变化，认为音系调整是由多因素驱动，而非感知这一单一因素驱动，这与 Aktürk-Drake ①以及 Smith② 等人的观点是一致的。而 LaCharité and Paradis ③与 Peperkamp 等人观点不同，他们认为借词的音系调整是由音系规则性（phonological regularity）驱动的。

另外，Kubozono ④以及 Labrune⑤关于日语借词的韵律调整研究对我们具有借鉴意义；Tsvetkov and Dyer ⑥的词典分层对于我们解释汉语字母词的词汇-形态地位提供了理论支持；Sproat ⑦所提出的 ORL 对我们对比汉字和字母字形系统影响提供了基础。借助 ORL，Huang ⑧以及 Huang and Hsieh⑨的观点，正是由于字形表征存在二分性（dichotomy of orthographic representation），汉语字母词才表现出一些特殊语言学特征。

最后，在语言科技与中文信息处理的研究方面，不能忽视字母词的辨识、合成、抽取、分类、搜寻等问题。作为这个议题的第一步，是字母词分词的问题，Huang ⑩以及 Huang and Xue ⑪分析了汉语字母词对汉语分词和语言处理所带来的一些挑战。

三、字母词的自动抽取和分类

本文对字母词的抽取是基于大规模语料库。《现汉》收录了 239 个字母词，尽管相比以前，这个数目已经增长了百倍，但是比起大规模语料库中字母词的分布来说则仍然过少，因而不容易观察字母词的语言学特征。本文所使用的语料库为“研究院语料库”（Sinica Corpus，Chen et al.1996）和带词类标注的中文亿词级语料库（Chinese Gigaword Corpus 2.0.，Huang，2009）。前者的规模为 1000 万词，后者的规模为 140 亿词；前者经过手工检查，数据质量高，后者为自动分词标注，质量不如前者；前者为平衡语料库，后者则主要为新闻语料。前者语料来源多样，后者规模庞大，保证了字母词抽取的充分性。

字母词的抽取采用基于规则的方法。字母词的特征比较明显，主要是含有 26 个英文字母（包括大写和小写）的词语，因此，字母词的抽取相对容易。从两个语料库中共抽取了 56833 个字母词（未经过手工检查）。中文存储存在全角和半角之分，本文在抽取时对两者进行了合并（合并为半角表示），即不区分全角和半角字母词。

---

① Aktürk-Drake，M.Language dominance as a factor in loanword phonology [Z] .International Journal of Bilingualism.2016

② Smith，J. Loan Phonology Is Not All Perception: Evidence From Japanese Loan Doublets [J] . Japanese/korean Linguistics，2006，(14) .

③ LaCharité，D.，& Paradis，C.Category Preservation and Proximity versus Phonetic Approximation in Loanword Adaptation [J] .Linguistic Inquiry，2006，(2) .

④ Kubozono，H.Where does loanword prosody come from? A case study of Japanese loanword accent [J] .Lingua，2006，(7) .

⑤ Labrune，L.The prosodic structure of simple abbreviated loanwords in Japanese: a constraint-based account [J] .Journal of the Phonetic Society of Japan，2002，(6) .

⑥ Tsvetkov，Y.，& Dyer，C. Lexicon Stratification for Translating Out-of-Vocabulary Words [A] . Paper presented at the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing. 2015.

⑦ Sproat，R.W..A computational theory of writing systems [M] .Cambridge: Cambridge University Presss，2000.

⑧ Huang，C.R.Semantics as an Orthography-Relevant Level for Mandarin Chinese [A] . Paper presented at the The 17th Annual Conference of the International Association of Chinese Linguistics，Paris.，2009.

⑨ Huang，C.R.，Hsieh，S.K.，& Chen，K.J.Chinese lexical semantics: from radicals to event structure [A] .In W.S.Y.Wang & C.F.Sun (Eds.) , The Oxford Handbook of Chinese Linguistics [M] .New York: Oxford University Press，2015: 290~305.

⑩ Huang，C.R.What You Need to Know about Chinese for Chinese Language Processing [A] .Paper presented at the Acl: Association of Computational Linguistics Meeting，2015.

⑪ Huang，C.R.，& Xue，N.Words without Boundaries: Computational Approaches to Chinese Word Segmentation [J] .Language & Linguistics Compass，2012，(8) .

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

为了进一步观察字母词的语言学特征,在自动抽取的同时,也应对其进行自动分类。见图 1。

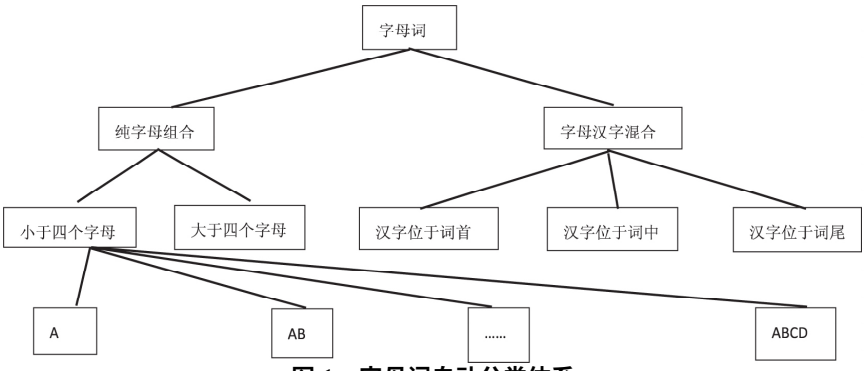

图 1 字母词自动分类体系

首先根据是否含有汉字,将字母词分成纯字母组合和字母汉字混合两种类型。纯字母组合的字母词一般表示简称或英文全称;字母汉字混合的一般表示人名、分类、分组等:

8) GG、WWW、BBC、CCTV

9) 甲 A、小 S、程 I 青、H5N1 型

纯字母组合根据字母多少进一步分类。小于 4 个字母的一般表示简称,大于 4 个字母的则一般表示英文原文或全称:

10) CC、OS、CCD、HTTP

11) CITES、INTERNET、CHINA、TAIWAN

小于 4 个字母的纯字母组合则根据字母多少和字母重复模式进一步划分为 23 个小类。下面只列出各个类别中词频前 4 位的词语:

12) A: A、B、C、O

13) AA: OO、CC、II、TT

14) AAA: OOO、WWW、III、BBB

15) AAAA: XXXX、AAAA、OOOO、YYYY

16) AAB: BBC、PPM、BBS、DDR

17) AABA: VVVF、AAAS、CCCP、IIIR

18) AABB: AAPP、IISS、DDBB、CCOO

19) AABC: CCTV、SSCI、CCID、IIPA

20) AB: IC、OK、CD、CI

21) ABA: DVD、CAC、TFT、SOS

22) ABAA: CICC、AIAA、AFAA、SPSS

23) ABAB: WEWE、YOYO、GIGI、COCO

24) ABAC: MDMA、ABAC、ITIS、FIFA

25) ABB: CNN、NII、NTT、FAA

26) ABBA: ELLE、ACCA、ABBA、OTTO

27) ABBB: IEEE、PIII、WEEE、FCCC

28) ABBC: SAAB、CMMI、GOOD、HTTP

29) ABC: WTO、NBA、IBM、GDP

30) ABCA: SARS、IFPI、CNBC、HIGH

31) ABCB: SOGO、NASA、IAEA、FAPA

32) ABCC: GATT、QFII、PECC、CALL

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

33) ABCD：APEC、DRAM、TVBS、CEPA

含有汉字的字母词根据汉字所处位置可以进一步划分为 3 个小类。下面同样列出每个类别中词频最高的 4 个词语：

34) 汉字位于词首：甲 A、小 S、三 T、甲 B

35) 汉字位于词首尾：朱 F 基、程 I 青、刘 X 福、简 X 燕

36) 汉字位于词尾：B 股、T 恤、H 股、B 组

四、统计与分析

本小结从一些统计数字上观察前文抽取的汉语字母词的规律。首先介绍一些统计量。语料库语言学通常区分词形频率（type frequency）和词例频率（token frequency）①前者是指在不重复计算的情况下，语料库中词的出现次数，其实也就是语料库的词库规模；后者是指在重复计算的情况下，语料库中出现的词的个数，其实也就是语料库的规模。比如，假设一个语料库中只有一句话：

37) 不同社会，有着不同性质、不同特点的教育。

在该句话组成的语料库中，共有“不同”“社会”“有”“着”“性质”“特点”“的”和“教育”共 8 个词，所以词形频率为 8。这 8 个词中，“不同”的词频为 3，其他所有词的词频都为 1，所以词例频率为 10。

词形/词例比（TTR，type/token ratio）在语料库语言学和计量文体学（quantitative stylistics）中常用来模拟语料库中的词汇变化（lexical variation）程度。两者比值越大，词汇变化越大，反之，则变化越小。比如，通常说来，书面语的词形/词例比都高于口语，因为书面语的思考时间长，可选词库大；而口语的思考时间短，无法选择比较丰富的词汇。② 有的也用词汇密度（lexical density）来衡量词汇丰富度，两者计算公式相似。③ 如果从使用率的角度来看，词形/词例比越高，表明每个词形的使用率越低。但是，如果从构词法的角度看，某个构词方式的词形/词例比越高，表示该构词法的能产性越高。

所有这些统计量都是基于词频，即某个词在语料库中出现的次数。本小结主要采用前文提到的这些统计量分析字母词。

（一）纯字母组合字母词

表 1　纯字母组合字母词的分布

模式	词例频率	词形频率	词形/词例比	模式	词例频率	词形频率	词形/词例比
A	239492	26	0.00	ABAB	848	69	0.08
AA	12620	26	0.00	ABAC	3253	375	0.12
AAA	2968	17	0.01	ABB	11121	276	0.02
AAAA	69	9	0.13	ABBA	267	45	0.17
AAAB	21	9	0.43	ABBB	289	13	0.04
AAB	7640	289	0.04	ABBC	2304	392	0.17
AABA	131	18	0.14	ABC	283938	6177	0.02
AABB	155	18	0.12	ABCA	16160	429	0.03
AABC	976	235	0.24	ABCB	6713	552	0.08
AB	129119	649	0.01	ABCC	10480	447	0.04
ABA	11695	302	0.03	ABCD	96494	7488	0.08
ABAA	45	16	0.36	大于五个字母	108264	38055	0.35

从上表可以观察到纯字母组合字母词的一些有意思的规律。

① Cortes，V.Longman Student Grammar of Spoken and Written English [J].Tesol Quarterly，2003，(4).

② Cortes，V.Longman Student Grammar of Spoken and Written English [J].Tesol Quarterly，2003，(4).

③ Ure，J.Lexical density and register differentiation [A].In G.Perren & J.L.M.Trim (Eds.)，Applications of Linguistics [M].London：Cambridge University Press.1971：443～452.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

第一,ABC 模式的字母词的词例频率最高。这表明在语料库中碰到 ABC 模式的词的概率最高,这与日常生活的经验相似,纯英文字母词一般用于表示简称,而简称中 ABC 模式的较多。图 2 是 ABC 模式中词频排名前 10 的字母词,它们的词频都在 5000 以上,而且都是大众非常熟悉的一些字母词,如 WTO、NBA、IBM、KTV、DNA 等。

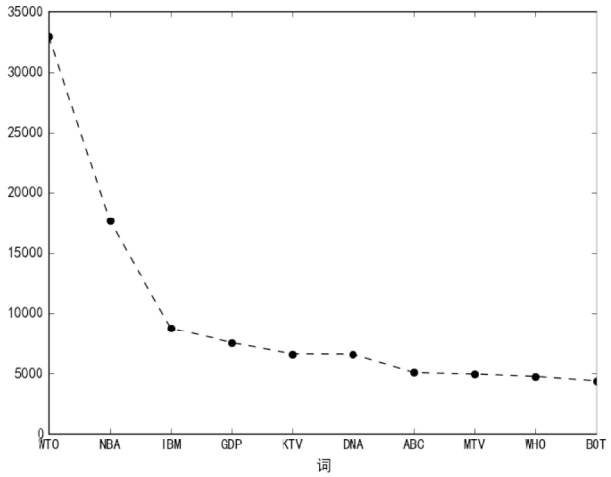

图 2 ABC 模式词频前 10 的字母词

事实上,进一步观察会发现,如果把同样字母数目的模式词例频率相加,会发现三字母模式(AAA、AAB、ABB、ABC)的词例频率是最高的;单字母模式(A)次之;双字母(AA、AB)跟四字母模式(AAAA、AAAB……ABCD)第三;最小的是 5 个字母以上的纯字母组合。见图 3。

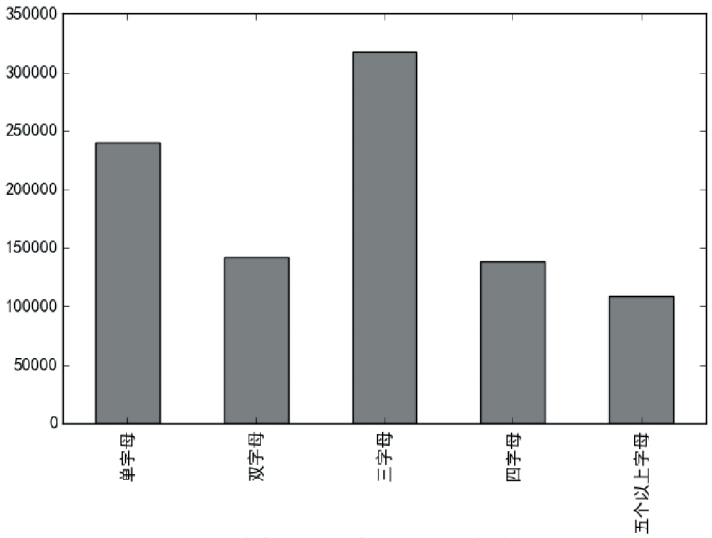

图 3 纯字母组合字母词词例频率分布

第二,大于 5 个字母的字母词词形频率最高。大于 5 个字母的纯字母组合字母词一般是直接引用英文专名或英文词,基本上来源几乎无限。所以不难理解其词形频率最高。语料库的词库中数量最多的是这类词,占据了所有抽取的字母词的 67%。相对而言,这些词形在语料中的出现,多半表现的是语言转码(code-switching)的现象,对中文词汇的直接影响较小。见图表。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

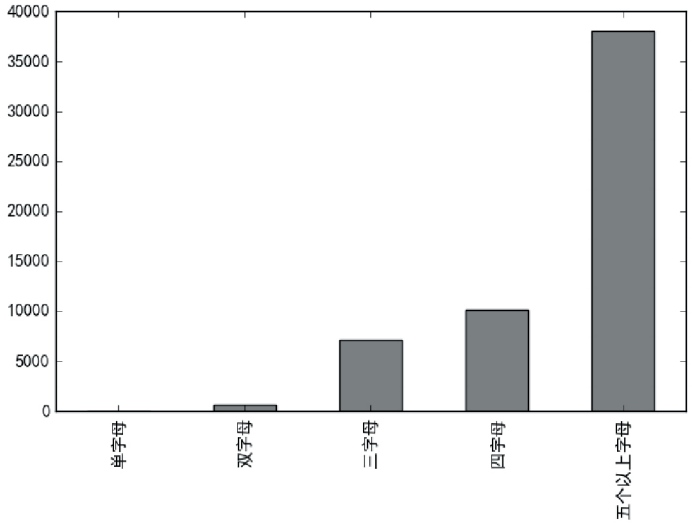

图 4　纯字母组合字母词词形频率分布

当与具体某个模式对比时可以发现，虽然 5 个字母以上的字母词在词库中占据了最大的比例，但是该模式的词例频率远远低于三字母模式中的 ABC 和单字母模式(A) ，也略低于双字母模式中的 AB。根据图 3 和图 4 可以看出，五字母模式词形数量虽大，但是每个词形的词频并不高，这导致其总体在语料库中的出现频率(即词例频率) 反而是最低的。

对比汉字词的词长分布会发现两者有很大不同。我们统计了研究院语料库中的汉字词的词长分布情况:

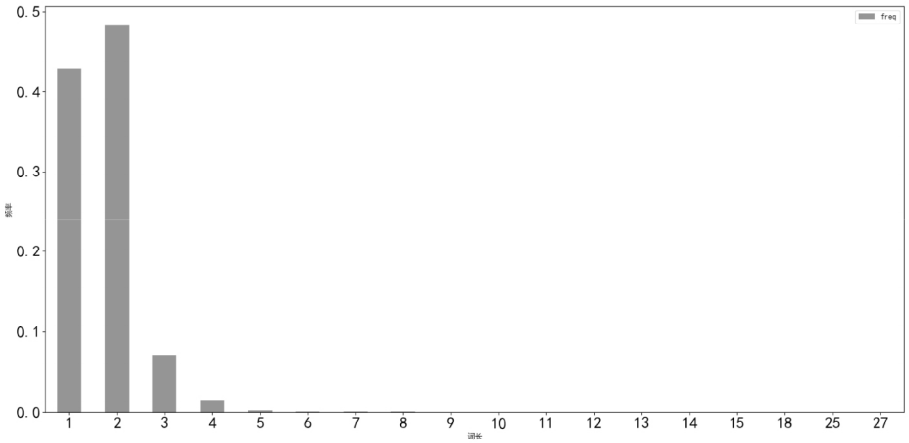

图 5　研究院语料库汉字词词长分布

从图 5 可看出，汉字词单字词和双字词的词形频率占绝对优势。这与 Chen et al.①和 Huang et al. ②的调查是一致的。

① Chen，C.Y.，Tseng，S.F.，Huang，C.R.，&Chen，K.J.Some Distributional Properties of Mandarin Chinese: A Study based on the Academia Sinica Corpus [A] .Proceedings of the Pacific Asia Conference on Formal and Computational Linguistics [M] . Taipei: Computational Linguistics Society，1993: 81～95.

② Huang，C.R.，Chen，C.J.，& Shen，C.C.C.The Nature of Categorical Ambiguity and Its Implications for Language Processing: A Corpus-based Study of Mandarin Chinese [A] .In Mineharu Nakayama (Ed.) .Sentence Processing in East Asian Languages [M] .Stanford，California: CSLI Publications，2001: 53～83.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

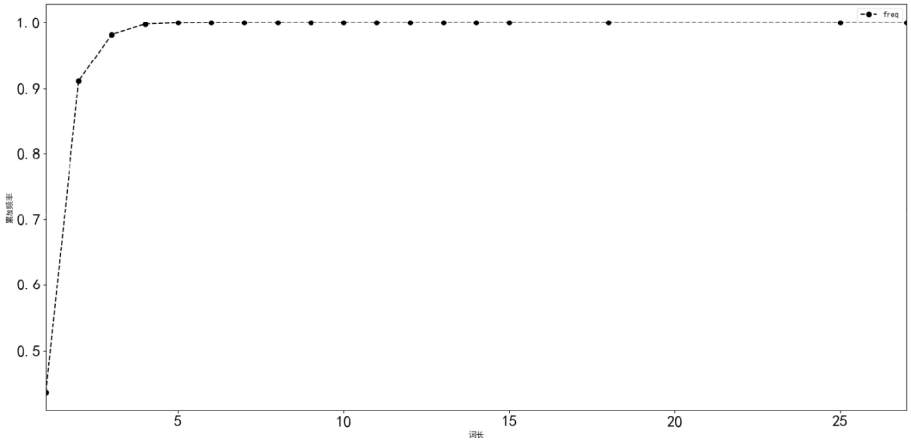

图 6　研究院语料库汉字词词长累加频率

我们发现,研究院语料库中单字词占比 43.8%,双字词占比 47.3%,三字词占 7%,四字词占比 1.6%。从图 6 可以看出,两字以下的汉字词占比超过 90%,三字及以下的汉字词占比超过 98%,四字词以上的汉字词非常罕见。这与字母词中 5 个字母以上的字母词占比最具优势非常不同。这种差别实际上表明,字母词的组合方式还没有明确的构词限制,所以每增加一个词长(增加一个字母位置),则词形可增加 26 倍不同的组合可能性,因此,对于字母词来说,词长越长,词形越多,即词形频率越高;而汉字词则不同,汉字词的词长越长,构词法对其限制越多,增加一个词长位置并不能增加更多的构词可能性。

大于 5 个字母的纯字母组合字母词,其词长也有规律可循:

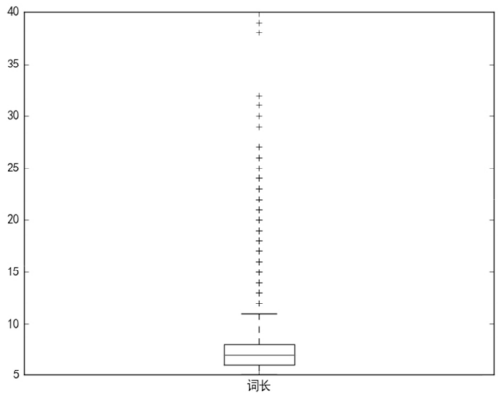

图 7　5 个字母以上字母词的词长分布

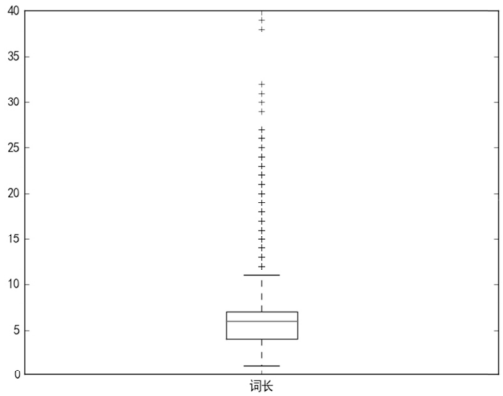

图 8　纯字母组合字母词的词长分布

图 7 为 5 个字母以上字母词的词长分布情况。箱型图用来描述数值分布非常直观方便。上下两端的分离横线是异常值的分界线,计算公式为:

上限: Q3 + 1.5 IQR

下限: Q1-1.5 IQR

Q1 和 Q3 表示第一和第三四分位数,即数据中处于 25%和 75%位置的数; IQR 为四分位距,即 Q3-Q1。在上下限之外的都是异常值,这里只有超出上限的异常值。

小方框的上下界限分别是 Q1 和 Q3。中间的线为中位数。

从箱型图可以看出,Q1 为 6,Q3 为 8,中位数为 7。所以绝大部分 5 个字母以上组合的字母词,词长都分布在 3~11 之间。异常值中最大的为 39。

可以进一步考察所有纯字母组合字母词的词长(图 8)。Q1 为 4,Q3 为 7,中位数为 6。所以纯字母组合字母词的词长绝大多数都分布在 1~11 之间。

第三,AAAB 模式的词形/词例比最高,A 以及 AA 模式的词形/词例比最低。AAAB 模式的字母词

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

（如 IISI、AAPA 等），词形频率和词例频率都不高，词形/词例比高表明词汇变化（lexical variation）程度比较高，每个词的使用率低。A 及 AA 模式的词汇变化小比较容易解释，因为可以出现在这两种模式中的词本身比较少（不区分大小写，每个位置只有 26 种可能，而 AA 由于是重复模式，因此也只有 26 种可能），而这些字母词的词频都非常高，因此，这两种模式的词形/词例比值都比较低，词汇变化少，每个词的使用率高。

四字母模式的组合有 15 种情况，从理论上讲，它们的出现概率都是一样的，但是在语料库中观察到的情况却并非如此。四字母模式中词形/词例比最低的是 ABCA 模式（如 SARS、CNBC 等），其词形频率、词例频率都远远高于 AAAB，词形/词例比低表明其每个词的出现频率也远远高于 AAAB，甚至高于或接近于某些三字母模式如 AAB（如 BBC、BBS 等）和 ABA（如 DVD、SOS 等）。

（二）含有汉字的字母词

含有汉字的字母词数量远远小于纯字母组合的字母词，只有 901 个。这个规模与目前辞典收词的规模类似。

表 2　含有汉字的字母词分布

模式	词例频率	词形频率	词形/词例比
X 【C】	4310	457	0.106032
【C】X	1344	252	0.1875
【C】X 【C】	332	175	0.527108
其他	39	17	0.435897

表 2 中 X 表示字母，【C】表示汉字。从上表可以看出，不管是词形频率还是词例频率，最高的都是 X+【C】模式（如 B 组、T 恤等）。词形/词例比值最高，词形平均使用率最低的是字母位于词中的模式（如三 K 党等）。

含有汉字的字母词，词长普遍低于纯字母组合：

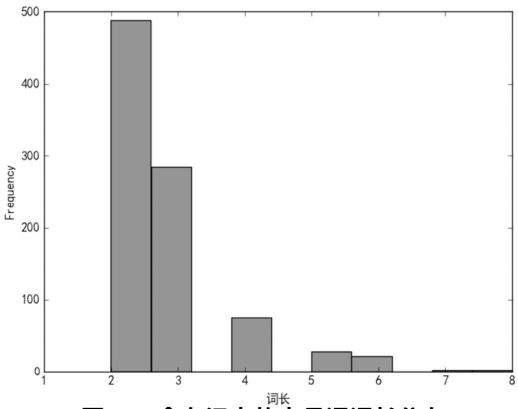

图 9　含有汉字的字母词词长分布

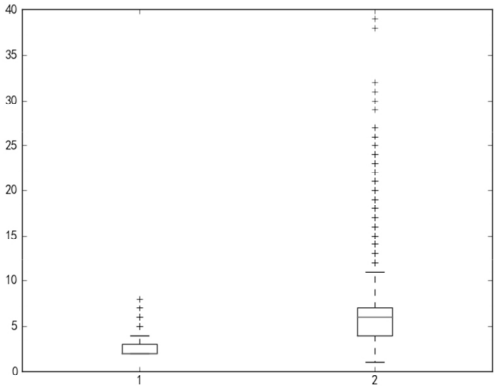

图 10　含汉字字母词与纯字母字母词词长对比

从图 9 可以看出，含有汉字的字母词词长大多为 2 或 3；图 10 箱型图中，1 代表含有汉字的字母词，2 代表纯字母字母词，可以看出，前者的词长大多分布在 2~4 之间，后者虽然分布在 1~11 之间，但是 Q1 和 Q3 以及中位数都大大超过前者。

纯字母组合和含有汉字的字母词的对比在语言分析上有相当重要的意义。Huang et al.①的研究显示汉语中二至四字词的词形占所有词词频的 88.51%；而如果只看名词则高达 90.47%。换句话说，含汉

① Huang，C.R.，Chen，C.J.，& Shen，C.C.C.The Nature of Categorical Ambiguity and Its Implications for Language Processing: A Corpus-based Study of Mandarin Chinese【A】.In Mineharu Nakayama（Ed.）.Sentence Processing in East Asian Languages【M】. Stanford，California: CSLI Publications，2001: 53~83.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

字的字母词的词形分布,与非字母词汉字词类似。而纯字母组合词的分布则与汉语现有词汇的分布差异相当大。从我们抽取的资料中看,出现词频最高的字母词,绝大多数都是含汉字的字母词,只有少数 2 至 3 个字母的纯字母词。这个分析,与我们在前文中提到 5 个字母以上的纯字母词基本上是直接引用外来词的语言转码是一体的两面。显示的是真正进入汉语词汇系统的字母词,以含汉字的字母词,或二至四字母的字母词为主。

五、字母词与构词字母的语言分析

为了更深入探讨字母词在汉语词汇系统中的地位,并深化字母词研究在汉语本体研究中的议题,我们在本节中以字母"K"为例,进行构词与词义的分析。

38) 由 K 组词的含汉字字母词

a.K 他命

b.拉 K

c.K 仔

d.K 金

e.K 歌

f.K 书(台湾用法)

g.三 K 党

39) 由 K 组词的纯字母词

a. K

b. OK

c. PK

d. KTV

e. YKK

f. LKK(台湾用法)

过去字母词研究中最常犯的错误,就是把字母词中的字母,当成英文字母来处理。这从语音、语义和构词来讲,都是错误的。初步观察的结果发现:

首先,以上这 13 个含 K 的字母词,"K"字的语音表现有相当大的差异。K 并不仅仅是一个简单的音节 [kʰeɪ]。最明显的区别是,39f) LKK 中的 K 读成 [kʰo] 而非 [kʰei]。虽然一些细微的音位差异需要更多资料来帮助分析,但是有的差别可以明显观察到,如 38a~d) 中的 K 通常读成接近阳平调,39c) 则接近去声。上海交通大学丁红卫研究小组初步成果显示,字母词声调可能因说话者的方言差异而不同。换句话说,一个用于组字母词的字母,可能代表多个音位。

其次,K 这个字母单位不但带有词义,并有词汇多义性。39a~c) 中 K 这个语素代表 ketamine 这种毒品。38d) 和 39a) 的 K 代表 karat,黄金纯度的度量(如"十六 K")。38e) 和 39d) 的 K 来自卡拉 OK 等等。换句话说,这些组字字母,和汉字类似,都可以有多义性,在不同构词环境中,代表多个不同的词义。

最后,组字字母能根据构词法构词。上面提到的 39a~c) 用 K 构词(其他还有 K 毒、K 粉、K 丸等)。39a) 的 K 用于数量复合词,起码有两个词义具衍生性,包括黄金与其他金属的 18K、24K 等;以及用于标示公路位置的 43K、105K 等(台湾用法)。而动词性的字母词则和其他汉字词相同,可以接时态词尾 40a)。含汉字的字母词,和其他动宾复合词相同,将时态直接依附在动词(即组字字母)上,如 40b)。

40) a. 林毅夫、张维迎北大朗润园 PK 了什么

b. K 了歌,K 过歌

简言之,组字字母在汉语字母词中呈现的语音、词义与构词行为,其实与汉字类似,而与拼音字母大不相同。我们可以大胆假设,组字字母在汉字书写系统中的地位,接近于汉字,是汉语以语义为文字系统相关层面的语言本体架构下创新的一环。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

六、结　论

本文以语料库为本，从语言本体分析的观点来看汉语的字母词。从语言本体分析来看，我们发现含汉字的字母词，其语言学表现与典型汉字词类似。而除了语言转码语境中出现的长纯字母词外，其他组字字母词的语音与构词特性，与汉字类似。而字母词中字母的读音，出现相当多不符合汉语音系的音位，完全违反了音韵学理论中的音位自主理论，以及外来新音位必须经过音系调整特征才能进入语言音系的理论。这个议题虽然不在本文处理范围，但字母词发音的声调呈现，受到发音者方言母语的影响（上海交通大学丁红卫教授研究小组初步成果）。这可以解释成音系调整特征在不同的语言层次呈现。这些讨论显示汉语字母词是当今语言学理论尚未能完善解释的特殊语言现象，字母词的研究有挑战并更新语言学理论的极大潜力。

最后，从字母词与科技的互动关系来看。从第一个字母词进入汉语词汇算起，已有 120 年。但是前 100 年，收入辞典的字母词，一共仅有 30 多个。而最近 20 年，新收入辞典的字母词不下数百个词条。我们由语料中抽取有成词潜力的含有字母的组合则有数万条，其中许多字母词沿袭了汉语词汇系统的大多数特性。这个重大的改变，合理的解释应该是数字载体（包括个人电脑与智能型手机）以及字母键盘输入法普及的影响。从语言科技影响语言生活角度看，汉语字母词更是一个不能不深入探讨研究的重要现象。

Corpus-based automatic extraction and analysis of Mandarin alphabetic words

HUANG Ju-ren & LIU Hong-chao

(*Department of Chinese and Bilingual Studies*, *the Hong Kong Polytechnic University*, *Hong Kong* 999077, *China*)

Abstract: This paper introduces a new perspective on the study of Mandarin alphabetic words (MAW). We first introduce the automatic extraction and quantitative analysis of Mandarin alphabetic words based on a large corpus. Second, we clarify the use of the term "alphabetic words" rather than "letter words", as all Chinese words written in Pinyin would be letter words. We further explicate the challenges and potential contributions to linguistic theories posed by an in-depth account of the phonology, morphology and writing system of Mandarin alphabetic words. Third, with a detailed analysis of the alphabet *K* in Mandarin alphabetic words, we suggest that the linguistic behavior of alphabet in Mandarin alphabetic words is similar to Chinese characters but different from letters. Finally, it is stressed that the study of Mandarin alphabetic words is a non-trivial subject on the influence of language life by language technology.

Key words: Mandarin alphabetic word; corpus linguistics; quantitative linguistics; language technology

责任编辑：　杨　恬

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net