# 现代汉语同音字家族系统的效率与易学性：来自计算与模拟的证据

**肖哲**
北京师范大学国际中文教育学院/ 北京，中国
xiaozhe@mail.bnu.edu.cn

**摘要**

信息论视角的语言研究揭示了语言系统中普遍存在的效率与易学性的认知约束。本研究探讨了现代汉语中同音字家族系统的认知约束，发现（1）在系统内部，家族效率与易学性正相关；（2）相比计算模拟系统和拼音化系统，同音字家族系统易学性虽较低，但效率更高；（3）无论是否考虑声调、声符和生僻字，家族系统均表现出上述特点。结果表明，汉语同音字家族系统对效率与易学性进行了权衡，揭示了其庞大规模形成背后的认知机制。

**关键词：** 同音字家族；效率；易学性；信息论；认知约束

# Efficiency and Learnability of the Homophone Family System in Modern Chinese Characters: Evidence from Language Computation and Simulation

**Xiao Zhe**
School of International Chinese Language Education, Beijing Normal University / Beijing, China
xiaozhe@mail.bnu.edu.cn

**Abstract**

From an information-theoretic perspective, linguistic research has shown that language systems are shaped by general cognitive constraints of efficiency and learnability. This study investigates these constraints in the homophone family system of Modern Chinese characters. Results show that: (1) within the system, family efficiency is positively correlated with learnability; (2) in comparison with computationally simulated systems and Pinyin-based systems, the attested homophone family system in Chinese exhibits lower learnability but higher efficiency; and (3) these properties are robust across different analytical conditions, including the presence or absence of tonal distinctions, phonetic radicals, and low-frequency characters. Taken together, the results suggest that the large-scale structure of the Chinese homophone family system reflects an underlying cognitive trade-off between efficiency and learnability.

**Keywords:** Homophone family , Efficiency , Learnability , Information theory , Cognitive constraints

## 1 引言



第二十四届中国计算语言学大会论文集，第247页-第258页，济南，中国，2025年8月11日至14日。

近年来，越来越多证据表明，人类语言系统不可避免地受到普遍认知规律约束 (Culicover and Jackendoff, 2012; Greenhill, 2023)。其中，信息论（information theory）视角认为，这种认知约束体现为语言系统的效率（efficiency）与易学性（learnability）间的统计依赖性 (Gibson et al., 2019)。例如，语音研究发现，相较于低效（或不易理解）的语音序列，儿童更易习得高效（或易于理解）的语音序列 (Coady and Aslin, 2004)。除语音系统外，类似的认知约束广泛存在于语义 (Kemp et al., 2018)、语法 (Koplenig et al., 2017)等其他语言子系统。然而，相关文献多聚焦于口语系统，文字系统是否同样受到效率与易学性的制约，尚缺乏系统性探讨。与此同时，相较于国际学界对此问题的持续关注，国内在汉语语境下的相关研究仍相对薄弱。对此问题的深入研究，不仅有助于拓展我们对语言认知约束适用边界的理解，也将为揭示汉语文字系统的形成机制提供理论依据。基于此，本文以现代汉语中极具代表性的同音字现象为切入点，探讨效率与易学性的认知约束是否同样适用于汉语的同音字家族系统。

### 1.1 效率与易学性的认知约束

现代信息论研究表明，人类语言系统所受到的认知约束主要涉及两个核心要素：效率（efficiency）与易学性（learnability） (Piantadosi et al., 2012; Gibson et al., 2019)。

信息论框架中，效率指语言信息传递的流畅程度，即信息的可预测性，可体现在语言的单元或整体系统上。效率越高，意味着语言单元或系统在传递信息时的可预测性越强，语言越易于解码或理解。信息论将效率量化为信息量（information content） (Gibson et al., 2019)。一个语言单元$I$ 的信息量$H(I)$ 通过其使用频率计算，见公式（ 1） (Gibson et al., 2019)：

$$H(I) = -\sum P(i) \log_2 P(i) \tag{1}$$

其中，$P(i)$ 表示元素$i$ 在单元$I$ 中的使用频率占比。整体语言系统的信息量则是系统内各语言单元信息量的平均值 (Kemp et al., 2018; Kemp and Regier, 2012)。信息量越小，表明系统越高效，信息传输越顺畅。

易学性指语言单元或系统在结构上的简约程度，可由语言单元或系统的规模进行量化，即其中包含的元素数量，例如，success（/skses/）的音素数为6，日语的音节数约为50。一般而言，越简洁的结构越易于学习掌握，因此易学性越高 (Gibson et al., 2019)。

信息论视人类语言为一种现代通信系统，该系统追求简洁、高效且稳健的传递机制。因此，语言在数千年、数亿人的实际使用中不断被优化。在此过程中，效率与易学性形成了内在的统计依赖性。这种关系既体现在语言系统的内部单元层面，也体现于整体语言系统层面。

首先，在语言单元层面，单元的大小受到其效率制约 (Gibson et al., 2019)，效率越高的语言单元，其结构往往越简约。其次，在整体语言系统层面，效率与易学性之间则表现出一种权衡关系。理论上，一个理想的通信系统应既高效又易学，但现实语言系统中二者往往难以兼得 (Kemp et al., 2018; Piantadosi et al., 2012)。高效率要求信息可预测性很高，使信源与信宿之间实现接近一一对应的映射，但这将导致系统极为庞杂（试想一人对应一个身份证号），进而降低其易学性。相反，高易学性要求系统结构简洁，但这往往以牺牲信息可区分性为代价（试想多人对应同一身份证号），降低了信息重构的准确度，从而影响传递效率。因此，语言系统在演化过程中往往需要在效率与易学性之间作出平衡。

### 1.2 效率与易学性认知约束的相关研究概述

效率与易学性对语言系统认知约束已在语言单元和整体系统层面都得到支持。首先，在语言单元层面，研究发现单元效率与单元大小之间的负向关系。例如，Piantadosi 等 (2011)研究了11种印欧语言中的单词，结果发现单词的信息量与词长呈显著正相关，即单词效率越高，词长越短。这表明词汇单元的结构受到效率的约束，语言系统可能通过压缩词长的方式实现效率优化。值得注意的是，该研究还发现单词信息量与词长的相关性显著高于词频与词长之间的相关性，进一步表明词长所受的约束主要源自效率压力而非使用频率。又如，Franzon和Zanini (2023)在研究意大利语名词的性范畴时也发现，阴性名词的信息量高于阳性名词，但其所指对象的范围更小，显示出效率对语言范畴结构的压缩。

在整体语言系统层面上，相关研究也发现了系统效率与易学性间的权衡关系。这些研究常见的策略是将真实语言系统与人造的计算模拟系统进行对比。例如，Kemp 和Regier (2012)对比多个自然语言与模拟语言中的亲属范畴系统，将各系统的效率量化为信息量，将其易学性量

化为系统中范畴的数量。结果发现，在487种自然语言中，亲属范畴系统的效率越高，则其结构越复杂，易学性越低；而在模拟系统中则未发现类似趋势。后续对颜色范畴的跨语言研究也进一步验证了这一现象 (Regier et al., 2015)。可见，效率与易学性之间的权衡关系是自然语言在长期演化过程中形成的稳定特征，反映了其作为人类认知系统的自然属性。

总之，无论是在语言单元层面，还是在整体语言系统层面，语言结构均受到效率与易学性的认知约束。然而，现有研究主要集中于口语系统，对于文字系统是否同样体现此类认知约束，尤其在口语与书面文字之间的对应关系方面，缺乏系统的实证考察。书面文字作为口语的记录形式，与口语共同演化数千年，很可能也在结构层面体现出相似的认知约束机制。有鉴于此，本文以现代汉语中具有代表性的同音字家族系统为研究对象，旨在探讨效率与易学性是否也制约着汉语口语系统与书写系统之间的映射关系。

### 1.3 现代汉语的同音字家族系统

同音字现象是指汉语中一个音节可对应多个汉字，体现了汉语语音与字形之间的一对多映射关系。共享同一读音的汉字形成一个同音字家族。例如，汉字“梦”与“孟”属于一个同音字家族“/meng4/”。

同音字数量庞大是现代汉语文字系统的一大特征 (邵敬敏, 2007)。据统计，现代汉语中高达80.49%的带调音节含同音字 (尹文刚, 2003)，同音字总数占全部汉字97.00%(苏新春和林进展, 2006)。从家族角度看，现代汉语中每个同音字家族平均含7个同音字，个别家族的同音字可达93个之多(肖哲等, 2024)。由此可见，现代汉语的同音字家族系统规模庞大，结构复杂。

如此庞大的同音字家族系统在一定程度上增加了汉字学习的认知负担。早在20世纪初，已有学者指出，同音字不利于汉字的识记与掌握，会干扰阅读理解，因而提出“汉字拼音化”的建议，试图借此缓解同音字所带来的学习难题 (李敏生, 1992; 王爱云, 2009)。当代研究也表明，大量同音字的存在是汉字相较于拼音文字更难学习和记忆的关键原因之一 (杨仕章, 1998; 叶蜚声和徐通锵, 2010)。无论是汉语母语者还是二语学习者，在汉字书写过程中均易出现同音别字的错误，反映出同音字对汉字学习过程构成的干扰 (杜同惠, 1993; 孟祥芝等, 2000)。

上述事实引发了一个重要理论问题：既然庞大的同音字家族不利于学习，为何现代汉语仍保有如此庞大的同音字家族系统？这是近年来汉语研究领域广受关注的问题之一。已有研究主要从语言学角度，基于汉语的语言结构特点及其历史演化展开解释。有学者指出，汉语的音节结构相对单纯，主要由声母和韵母构成，音节组合空间有限 (万业馨, 2021)；与此同时，汉语倾向于通过创造新字的方式应对指称新概念的需求，而非引入新的音节 (张学新, 2011; 柳英绿和关黑拽, 2014)。在音节数量有限而汉字总量不断增长的背景下，单个音节往往需对应多个汉字，从而导致同音字家族规模持续扩张。

然而，上述解释多基于汉语内部的结构性机制，对于同音字家族系统是否还受到更为根本、具有普遍性的认知约束，目前仍缺乏系统性的探讨。深入理解这一问题不仅有助于揭示现代汉语同音字家族现象的认知动因，也有望为重新审视“汉字拼音化”问题提供新的解释视角。

### 1.4 本研究

本研究旨在探讨现代汉语同音字家族系统是否受到效率与易学性的认知约束，并据此提出两个主要假设。

假设（1）：在系统内部单元层面，同音字家族单元的大小受其效率的约束。若假设成立，则同音字家族单元的大小与其效率呈负相关，即家族易学性与效率正相关。

假设（2）：在整体系统层面，同音字家族系统在效率与易学性之间表现出权衡关系。验证该假设需将真实的同音字家族系统与计算模拟系统进行对比 (Gibson et al., 2019)。根据信息论中关于效率与易学性的权衡原则，缺乏效率的系统通常具有较高的易学性 (Gibson et al., 2019)。因此，若该假设成立，那么与缺乏效率的计算模拟系统相比，汉字形式的同音字家族系统应表现出更高的效率，但其易学性则相对较低。

此外，围绕“汉字拼音化”这一议题，本文进一步比较了汉字形式的同音字家族系统与拼音化系统在效率与易学性方面的差异。拼音化系统中，每个音节对应一个唯一的正字法形式（即一个拼音符号），因此结构更为简化而更易学习。从效率与易学性之间的权衡视角来看，更高的易学性往往意味着效率的降低 (Gibson et al., 2019)。据此，本文进一步预期：与拼音化系统相比，真实的同音字家族系统在效率上应具有优势，而在易学性上则处于劣势。

## 2 数据来源与同音字家族构建

本研究所用数据库基于第七版《现代汉语词典》（以下简称《现汉》）(中国社会科学院语言研究所词典编辑室, 2016)与《现代汉语频率词典》（以下简称《频率词典》）(北京语言学院语言教学研究所, 1986)构建。

首先，根据《现汉》提供的带调音节表，录入所有带调音节及其所对应的汉字，一个带调音节对应的所有汉字构成一个同音字家族。例如，音节“/hai2/”对应汉字“孩”“骸”和“还”，因此这些汉字共同构成同音字家族“/hai2/”。若某一汉字为多音字，则可归入多个同音字家族，如“还”既属于“/hai2/”，又属于“/huan2/”。随后，根据《现代汉语频率词典》录入所有汉字的字频信息。例如，“孩”“骸”和“还”的字频分别为1249、5 和6034。若某字在《现汉》中出现而未被《频率词典》收录，则默认其字频为1。对于多音字，我们采用穷尽式方法，在《频率词典》中检索其在不同音节下的实际使用频率。例如，“还”的总频率为6034，其中“/hai2/”用法为5902，“/huan2/”为132。经上述处理，最终数据库包含1340 个带调同音字家族、10333 个汉字，总字频为1815583。数据库可在线获取（https://osf.io/96j4a/）。

## 3 同音字家族内单元效率与易学性的关系

本节检验假设（1），观察同音字家族系统内，家族单元的效率与易学性是否正相关。

### 3.1 家族单元效率与易学性测量

先前研究将单元的易学性量化为单位内包含的元素 (Franzon and Zanini, 2023; Piantadosi et al., 2011)。与先前研究一致，本文将同音字家族单元的易学性反向编码为家族内部单元的元素数，即家族汉字数（本文称为“家族大小”）。家族汉字数越多，其易学性越低。例如，同音字家族“/hai2/”对应汉字“孩”“骸”和“还”，因此其家族大小为3。

家族单元效率基于家族的使用频率计算 (Gibson et al., 2019)。家族的使用频率（以下简称“家族频率”）即家族内汉字字频之和，如同音字家族“/hai2/”的汉字（“孩”“骸”和“还”）字频分别为1249、5和5902，因此该家族频率为$1249 + 5 + 5902 = 7156$。

家族单元的效率反向编码为家族的信息量，家族信息量越大，其效率越低 (Piantadosi et al., 2011; Gibson et al., 2019)。若某同音字家族单元$F$(Family)含元素（即汉字）$C$(Character)，那么单元$F$的信息量$H(F)$ 可表示为公式（2），由 1.1节的信息量计算公式（1）改编：

$$H(F) = -\sum P(C|F)\log_2 P(C|F) \tag{2}$$

其中$P(C|F)$ 表示汉字$C$ 在家族$F$ 中的使用频率占比。例如，同音字家族“/hai2/”中各汉字（“孩”“骸”和“还”）的使用频率分别分别为：17.45%、0.07%和82.48%。调用公式（2）可得该家族信息量为：$-[17.45\%\times\log_2(17.45\%)+0.07\%\times\log_2(0.07\%)+82.48\%\times\log_2(82.48\%)] = 0.68$。

### 3.2 数据分析

首先计算家族信息量与家族大小之间的Spearman 相关系数以检验二者是否存在正相关 (Piantadosi et al., 2011)。为进一步明确家族大小的变化是否主要受家族效率（信息量）而非家族频率的影响，还计算了家族频率与家族大小之间的相关系数，并将其与家族信息量与家族大小的相关系数进行对比，检验族这两个相关系数是否存在差异 (Piantadosi et al., 2011)。

需指出的是，由于家族信息量本身基于家族频率计算，因此二者显著相关（$r = 0.21$，$p < 0.01$）。为剥离二者的潜在混淆，在分析家族大小与信息量或频率之间的关系时，采用了偏相关：控制家族频率，计算家族信息量与家族大小之间的偏相关；控制家族信息量，计算家族频率与家族大小之间的偏相关。

### 3.3 结果

如表 1 所示：在控制家族频率时，家族信息量与家族大小呈显著强正偏相关（$r = 0.83$，$p < 0.01$）；在控制家族信息量时，家族频率与家族大小亦呈显著中等正偏相关（$r = 0.66$，$p < 0.01$）。

为比较上述两个偏相关系数的差异，进行$Z$ 检验。结果显示，家族信息量与家族大小的偏相关系数显著大于家族频率与家族大小的偏相关系数（$r$ 差值$= 0.18$，$Z = 8.78$，$p < 0.01$）。

第二十四届中国计算语言学大会论文集，第247页-第258页，济南，中国，2025年8月11日至14日。

| 控制变量 | 计算变量 | 偏相关系数 |
|---|---|---|
| 家族频率 | 家族信息量，家族大小 | 0.83** |
| 家族信息量 | 家族频率，家族大小 | 0.66** |

** 表示$p < 0.01$。下同。

Table 1: 家族大小与家族信息量、家族频率的偏相关系数

综上，现代汉语同音字家族的信息量与家族大小呈显著正偏相关，即家族效率越高，则家族越小，显示出家族效率对家族大小的制约。同时，二者的相关强于家族频率与家族大小之间的相关，进一步支持家族大小主要受到效率而非频率的约束。这些揭示了现代汉语同音字家族单元的效率与易学性之间的统计依赖性，支持了假设（1）。

## 4 同音字家族系统效率与易学性的权衡

本节检验假设（2），通过比较真实的同音字家族系统与计算模拟系统及拼音化系统在效率与易学性方面的差异，考察同音字家族系统在整体层面上是否体现出效率与易学性的权衡。

### 4.1 建立计算模拟系统与拼音化系统

根据信息论关于效率与易学性的研究范式，一个语言系统在效率与易学性上的表现需通过与其他系统的比较加以评估 (Kemp and Regier, 2012; Gibson et al., 2019)。本研究以真实的同音字家族系统为蓝本 (Kemp and Regier, 2012; Regier et al., 2015)，分别构建了计算模拟系统与拼音化系统，以进行对比分析。

为构建计算模拟系统，我们首先设定一组与真实系统一致的基本参数作为约束条件：音节总数（即同音字家族数）为1340，汉字总数为10333，总字频为1815583。模拟系统的生成包括两个阶段。第一阶段为汉字分配至音节：将全部汉字打乱顺序后，依次分配至1340 个音节，每个音节初始分配1 个汉字，以确保所有音节均有对应的汉字。剩余汉字以每次一个为单位，均匀随机分配至各音节，直至全部分配完毕，由此形成模拟同音字家族结构。第二阶段为字频分配至汉字：在上述结构基础上，首先为每个汉字分配初始字频1，以确保频率非零；其余$1815583 - 10333 = 1805250$ 个字频单位则逐一随机分配至任一汉字，直至总频率分配完成。上述过程构成一个完整的模拟系统。我们据此重复该过程1000 次，生成1000 个模拟样本系统。该采样量通常足以使相关指标分布稳定并近似正态，从而支持对真实系统与模拟系统之间差异的统计检验 (Dautriche et al., 2017)。模拟系统构建代码可在线获取（https://osf.io/96j4a/）。

随后，我们构建了拼音化汉字系统。该系统中，每个汉字对应的拼音符号为其在《现汉》中所标注的拼音。例如，同音字家族“/hai2/”中的所有汉字（如“孩”“骸”“还”）在《现汉》中的拼音均为“hái”，因此在拼音化系统中，它们均由拼音符号“hái”表示。[1]对于多音字，其在不同语音下将分别对应不同拼音符号，例如“还”读作“/hai2/”时拼音化为“hái”，读作“/huan2/”时拼音化为“huán”。据此构建的拼音化系统共包含1340 个拼音符号，与1340 个同音字家族一一对应，各拼音符号的使用频率为其所对应全部汉字的字频之和。

### 4.2 测量指标

将各系统的效率反向编码为系统的信息量 (Kemp and Regier, 2012)。系统信息量越大，系统效率越低。对于真实的同音字家族系统及1000 个计算模拟系统，其系统信息量以各家族信息量的加权平均值计算，如公式（3）所示：

$$H(L) = \sum P(F|L) \cdot H(F|L) \tag{3}$$

其中，$H(L)$ 为系统$L$ （Language）的信息量，$P(F|L)$ 为家族$F$ （Family）的频率在系统$L$ 中的占比，$H(F|L)$ 是该系统$L$ 中家族$F$ 的信息量。

对于拼音化系统，其系统信息量由各拼音符号的信息量的加权平均值计算而得，见公式（4）：

[1]为便于区分，本文规定：凡指“同音字家族”者，使用“//”括起音节，并采用数字标注声调（如“/hai2/”）；凡指拼音符号者，不加“//”，声调使用拼音符号（如“hái”）。

$$H(R) = \sum P(S|R) \cdot H(S|R) \tag{4}$$

其中，$H(R)$ 为拼音化系统$R$（Romanized）的信息量，$P(S|R)$ 为拼音符号$S$ （Symbol）的使用频率在系统$R$ 中的占比。

各系统的易学性通过系统大小的反向编码表示，即系统越大，其易学性越低 (Kemp and Regier, 2012)。

对于真实的同音字家族系统及1000 个计算模拟系统，系统大小为各家族大小的加权平均值，见公式（ 5）：

$$N(L) = \sum \frac{n(F|L)}{\sum n(F|L)} \cdot n(F|L) \tag{5}$$

其中，$N(L)$ 为系统$L$ 的大小，$n(F|L)$ 为系统$L$ 中同音字家族$F$ 的家族大小，$\sum n(F|L)$ 为系统$L$ 中各家族大小之和（即系统的总汉字数：10333）。

本研究采用加权平均值而非算术平均值来衡量系统大小，主要基于以下两个考虑：（1）系统内部的家族大小不同，其在构成系统总体规模中的权重不同，加权平均值能够更准确地反映这种结构差异；（2）实操上，真实的同音字家族系统与计算模拟系统在家族总数与汉字总数上一致，采用算术平均值将难以区分两类系统的实际规模差异。

对于拼音化系统，由于系统中每一音节仅对应一个拼音符号，因此其系统大小为1。

### 4.3 数据分析

首先采用单样本$Z$ 检验，比较真实的同音字家族系统与计算模拟系统在效率与易学性指标上的差异 (Kemp and Regier, 2012; Regier et al., 2015)。由于研究假设明确预期真实系统的效率高于模拟系统、易学性低于模拟系统，故采用单尾的单样本$Z$ 检验。随后，进一步比较真实的同音字家族系统与拼音化汉字系统在效率与易学性方面的不同。

### 4.4 结果

结果显示，真实的同音字家族系统信息量为0.93，而计算模拟系统的信息量平均值为3.03（$SD < 0.01$）；同时，前者的系统大小为16.79，而后者的系统大小平均值为8.58（$SD = 0.03$）。单样本$Z$ 检验结果表明，与计算模拟系统相比，真实系统的信息量显著更小（$Z = -673.44$，$p < 0.01$；图 1 左），但系统显著更大（$Z = 251.04$，$p < 0.01$；图 1 右）。这些结果说明，真实系统虽然易学性更低，但效率却更高。

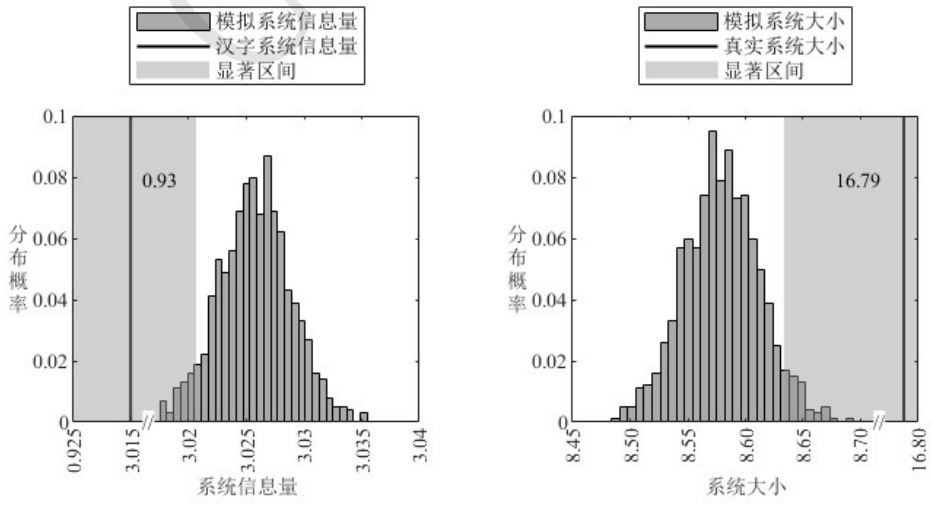


Figure 1: 同音字家族系统与计算模拟系统的信息量（左）与系统大小（右）对比

进一步比较发现，拼音化系统的信息量为8.54，远高于真实的同音字家族系统（0.93）；其系统大小为1，远小于汉字形式的系统（16.79）。这一结果表明，拼音化系统在易学性上具有优势，但在效率上则处于劣势。

综上所述，无论是与计算模拟系统还是与拼音化系统相比，真实的同音字家族系统在效率与易学性之间均表现出权衡关系：系统效率更高，易学性相对较低。该结果支持了假设（2），表明汉语的同音字系统在演化过程中可能受到了效率与易学性的双重约束，并在二者之间实现了结构性平衡。

## 5 验证性分析：带调/不带调音节、整字/声符、有/无生僻字

为检验第 3节与第 4节中效率与易学性统计依赖性的稳健性，本节从系统构建角度出发，考察三类潜在的影响因素，并在原有分析框架上构建对比系统进行验证性分析。具体而言，这些因素分别涉及音节颗粒度、正字法层级与字表覆盖范围，分别对应如下三种系统变体。

首先，在音节维度上，汉语单音节系统可细分为携带声调的“带调音节系统”与去除声调的“不带调音节系统”，两者对应不同的语音颗粒度，可能影响同音字家族的结构划分 (Chao et al., 2019; Lee et al., 2015)。其次，在正字法维度上，同音字家族既可在整字水平上构建，也可在亚字水平（如声符层级）展开。声符作为汉字中主要的语音提示部分，其结构性或许对同音字的分布与加工存在调节作用。最后，语言使用者的心理词典结构会受到实际语料暴露的影响 (Bogaerts et al., 2022)，尤其是否接触生僻字可能影响其对同音字家族的认知表征。因此，我们另构建一个排除生僻字的系统，以检验字表范围对研究结果的可能影响。

综上，本节将基于三种系统构建方式：（1）不带调音节系统，通过合并带调系统中所有同声同韵但声调不同的同音字家族构建而成；（2）声符水平的家族系统，其中汉字声符依据《汉字声旁读音便查》 (周有光, 1980) 标注；（3）不包含生僻字（即字频等于1的汉字）的系统，分别重复前述核心分析，检验效率与易学性关系的跨系统稳健性。各系统的构建方法和系统规模等参数可见：https://osf.io/96j4a/。

对每个类型的系统进行验证性。分析步骤重复了第3节与第4节中的所有分析步骤，包括相关分析及其$Z$ 检验、1000个计算模拟系统构建、以及真实系统与模拟系统之间的单样本$Z$ 检验。出于简洁，本节未再纳入拼音化系统的比较。

结果显示，无论在带调或不带调音节层面，在整字或声符分析水平上，亦或在包含或剔除生僻字的条件下，同音字家族系统在效率与易学性之间均呈现出与第3节和第4节一致的统计特征，即系统内部，家族效率越高，则家族越小，显示出家族效率对家族大小的制约（表 2）；相比于计算模拟系统，真实系统的易学性显著更低但效率显著更高，表现出对效率与易学性的权衡（表 3）。这些发现进一步支持了效率与易学性之间的统计依赖关系在同音字家族系统中的稳健性与普适性。

| 家族类型 | 控制变量 | 计算变量 | 偏相关系数 |
|---|---|---|---|
| 不带调+汉字+含生僻字 | 家族频率 | 家族信息量，家族大小 | 0.80** |
| | 家族信息量 | 家族频率，家族大小 | 0.71** |
| 带调+声符+含生僻字 | 家族频率 | 家族信息量，家族大小 | 0.84** |
| | 家族信息量 | 家族频率，家族大小 | 0.53** |
| 不带调+声符+含生僻字 | 家族频率 | 家族信息量，家族大小 | 0.78** |
| | 家族信息量 | 家族频率，家族大小 | 0.60** |
| 带调+汉字+不含生僻字 | 家族频率 | 家族信息量，家族大小 | 0.91** |
| | 家族信息量 | 家族频率，家族大小 | 0.54** |
| 不带调+汉字+不含生僻字 | 家族频率 | 家族信息量，家族大小 | 0.85** |
| | 家族信息量 | 家族频率，家族大小 | 0.73** |

Table 2: 各家族类型下同音字家族的家族大小与家族信息量、家族频率的偏相关系数

## 6 讨论

本研究系统考察了效率与易学性对现代汉语同音字家族系统的认知约束，结果发现，无论是在系统内部的家族单元层面，还是在与其他系统的整体比较层面，均观察到了显著的统计依

| 家族类型 | 指标 | 真实系统值 | 计算模拟系统值[a] | $Z$值 |
|---|---|---|---|---|
| 不带调+汉字+含生僻字 | 系统信息量 | 1.91 | 4.59（< 0.01） | -1311.77** |
| | 系统大小 | 44.64 | 24.55（0.07） | 304.38** |
| 带调+声符+含生僻字 | 系统信息量 | 0.78 | 1.87（0.03） | -37.18** |
| | 系统大小 | 8.09 | 4.80（0.09） | 38.38** |
| 不带调+声符+含生僻字 | 系统信息量 | 1.61 | 3.20（0.06） | -28.80** |
| | 系统大小 | 19.05 | 12.78（0.35） | 17.86** |
| 带调+汉字+不含生僻字 | 系统信息量 | 0.90 | 2.01（< 0.01） | -195.96** |
| | 系统大小 | 6.74 | 4.38（0.03） | 76.86** |
| 不带调+汉字+不含生僻字 | 系统信息量 | 1.87 | 3.46（< 0.01） | -362.77** |
| | 系统大小 | 18.27 | 11.48（0.06） | 107.01** |

[a] 1000个系统的平均值，括号内为标准差。

Table 3: 各家族类型下真实同音字家族系统与计算模拟系统的信息量与大小差异

赖关系。以下将围绕主要研究发现展开讨论。

### 6.1 效率与易学性对同音字家族的认知约束

本研究发现，无论是在同音字家族单元层面，还是在整个同音字家族系统层面，均观察到效率与易学性的认知约束。这一认知约束在口语系统中已有较为充分的证据支持 (Piantadosi et al., 2012; Koplenig et al., 2017; Kemp et al., 2018)，但针对文字系统的实证研究仍相对有限。当前研究将效率与易学性的认知约束扩展至汉语中的语音—字形映射系统，拓宽了该认知约束原则的适用范围。

首先，在系统内部层面，同音字家族单元的效率与易学性呈显著正相关，即效率越高的家族单元，其易学性越高（即家族规模越小）。从信息论看，效率本质是信息的可预测性。已有研究表明，语言单元在可预测性压力下会受到压缩，呈现出更简约的形式，如口语研究发现，具有较高可预测性的音节往往发音时长更短 (Jurafsky et al., 2001; Aylett and Turk, 2004)。本研究在汉语同音字家族系统中也观察到类似规律：可预测性较高的汉语音节往往对应数量较少的同音字。这一发现表明，同音字家族单元的规模可能也受到信息压缩机制的约束。这一压缩机制对于语言认知具有重要意义。一方面，较小规模的同音字家族有助于降低认知负荷，语言使用者在处理同音解歧方面所需的资源减少，从而更高效地提取和理解单音节所指代的具体字形；另一方面，高效率音节本身提供的信息较为集中明确，其对应的字形数量较少，使学习者能够凭借较少的输入线索，快速、准确地掌握语音到字形的映射关系 (Mahowald et al., 2018)。

值得注意的是，本研究还发现，家族效率与家族大小之间的偏相关系数显著高于家族频率与家族大小之间的偏相关系数。这一结果表明，同音字家族单元所受的认知压力更可能源自效率而非频率。这一发现与Piantadosi et al. (2011)的研究相符，即语言系统中的最优编码机制倾向于压缩最具可预测性的单元，而非压缩最常用的单元。

综上可见，家族单元层面的效率—易学性关系反映出一种信息压缩机制，该机制对高可预测性同音字家族进行结构简化，从而在语音到字形的映射过程中实现更高效的信息传递，并促进同音字在心理词典中的最优编码。

其次，在整体系统层面，汉字形式的同音字家族系统表现出效率与易学性之间的权衡关系。与计算模拟系统和拼音化系统相比，真实的同音字家族系统虽然在易学性上表现不高，但其效率显著更高。效率与易学性的权衡被认为是语言系统演化中普遍存在的认知原则 (Kemp and Regier, 2012; Kemp et al., 2018)，本研究表明该原则同样适用于汉语语音与文字的映射系统。值得强调的是，不同语言系统在效率与易学性权衡中的具体表现可能存在差异：某些语言可能追求极高的效率而牺牲一定程度的易学性，而另一些语言则可能趋向结构简单、易于习得 (Kemp et al., 2018)。当前研究中，真实的同音字家族系统效率高于计算模拟系统，易学性低于计算模拟系统，显示出其更倾向于优化效率，而这以一定学习难度为代价。至于为何汉语系统更偏向于提升效率而非易学性，可能与汉字自身的历史演化路径与文化功能相关。例如，汉字系统长期承担记录文化、区分词义的精细功能，这可能促使其在效率上不断优化，即使这意味着更复杂的学习负担。这一假设值得在未来研究中进一步探讨。

### 6.2 同音字家族系统规模的认知解释

为何现代汉语的同音字家族系统具有如此庞大的规模？从效率与易学性的角度来看，这一现象体现了认知约束对汉语文字系统中语音—正字法映射的深刻影响。

一方面，同音字家族系统在整体层面需在效率与易学性之间实现权衡。如前所述，现代汉语同音字家族系统表现出以提高系统效率为优先目标的倾向，往往以牺牲易学性为代价来提升系统中音节信息的可预测性。在语言历史演化过程中，汉语使用者倾向于通过造字方式将新出现的概念纳入新的汉字编码之中，而非引入新的音节形式。这种做法在提升系统效率的同时，也增加了系统的结构复杂度与学习负担。此外，人类认知系统具有“连接偏好”（preferential attachment）倾向，即更倾向于将新加入的节点连接至已有较多连接的节点上 (Kello et al., 2010)。这一机制在汉字系统中可能表现为新造汉字更易被归入已有较大规模的同音字家族，从而进一步推动同音字家族系统的整体规模不断扩大。这一趋势可从形声字的构造规律中获得支持。文字学研究表明，指称的新义的新造字更倾向于归入已有音节。这一偏好源于形声构造在造字中的广泛使用，其中声符不仅指示读音，往往亦承载一定的语义信息。因此，许多新造字在语义上继承声符所表达的意义，在语音上则复用其读音，从而推动同音字家族的扩展 (周克庸, 2009)。例如，“仑”本义为竹简编册，引申为“条理”“有序”之义，后续造字也蕴含此义，如“论”（有“理”之言）、“伦”（人际“秩序”）、“轮”（有“辐条”的车部件）、“纶”（“理”丝）与“沦”（水之“纹理”）等，它们也在古汉语中保持相同或相近的读音 (周克庸, 2009)。

另一方面，在系统内部，效率对语言单元大小存在压缩效应，即高效单元因其高度可预测性而更易受到压缩，进而形成更小规模、易于学习的语言结构 (Piantadosi et al., 2011)。同音字家族系统亦体现出该规律：效率较高的家族单位承受更强的信息压缩压力，其规模更小；而效率较低的家族单位则因压缩压力较小，往往形成更庞大的结构。

具体而言，如图 2 所示，若以所有家族信息量的算术平均值（0.91）为界，将信息量低于该值的视为高效家族，信息量高于该值的视为低效家族，则现代汉语中共有697 个高效家族（平均信息量为0.25，平均家族大小为3.83），其中607 个家族的大小低于全体家族的平均大小（7.71）。相对地，共有643 个低效家族（平均信息量为1.62，平均家族大小为11.92），其中402 个家族的大小高于总体平均值。从总量上看，高效家族的规模总体较小，累计仅包含2671 个汉字；而低效家族的总体规模很大，累计汉字数达7662。可见，在同音字家族系统内部，效率对家族大小的压缩作用主要集中施加于高效家族，而低效家族则因缺乏压缩压力而更易膨胀（可能在历史演化中更易与新出现的汉字产生连接）。这种结构上的不均衡反映出效率与易学性间的动态平衡，在保障系统效率的同时，也导致了家族系统总体规模的持续扩展。

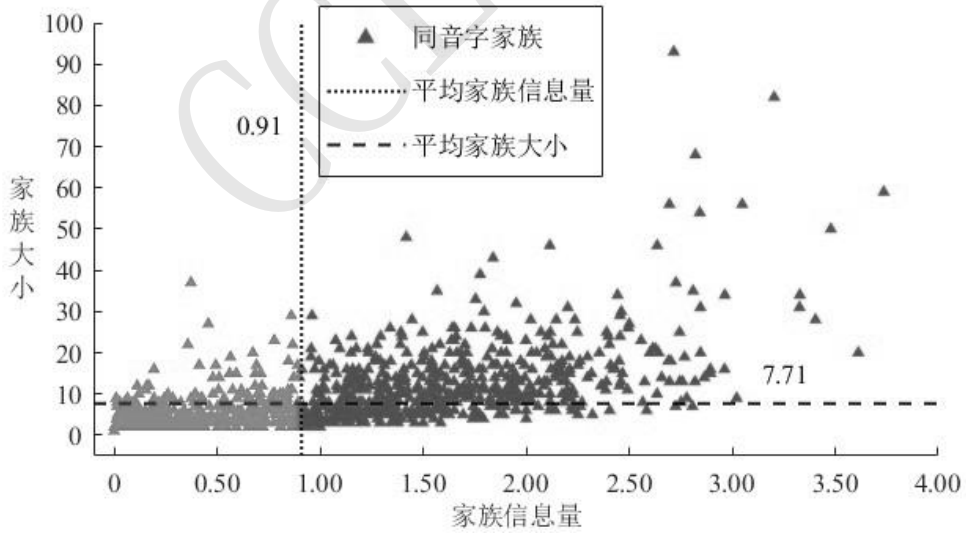


Figure 2: 同音字家族的家族信息量与家族大小分布

综上所述，现代汉语同音字家族系统并非汉字系统演化中的“缺陷”，而是效率与易学性这两种认知约束在语言系统内部单元与整体系统双重层面上共同作用的产物。其大规模的形成体现了语言系统在长期使用过程中对传递效率与认知负担的深层权衡与动态调节。

### 6.3 从效率与易学性角度看汉字拼音化问题

对汉语同音字现象的深入理解不仅有助于全面把握汉字系统的结构特征，也有助于为我们对语言文字战略的思考提供更多样的视角。例如，早在20世纪初，一些学者便认为同音字是汉字系统的重大缺陷，因而提出“汉字拼音化”主张，力图通过消除同音字现象来简化汉字体系 (李敏生, 1992)。这一观点在新中国成立初期曾产生较大影响 (李敏生, 1992; 王爱云, 2009)。尽管自1986年全国语言文字工作会议召开以来，拼音化的声音逐渐式微，但其所指涉的根本理论问题未得到完全回应。当前研究从人类普遍认知约束的角度出发，探讨同音字家族系统的形成机制，或可为汉字拼音化议题提供新的理论视角与解释路径。

结果表明，尽管拼音化系统的结构较为简洁、易于习得，但效率远低于同音字家族系统。这一发现与黎锦熙 (1949)的早期观察契合。黎氏曾在近三十年间尝试使用拼音化汉字，他指出，虽然拼音符号确实简化了书写，但阅读效率却大大下降，甚至减少三倍以上。可见，从效率与易学性的认知权衡出发，拼音化虽解决了部分易学性问题，却在效率层面明显不足。这可能正是历史上汉字拼音化改革难以持续推进的重要原因之一：若通过拼音化提升汉字易学性很可能削弱汉语文字系统的效率，降低汉字在实际语言使用中的功能性与实用性。

此外，关于汉字是否应当拼音化的另一关键问题，是现有同音字家族系统是否违背了人类认知的基本规律——即效率与易学性之间的认知约束。若系统设计偏离了这一规律，极有可能引发认知失衡甚至系统性崩溃。然而，本研究的结果显示，同音字家族系统内部与系统层面的结构分布均体现出效率与易学性之间的统计依赖性，该系统在现今的实际语言使用中仍在高效运作，这进一步佐证了其认知合理性。因此，当前的同音字家族系统并未违背人类的基本认知规律，而是体现出汉字系统伴随着汉语口语系统在长期演化中所达成的认知最优化结构。

综上所述，本研究从效率与易学性的角度出发，论证现代汉语同音字家族系统符合人类语言认知的基本约束机制，是一种具有高度结构效率的文字系统。

## 7 结论与未来研究方向

本研究从语言系统的效率与易学性出发，系统考察了现代汉语同音字家族系统所体现的认知约束机制。研究结果表明，现代汉语的同音字家族系统是在效率与易学性的共同作用下形成的，体现出二者间的动态平衡。这一发现表明，效率与易学性的认知约束不仅适用于口语系统，也存在于口语语音系统与书面文字系统的映射关系之中。尽管在人类语言文字体系中，汉字与拼音文字在形式、结构与编码方式上存在重大差异，且汉语的同音字现象具有高度的语言特异性，但本研究发现，从效率与易学性的视角出发，汉语同音字家族系统同样受到人类认知普遍规律的约束，具有一定的跨语言适用性与认知普遍性。这一结论为理解汉语文字系统的结构特征提供了新的理论认识，也为探讨人类语言系统演化中的认知机制贡献了来自汉语书写系统的证据。本研究亦对自然语言处理与国际中文教育等应用领域具有潜在启示。例如，在自然语言处理中，不同拼音对应的汉字分布在家族中的结构复杂度不同，本文所量化的效率与易学性指标可作为输入法候选排序的先验约束或加权因子。在国际中文教育方面，教师可根据汉字所在家族的效率与易学性特征进行分层教学，如优先教授高效率家族，重点讲解高效率但低易学的家族，针对低效率且低易学的家族则作为难点内容有针对性地强化练习。

本研究从效率与易学性的视角对汉语同音现象进行了初步探讨，为后续研究奠定了基础。首先，本研究聚焦于字水平的家族结构特征。考虑到单音节是现代汉语的核心语音单位 (李如龙, 2009)，而汉字在母语者心理词典中占据关键地位 (张玲燕等, 2013)，从字层面考察同音现象所体现的认知约束具有重要的理论与实证意义。然而，本研究未涉及同音现象在词水平的效率与易学性表现，亦未探讨语境因素对其的作用。鉴于汉字上下文和语境对汉字信息预测性与歧义解决机制具有重要影响，未来可进一步考察其在同音字系统构建与加工中的功能机制。其次，字义因素对同音字家族系统效率与易学性的潜在影响亦值得关注。尽管本研究引入了亚字水平的声符分析以控制语义干扰（因声符本身不表义），但字义在信息压缩与效率结构中的具体作用仍有待深入阐明。此外，由于声符在不同汉字中的示音能力（如提示整字音、部分音）及其与其他形符组合成字的多产性存在差异，其在调节同音系统效率与易学性之间权衡机制中的作用亦应成为未来关注的重点。最后，本研究基于共时视角揭示了当前同音字家族系统所体现的认知约束机制。未来研究可引入历时语言材料，探讨系统演化过程中的效率与易学性变迁，进一步揭示认知约束如何在语言系统的长期演变中持续发挥作用。这不仅有助于深化对效率—易学性认知机制的普遍理解，也有助于构建对汉语同音字系统发展规律的系统性解释。

第二十四届中国计算语言学大会论文集，第247页-第258页，济南，中国，2025年8月11日至14日。

## 参考文献

北京语言学院语言教学研究所（编）. 1986. 现代汉语频率词典. 北京语言学院出版社.

杜同惠. 1993. 留学生汉字书写差错规律试析. 世界汉语教学, (1):69–72.

黎锦熙. 1949. 国语新文字论. 师范大学, 北京.

李敏生. 1992. 废除汉字论和汉字落后论的由来及其理论基础. 汉字文化, (3):2–12.

李如龙. 2009. 汉语和汉字的互动与和谐发展. 吉林大学社会科学学报, 49(2):108–116+160.

柳英绿 和 关黑拽. 2014. 汉语拼音化运动的历史进程与现实困境. 吉林大学社会科学学报, 54(2):160–167+176.

孟祥芝 舒华 周晓林 和 罗晓辉. 2000. 不同阅读水平儿童的汉字字形输出与再认. 心理学报, 32(2):133–138.

邵敬敏. 2007. 现代汉语通论（第二版）. 上海教育出版社.

苏新春 和 林进展. 2006. 普通话音节数及载字量的统计分析——基于《现代汉语词典》注音材料. 中国语文, (3):274–284+288.

万业馨. 2021. 谈汉字形声化与汉语词汇双音化. 古汉语研究, (3):67–82+127.

王爱云. 2009. 中国共产党与新中国文字改革(1949—1958). 党史研究与教学, (6):11–24.

肖哲 徐彩华 和 邓娟. 2024. 现代汉语同音字家族属性的计量研究. 语言文字应用, (4):95–109.

尹文刚. 2003. 汉字同音率、同音度及同音字音节个数随同音度增加而递减的规律. 语言科学, (4):3–6.

杨仕章. 1998. 论谐音及其功能. 中国俄语教学, (3):31–36.

叶蜚声 和 徐通锵. 2010. 语言学纲要（修订版）. 北京大学出版社.

张学新. 2011. 汉字拼义理论: 心理学对汉字本质的新定性. 华南师范大学学报（社会科学版）, (4):5–13+160.

张玲燕 田朝霞 和 金檀. 2013. 汉语合成词的整词词形表征. 心理与行为研究, 11(4):569–574+576.

中国社会科学院语言研究所词典编辑室（编）. 2016. 现代汉语词典. 商务印书馆, 北京, 第7版.

周克庸. 2009. “会意兼形声”是拥有大量字例的重要汉字结构类型. 文史哲, (1):147–453.

周有光. 1980. 汉字声旁读音便查. 吉林人民出版社.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Louisa Bogaerts, Noam Siegelman, Morten H. Christiansen, and Ram Frost. 2022. Is there such a thing as a ‘good statistical learner’? *Trends in Cognitive Sciences*, 26(1):25–37.

Pei-Chun Chao, Wei-Fan Chen, and Chia-Ying Lee. 2019. The second-order effect of orthography-to-phonology mapping consistency on Chinese spoken word recognition. *Journal of Neurolinguistics*, 51:1–16.

Jeffrey A. Coady and Richard N. Aslin. 2004. Young children’s sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, 89(3):183–213.

Peter W. Culicover and Ray Jackendoff. 2012. Same-except: A domain-general cognitive relation and how language expresses it. *Language*, 88(2):305–340.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, Anne Christophe, and Steven T. Piantadosi. 2017. Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.

Francesca Franzon and Chiara Zanini. 2023. The entropy of morphological systems in natural languages is modulated by functional and semantic properties. *Journal of Quantitative Linguistics*, 30(1):42–66.

Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407.

Simon J. Greenhill. 2023. A shared foundation of language change. *Science*, 381(6656):374–375.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan L. Bybee and Paul J. Hopper, editors, *Frequency and the Emergence of Linguistic Structure*, pages 229–254. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Christopher T. Kello, Gordon D.A. Brown, Ramon Ferrer-i-Cancho, John G. Holden, Klaus Linkenkaer-Hansen, Theo Rhodes, and Guy C. Van Orden. 2010. Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, 14(5):223–232.

Charles Kemp and Terry Regier. 2012. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1):109–128.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLOS ONE*, 12(3):e0173614.

Chia-Ying Lee, Chia-Hao Hsu, Yi-Ning Chang, Wei-Fan Chen, and Pei-Chun Chao. 2015. The feedback consistency effect in Chinese character recognition: Evidence from a psycholinguistic norm. *Language and Linguistics*, 16(4):535–554.

Kyle Mahowald, Isabelle Dautriche, Edward Gibson, and Steven T. Piantadosi. 2018. Word forms are structured for efficient use. *Cognitive Science*, 42(8):3116–3134.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Terry Regier, Charles Kemp, and Paul Kay. 2015. Word meanings across languages support efficient communication. In Brian MacWhinney and William O'Grady, editors, *The Handbook of Language Emergence*, pages 237–263. John Wiley & Sons, Inc, Hoboken, NJ, USA.