从多语种词表推断部分共词化

Johann-Mattis List[1,2]

[1] 多语种计算语言学教席 / 帕绍大学

[2] 语言与文化演化系 / 莱比锡马克斯·普朗克进化人类学研究所

2023

arXiv:2302.00739v1 [cs.CL] 1 Feb 2023

过去几年，专门研究特定语言家族中的共词化模式，尤其是世界各地语言中的共词化模式的研究大幅增加。具体而言，计算研究得益于这样一个事实：共词化作为一种科学构念，易于操作化，使学者能够从大量跨语言数据中推断共词化模式。然而，迄今为止，专门研究部分共词化——即不涉及完整词语，而是涉及词语各个部分的共词化模式——的工作仍然很少。这并不令人意外，因为部分共词化在计算方法中较难处理，而且很容易受到各种噪声的影响，例如假阳性匹配。为了解决这一问题，本研究提出了处理部分共词化的新方法：（1）提出新的模型，用以表示部分共词化模式；（2）开发高效的新方法和工作流程，以帮助从多语言词表中推断各种类型的部分共词化模式；（3）展示如何对推断出的部分共词化模式进行计算分析和交互式可视化。

关键词： 部分共词化、宽松共词化、共词化网络、计算比较语言学

1 引言

近年来，专门研究单个语系及世界各地语言中共词化模式的研究大幅增加。共词化 这一概念已被证明特别适用于词汇类型学中的计算方法和定量方法。术语 共词化 最初由 Fran.cois（2008）提出，作为一个统称，涵盖多个义项由同一词形表达的所有情况，

无论多个义项是源于多义还是同音异义。共词化可以从大量词汇数据中轻松计算得出，尤其是从多语言词表中；这类词表将一定数量的概念翻译成多种目标语言（见 List 2014，22-24）。通过汇集多个多语言词表，可以轻松整理出大量跨语言共词化数据；跨语言共词化数据库（CLICS，List 等，2018；Rzymski 等，2020，https://clics.clld.org）近期版本的增长，以及利用不同来源汇编的共词化数据开展的研究不断增加，均可证明这一点（Di Natale 等，2021；Bao 等，2021）。关于共词化模式的定量研究还表明，通过查找在多个语系中反复出现的共词化，可以轻松提取最有可能源于多义的共词化；与之相对的是，在同一语系内部频繁出现的共词化，这类现象可能反映了广泛存在的同音异义（List 等人，2013）。这也意味着，大型共词化网络可以被视为多义网络，从而让我们直接洞察词汇语义学的某些方面（Youn 等，2016；Jackson 等，2019；Harvill 等，2022）。

然而，迄今为止，大多数研究共词化的研究都聚焦于整个词的共词化。仅涉及某种语言中词语特定部分的共词化——部分共词化，也称为松散共词化 （Fran.cois，2008）——鲜少有人研究（例外情况见 Urban，2011），也很少有人尝试从大型跨语言数据集自动计算这类共词化（初步尝试见 List 等，2022）。

部分共词化或松散共词化研究面临的问题似乎主要由两个因素造成。一方面，在网络中对部分共词化进行建模并不那么直接，因为共享相同部分的词语之间的关系有时可能是不对称的：一个词可能完整地出现在另一个词中。对部分共词化网络进行建模不仅需要不同的网络类型，而且这类网络也更难解释。另一方面，从大型跨语言数据集中推断部分共词化网络并不容易，因为词语之间的部分相似性很容易偶然出现，也可能反映语法区别（名词类别、性别标记、词性）。因此，在大型语料库中，对同一语言变体内的词语进行相似性搜索这一简单方法，通常会产生连接极为密集的噪声网络，其中几乎找不到任何具有语义或认知视角研究价值的信号。因此，尽管在完全共词化网络中，可以通过对特定共词化的出现频率采用严格阈值并结合归一化权重，轻松处理同音异义造成的噪声，但在创建部分共词化网络时，却很难采用同样的标准。

本研究试图通过提出新模型来解决其中至少一部分问题，使某些类型的部分共词化模式能够以网络形式表示；同时开发新的高效方法和工作流程，以帮助从多语言词表中推断不同类型的部分共词化模式。在推断出这些模式后，本研究还进一步展示了如何对其进行可视化和分析。

2 材料与设备

2.1 多语言词表

我们用于推断部分共词化的新工作流程以多语言词表为起点。这里的词表指的是按含义编排的一组词形。词典以词形（词头）为主要语言单位，并据此排列数据；与此不同，词表按含义排列词语。词典从形式出发，遵循语义学视角或基于形式的视角；词表则从含义出发，遵循命名学视角或基于概念的视角。因此，多语言词表使我们能够比较某些概念 （人们认为这些概念通常可以跨语言进行比较，尽管在实践中这可能存在问题）如何被译成特定语言。

过去十年间，多语言词表的编纂与汇集取得了显著进展，数字化词表集的数量也在不断增加。一方面，近年来已有研究提出大型的统一多语言词表集（Key and Comrie，2016；Dellert 等人，2020；Haspelmath and Tadmor，2009）；另一方面，跨语言数据格式的标准不断完善（Forkel 等人，2018），并应用于许多规模较小或不断扩大的数据集（Ferraz Gerardi 等人，2021），以及追溯标准化的工作追溯标准化 （Geisler 等人，2021）。

2.2 跨语言数据格式

为了探索多种语言中的部分共词化模式，我们整理了一个广为人知的洲际词典系列 (IDS)的修订版（Key and Comrie，2016）。原版将语音转写、各语言特有的音系转写和正字法词条混杂在一起，而修订版中的词条则通过半自动方式转换为国际音标，并采用跨语言转写系统（CLTS）参考目录提出的变体（https://clts.clld.org，List 等人，2023；另见 Anderson 等人，2018）。这一转换是通过应用 Lexibank 工作流完成的，该工作流旨在以跨语言数据格式（List 等人，2022）创建标准化词表。在这一工作流中，原本未经标准化的数据集通过结合软件工具（基于 CLDFBench，Forkel 和 List，2020）与人工标注，以半自动方式进行标准化，从而将数据转换为跨语言数据格式倡议（Forkel 等人，2018）推荐的格式。

IDS 的更新版本提供了 329 种语言变体的词表，涵盖最多 1310 个概念。标准化的语音转写总共包含 558 种不同的语音（类型），这些语音在数据中出现了 2 902 306 次（词例）；每种变体的音位库存平均包含 50.76 个语音。严格来说，虽然理论上也可以从正字法数据中识别部分共词化，但即使转写中可能存在某些错误，能够使用规模更大的语音转写多语言词表仍有两大优势。一方面，

如果针对同一数据集对转写进行统一，研究结果就更容易评估；另一方面，既然语音是以分段形式表示的，就更容易选定阈值，据此初步判定是否接受或舍弃部分共词化。《洲际词典系列》的修订版目前由 GitHub 管理，可通过 https: //github.com/intercontinental-dictionary-series/ids-segmented 访问。本研究使用的版本为 v0.2（ https://github.com/intercontinental-dictionary-series/idssegmented/tree/0.1）。

出于开发目的，并为了测试本文提出的新方法的某些技术方面，我们使用了 Allen（2007）编制的一份较小词表。这份词表也已转换为跨语言数据格式（参见 https: //github.com/lexibank/allenbai），提供了9种白语方言变体的标准化语音转写数据。

3 方法

跨语言的完全共词化可以通过一种高效的方法进行处理，这种方法已被证明能够为语义关系研究带来非常有趣的见解。然而，部分（或宽松）共词化存在噪声，因为同一语言中词语之间的部分相似性可能由许多因素造成（如偶合、语法标记），而这些因素并不能反映相关词语之间特定的语义关系。因此，用于推断完全共词化网络的成熟工作流程无法用于推断部分共词化网络。为了解决这一问题，我提出了一种三阶段方法：首先对部分共词化进行建模，以缩小搜索空间，并为网络中不同类型的部分共词化提供一致的表示；然后提供高效的方法来推断 特定类型的部分共词化；最后使我们能够以多种方式分析 不同种类的部分共词化网络。在此背景下，建模、推断和分析体现了历史科学中解决科学问题的一般方法，这一方法的灵感来自其在进化生物学中的应用（Dehmer 等，2011）。

3.1 跨语言的部分共词化建模

3.1.1 部分共词化的主要类型

将词建模为语音序列时，我们可以用形式化的方式定义主要的序列关系。由于序列在许多科学领域中都发挥着关键作用——从计算机科学（Gusfield，1997年），到生物信息学（Durbin 等， 2002[1998]），再到物理学（Kruskal 和 Liberman， 1999[1983]）——序列之间的基本关系很早以前就已被独立识别并加以讨论。下文将区分部分共词化 与宽松共词化 （后者最初由 Franc̦ois 于 2008 年提出）。根据这一区分，部分共词化仅限于串联形态，而宽松共词化也允许非串联（范式）形态。较狭义的部分共词化概念有一项优势：我们可以利用既有模型以及早期关于序列关系和字符串关系的研究成果，并将其应用于部分共词化这一概念

部分共词化的

比较三个虚构序列 ABC、XYABCD 和 ZABCEF 时，很容易看出，第一个序列 ABC 在第二个和第三个序列中都会出现。在计算机科学和生物信息学中，ABC 被称为 XYABCD 和 ZABCEF 的公共子串。由于不存在比 ABC 更长的子串，因此它还是这两个序列之间的最长公共子串。就 XYABCD 和 ZABCEF 之间的具体关系而言，我们可以说，它们共享一个长度为 3 的公共子串。序列 ABC 与另外两个序列 ZABCEF 和 XYABCD 也共享长度为 3 的子串。此外，我们还可以看到，序列 ABC 与该公共子串相同，因此可以说 ABC 是 XYABCD 和 ZABCEF 的一部分。序列之间的前一种关系（共享给定长度的子串）具有对称性，而后一种关系则不具有对称性：说序列 A 是序列 B 的一部分，并不等同于说序列 B 是序列 A 的一部分。

鉴于它们在广泛的科学和工业应用中的重要性，人们提出了许多高效算法，用于计算序列中的公共子串以及识别部分关系（或称部分关系，参见 Hoehndorf 等，2009年）（综述参见 Gusfield，1997年，第89—121页）。公共子串和部分关系是序列之间的两种基本关系，在一定程度上涵盖了部分共词化这一概念。公共子串和部分关系与部分共词化概念之间的确切关系，取决于我们如何定义后者。如果我们坚持认为，同一语言中不同词之间共享的材料应当体现词汇语素 ——也就是说，某种语言中最小的形式—意义配对，且具有词汇意义——那么可以说，同一语言中两种词形之间的任何部分共词化实例，也对应着表示这两种词形的音序列之间的公共子串关系。然而，如果两个音序列之间存在公共子串关系，并不意味着它们必然构成这种较狭义意义上的部分共词化，因为词形可能出于其他原因而具有公共子串。公共子串可能体现的纯粹是语法语素（各种形式的词缀；例如德语动词都带有相同的不定式词尾，如 *lauf-en* “跑”、*geh-en* “走”等），也可能只是偶然相似（例如德语 *Herbst* “秋天”与 *Wurst* “香肠”共享公共子串 *st* ）。

虽然学者们似乎会不言自明地使用术语 宽松共词化，将这种关系限定为同一语言中不同词语共享词汇语素（Fran.cois，2008），但需要指出的是，实际上并不必采用如此狭窄的定义。选择“共词化”这一术语，原本就是为了描述意义不同的词语共享相同发音的关系。因此，诸如部分共词化或宽松共词化等术语，也应以中性、概括的方式使用；至于语义上更有趣的关系，则应在第二步中从部分共词化模式中推断得出。

3.1.2 词缀与重叠共词化

在尝试开发能够寻找有意义的部分共词化的方法时，也就是说，寻找能够反映某种语言中词语形成背后相似词汇动因过程（Koch，2001）的部分共词化，采用比前文所述序列间公共子串和部分关系所体现的相似性更狭义的共享相似性概念会很有帮助。因此，我们不必搜索表达概念 X 的词形 A 是否属于表达概念 Y 的词形 B，而可以询问 A 是否为 B 的前缀或后缀，也就是说，A 是否与 B 的开头或结尾相同。同样，我们可以询问 A 和 B 是否重叠，即它们是否共享一个公共子串，而该子串既是两个序列的前缀，也是它们的后缀。

通过设定共词化子串长度的阈值，以及设定序列中未共词化的剩余部分长度的阈值，可以进一步缩小对语言学相关的词缀共词化和重叠共词化的搜索范围。此处，多语言词表现已采用完全切分且标准化的语音转写形式，这一特点派上了用场，因为它使我们能够按一定数量的语音而非一定数量的符号来设定阈值；后者往往只有组合起来才能表示单个语音。

3.1.3 在网络中表示部分共词化

与子串共词化和部分共词化类似，词缀共词化与重叠共词化的根本区别在于：前者包含一种有向关系（在此特定情形下，一个序列是另一个序列的一部分，出现在其开头或结尾），而后者（两个序列共享一个部分）的方向无法直接确定。在网络中表示词缀共词化时，我们可以使用有向网络模型来体现这种关系的方向性。

与众所周知、用于表示完全共词化网络且易于可视化的无向加权网络模型（List 等人，2013）相比，这类网络既可以交互式地（Mayer 等人，2014）也可以静态地（通过边的粗细体现连接各个概念的连边权重差异，见 List 等，2018）进行可视化；而加权有向网络则仅在能够证实存在从概念 A 到概念 B 的词缀共词化时，才绘制一条从 A 指向 B 的连边。以德语为例，看看德语单词*Finger* “手指”和*Fingernagel* “指甲”。前一个词形是后一个词形的词缀（因为后一个词以它开头），因此我们可以在词缀共词化网络中绘制一条从“FINGER”概念指向“FINGERNAIL”概念的连边；这两个词在德语中都有实例。可视化这些网络时，连边可能会双向指向（因为某些语言中，“指甲”的词可能是“手指”一词的词缀，尽管我们预计这样的例子不会很多），而我们可以在连边末端使用箭头来指示网络中各条连边的方向。

重叠共词化 可以采用处理完全共词化的相同方式来处理 s.

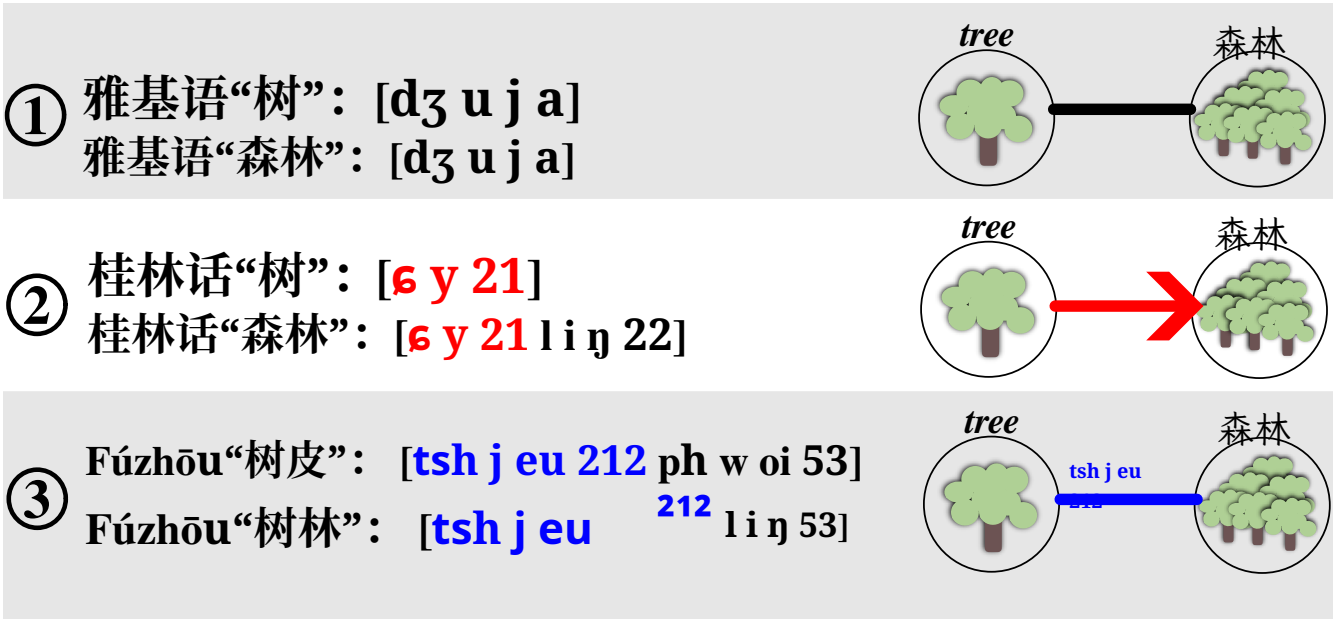

图1：本研究讨论的三种主要共词化类型概览。（1）以雅基语为例展示完全共词化（数据来自 CLICS3，Rzymski 等，2020）；（2）以 Gu˙ıl´ın 汉语为例展示词缀共词化的方向性表示（数据来自 Li´u L˙ıl˘i 等，2007）；（3）以 F´uzh¯ou 汉语为例展示重叠共词化（数据来自 Li´u L˙ıl˘i 等，2007）。

区别在于，应在内部存储构成重叠连接的各个后缀。因此，在完全共词化网络中，共词化只需存储一种词形（因为按照定义，共词化两个或更多概念的词语彼此相同）；而在重叠共词化网络中，为了保证透明度，标明两个词语中重复出现的实际后缀则十分重要。图1对比了三种类型的共词化网络，并列举了一些简化示例。

3.2 完全与部分共词化网络的高效推断

一种朴素的实现方式是：对部分共词化（无论是词缀共词化还是重叠共词化）进行简单搜索，取出某种语言中的所有词形，再将样本中的每个词与其他所有词逐一比较，并记录观察到的共性。虽然这一过程易于理解，也确实能得到所需结果，但效率远远不够。因此，尤其是在处理大型跨语言数据集时，建议采用高效的搜索策略。

对于完全共词化的计算，一种高效的搜索策略是将关联数组 用作主要数据结构。关联数组由一个键组成，可用于访问给定的值。在 CLICS 数据库采用的 Python 实现中（List 等人，2018；Rzymski 等人，2020），键是某种语言中的各个词形，而值则是一个概念列表，这些概念是该词形

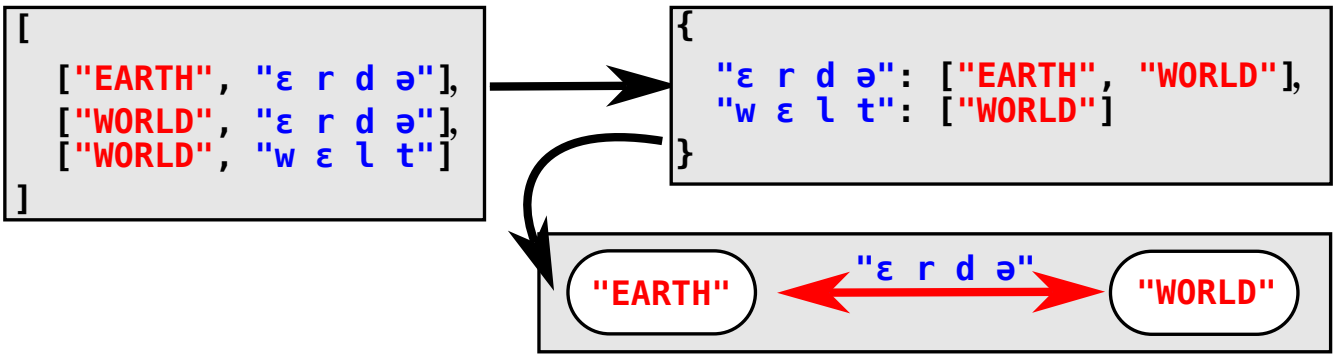

图2：使用关联数组高效搜索完全共词化。数据以 JSON 格式表示，包含三个德语词形的词表：*Erde* “地球”、*Erde* “世界”以及 *Welt* “世界”。左上方的方框展示了数据的初始格式（一个包含两列的词表，一列存储概念，另一列存储以国际音标表示的词形）。右上方的方框展示了生成的关联数组，其中词形用作键，表达该词形的概念则作为值添加到同一个数组中。右下方的方框展示了根据此示例推断出的共词化。

所关联的概念。为了推断某种语言中的共词化，该方法会遍历词表中该语言的所有词，随后将它们添加到关联数组中，并把词形所表达的概念存储在作为值的列表里。如果某个词形已经添加到数组中，就会将相关列表扩展，加入相应的新概念。第二阶段，该方法遍历关联数组中的所有键，并将列表中所有彼此不同的概念对添加到逐渐构建的跨语言共词化网络中。Jackson 等（2022）附带的教程以及 List（2022）中对这一过程有详细说明。图2展示了将此方法应用于包含三个德语词的小型词表后所得到的结构。

在我们处理部分共词化的方法中，我们采用类似的方式，对每种语言的词表各遍历两次。不过，为了寻找词缀共词化，我们会在关联数组中填入不同长度的词缀，而作为值的列表则填入相应完整词形及其概念组成的元组。我们通过从表示词形的音序列的左侧和右侧遍历来计算词缀长度。词缀长度受两个阈值限制。第一个阈值（默认设为 2）将词缀的最小长度限制为 3 个语音。第二个阈值确保剩余词语部分的长度大于某个最小值（默认设为 2）。两个阈值结合起来，既能保证推断出的词缀长度足够合理，也能确保与之关联的完整词形足够长，从而提高我们检测到复合结构而非屈折形式的可能性。借助这些阈值，我们可以在第一轮遍历中检测给定词的所有潜在词缀候选项，并将其存入关联数组。第二步，我们按长度对数据中的所有原始词语排序，从最长的词语开始依次遍历。接着，对于每个词形，我们检查它是否出现在词缀候选项数组中。如果是，这意味着该词作为词缀出现在某个词形中，而该词形已作为值关联到这个数组；因此我们可以

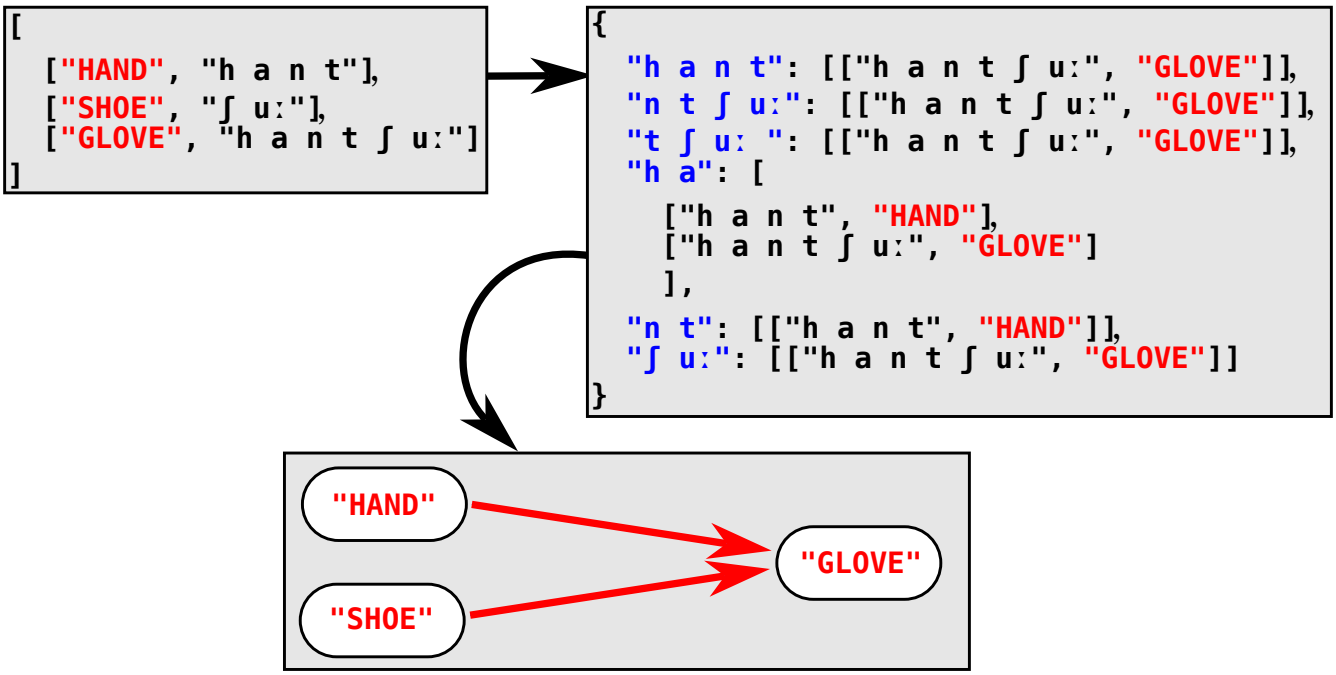

图3：以三个德语词组成的词表为例，展示如何高效搜索词缀共词化：*Hand*“手”、*Schuh* “鞋”和*Handschuh* “手套”（字面意思为“手-鞋”）。从左上方方框中的原始词表出发，在右上方方框的关联数组中，每个词形都由所有符合双阈值标准（见正文）的可能前缀和后缀表示。在遍历原始概念列表中的词形时，我们发现有两个词*Hand* 和*Schuh* 已存储在数组中，因此可以推断这两个词与*Handschuh*之间存在词缀关系，并在图底部的方框中以有向图表示。

将它们添加到网络中，方法是从作为词缀重复出现的词语向包含该词缀的词语添加一条连接。

在计算重叠共词化时，第一阶段采用与词缀共词化相同的策略：在关联数组中填入词表中各词形的候选词缀。由于搜索重叠共词化时噪声增加，词缀长度的默认阈值设为4，剩余部分长度的阈值设为3。第二阶段，我们遍历候选词缀数组本身；该数组按作为键的词缀长度降序排列（从给定语言数据中发现的最长词缀开始）。对于每个词缀，我们比较所有包含该词缀的词对，并检查其中任一形式都不是另一形式的后缀或前缀。如果满足这些条件，且两种形式也不完全相同（完全相同则对应完全共词化），我们就将这些形式连同使其发生重叠的词缀一起记录为重叠共词化。

从描述中可以很容易看出，用于推断完全共词化、词缀共词化和重叠共词化的三种方法，其复杂度各不相同。搜索完全共词化

所需的计算时间最少，其次是搜索词缀共词化，再其次是搜索重叠共词化。

利用上述针对单个语言中完整共词化和部分共词化的推断方法，我们可以将搜索策略应用于多种语言，并通过迭代扩展共词化网络，构建完整共词化和部分共词化网络：当推断出某种语言中新的边时，我们就添加这些边；如果迭代过程中已发现这些边，则增加其边权重。以这种方式计算出的网络都以多种方式进行了标注。对于节点，我们记录数据中可找到的词形数量、这些词语所属的语系数量，以及各语言中的实际词形。对于节点之间的连边，我们记录体现共词化关系的具体词形数量、能够找到这些共词化关系的语系数量，以及发生共词化的实际词形（包括部分共词化中发生共词化的部分）。对于词缀共词化，我们推断出一个有向网络；而完整共词化和重叠共词化对应的网络则是无向的。

3.3 分析部分共词化网络

为了理解完全共词化网络与本文引入的两种新类型网络之间的主要差异，可以比较它们的度分布。网络中节点的度是其边的数量（Newman，2010：133–135）。网络中节点的加权度是其各条边的边权重之和。对于无向网络，我们只有一种度；而对于具有有向边的网络，则有两种可能的度，即入度 和出度。前者表示给定节点的入边数量（或边权重之和），后者表示网络中给定节点的出边数量（或边权重之和）。为了比较由同一组节点构建的两个网络的度分布，我们可以计算斯皮尔曼秩相关（Spearman，1904）。该指标可以告诉我们：在一个网络中度非常高的节点，在另一个网络中是否也具有较高的度；以及在一个网络中度较低的节点，在另一个网络中是否也往往具有较低的度。

除了比较度分布之外，对网络进行可视化并放大查看能够体现主要差异所在的有趣部分也很有帮助。借助 Gephi（Bastian等人，2009年）或 Cytoscape（Shannon等人，2003年；Smoot等人，2011年）等网络可视化软件包，现在可以相当便捷地完成这一操作。本文所报告的可视化结果均使用 Cytoscape。

3.4 实现

本文报告的方法以 Python 实现，并以 Python 库的形式共享，可作为插件用于 CL Toolkit 软件包（https://pypi.org/project/cltoolkit，List 和 Forkel 2021年）。CL Toolkit 旨在让用户能够通过 Python 脚本或 Python 交互式控制台，便捷地访问符合

Lexibank 语料库所提出标准的 CLDF 词表（List 等人，2022年）。图的处理使用了 NetworkX 软件包（Hagberg，2009年），社群推断则使用了 Igraph 软件包（Cs´ardi 和 Nepusz，2006年）。秩相关的计算使用 SciPy 完成（Virtanen 等人，2020年）。补充材料提供了复现本文所报告结果所需的全部数据和代码。

4 结果

4.1 高效共词化推断的计算时间

为了检验新提出的词缀共词化推断方法是否确实比一种概念上简单得多的做法更高效——即将词表中的每个词与其他所有词逐一比较——我们设计了一项小型实验，对源自 Allen（2007）的白语方言 CLDF 数据集进行了多次分析，并计算了所需的计算时间。测试结果表明，与朴素迭代相比，新方法在计算时间方面确实高效得多。在不同 Linux 机器上进行的多项实验中，计算时间的差异显示，朴素的全对全词语比较所需时间是新型高效方法的五倍以上，而两者得出的结果完全相同。对于仅包含十几种语言的小型数据集，计算时间或许不那么重要；但处理《洲际词典系列》等大型数据集时，计算时间可能会成为瓶颈。因此，本文提出的高效解决方案证明非常实用。不过，这并不意味着该方案十全十美；也可能存在更高效的解决方案（例如使用后缀树，参见 Gusfield，1997年，122–180 页），未来或可将其实现。

4.2 比较度分布

计算出完全共词化、词缀共词化和重叠共词化的共词化网络后，我们针对这三种共词化类型的加权度分布计算了斯皮尔曼秩相关，并将词缀共词化拆分为两种度分布：入度和出度。比较结果见表 1。如表中所示，在总共六种配对中，可以观察到两种中等程度的相关性。完全共词化网络的度分布与词缀共词化网络的出度分布呈中等相关（$r = 0.50, p < 0.0001$）；重叠共词化网络的度分布与词缀共词化网络的入度分布呈中等相关（$r = 0.42, p < 0.0001$）。

乍看之下，对这些结果的解读似乎并不直接。完全共词化网络中概念的加权度与词缀共词化网络中概念的出度之间的相关性表明，完全共词化频率较高的概念，似乎也经常被重新使用为复合词或复杂词中的词缀。尽管这一发现可能显得相当合理，甚至显而易见，但迄今为止，跨语言研究尚无法证实这一点

A 型共词化	B 型共词化	节点	R	P 值
完全共词化	词缀共词化（入度）	1308	0.0960	< 0.0001
完全共词化	词缀共词化（出度）	1308	0.5034	< 0.0001
完全共词化	重叠共词化	1307	0.1179	< 0.0001
词缀共词化（入度）	词缀共词化（出度）	1308	-0.0830	< 0.0001
词缀共词化（入度）	重叠共词化	1307	0.4212	< 0.0001
词缀共词化（出度）	重叠共词化	1307	-0.0488	0.0104

表 1：比较四种不同度分布的斯皮尔曼秩相关。可以看出，完全共词化与词缀共词化的出度（0.5）之间，以及词缀共词化的入度与重叠共词化（0.42）之间，都存在显著的中度相关。对于其他配对，则未观察到显著相关。

。因此，部分共词化网络让我们注意到这样一个重要的概念属性：在世界各语言中倾向于频繁共词化的概念，具有通过构词过程被重新利用以形成新词的倾向。我提议将这一属性称为词根生产力（这一术语源于我与 Alexandre Fran̦cois 的一次讨论，参见 List 2019a 和 List 2019b）。词根生产力在词汇动因中发挥着关键作用；词汇动因是世界各语言中形成新词形的过程（Koch，2001年）。

重叠共词化的加权度分布与词缀共词化的出度分布之间的相关性，还有一个更为直接的解释。在更大规模的语言样本中，表现出许多重叠共词化的概念，往往是借助复合词或形态复杂词来表达的概念。词缀共词化网络中有许多入边的概念也是如此。因此，这两种度分布之间的相关性并不十分令人意外。不过，这表明，词缀共词化网络的加权入度和重叠共词化网络的加权度都可以作为衡量概念复合性的指标（这一术语的灵感来自 Martin Haspelmath，个人通信），也就是说，用复合词或形态复杂词表达概念的倾向。

4.3 通过子图考察共词化

尽管对度分布的考察已经让我们很好地了解了不同类型共词化网络之间的共性与差异，但更仔细地考察图中的较小部分，有助于我们更清楚地看出这些差异。为了提供有价值的样本，我们使用Infomap算法（Rosvall和Bergstrom，2008）从完全共词化网络中计算社群。第二步，选取了23个在完全共词化和部分共词化方面呈现不同特性的概念，并提取了相应的完全共词化、词缀共词化和重叠共词化网络子图，再借助Cytoscape（Shannon等人，2003；Smoot等人，2011）将其可视化。

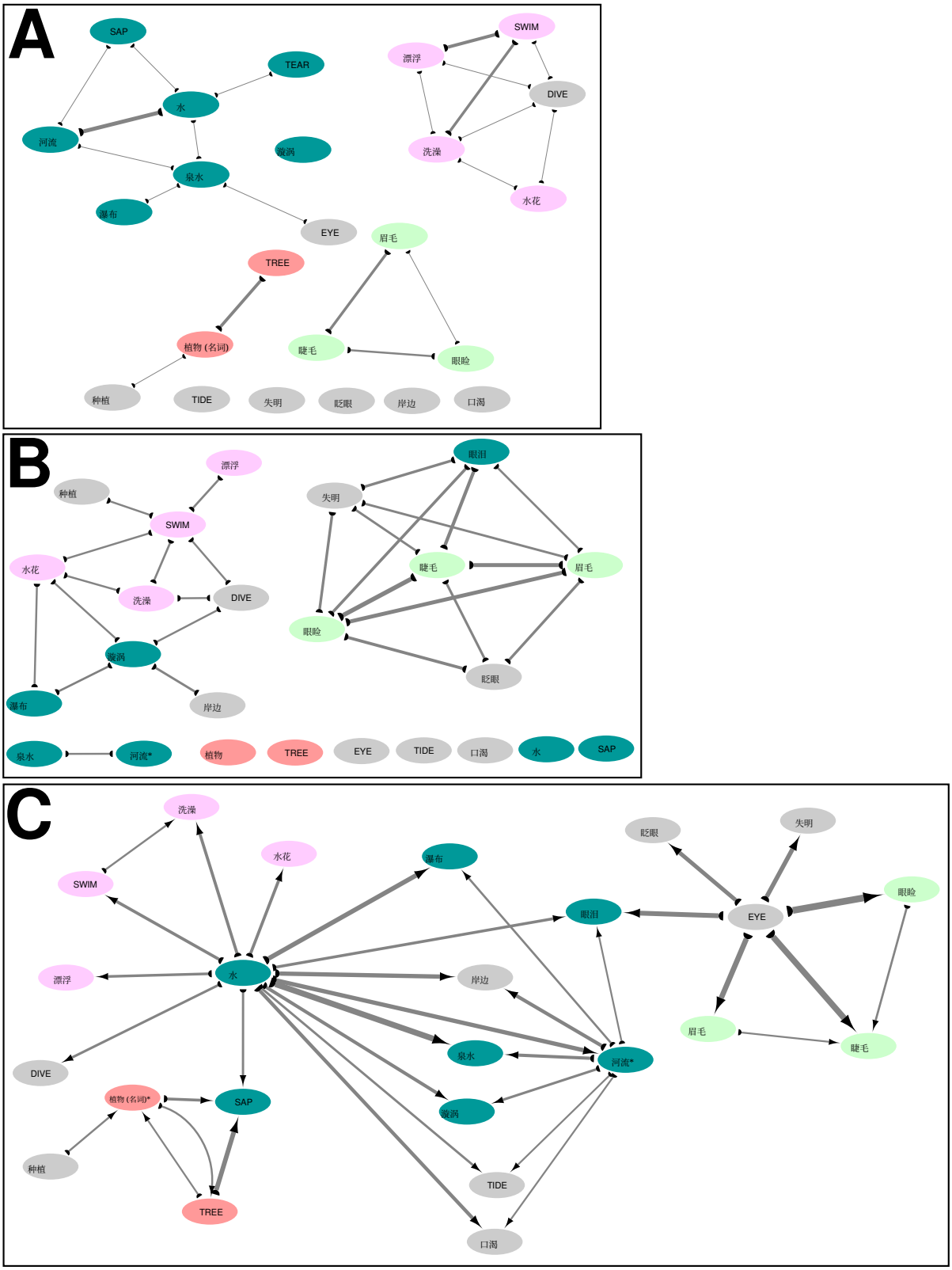

图4：比较 IDS 数据集子图中的完全共词化（A）、重叠共词化（B）和词缀共词化（C）。线条宽度表示共词化的权重；浅灰色以外的颜色表示为完全共词化网络推断出的群落；词缀共词化网络（B）中的连接方向借助箭头显示。概念标签取自 Concep- ticon 项目。带星号的概念标签经过修改，以增强可视化效果。

从图4所示的可视化结果可以看出，尽管这三个网络都涉及相同的概念，但它们各自的结构却有显著差异。例如，概念“眼睛”在完全共词化网络中仅与“泉水”存在一条虚假连接（A），在重叠共词化网络中则完全孤立（B），而在词缀共词化网络中却是一个相当核心、出度较高的概念（C）。考察这三个网络中的连通分量时，我们发现彼此完全连接的概念之间存在巨大差异；与此同时，也很容易发现促成这些模式的语义或形态联系。在重叠共词化网络中，我们发现了一个由“失明”“眼泪”“睫毛”“眉毛”“眼睑”和“眨眼”组成的簇，这显然是因为表达这些概念的词语都包含表示“眼睛”的语素。眼睛在词缀共词化网络中的核心位置印证了这一作用；在该网络中，“水”也呈现出类似的结构，成为另一个核心概念。系统地比较这些不同类型的共词化网络，有助于我们识别出在为世界上许多语言的词汇库构建提供形态材料方面发挥重要作用的语义关键参与者。

5 讨论

本研究提出了从多语言词表推断部分共词化网络的新思路，介绍了可用于处理部分共词化模式的新模型，并提出了用于推断部分共词化网络的新型高效方法和工作流程。研究介绍了在网络中处理部分共词化模式的两种新方式，即词缀共词化和重叠共词化。利用这些新型共词化模式，从大型多语言词表中推断词缀共词化网络和重叠共词化网络，揭示了这两种网络类型的一些有趣特性。重叠共词化网络能够帮助我们衡量世界各语言中各个概念的复合性；词缀共词化网络则可作为初步指标，用于衡量跨语言的词根生产力。除了对词汇类型学领域的研究者具有意义外，我们认为，这些新型共词化网络未来还可广泛应用于其他科学领域，尤其是计算机科学（以及计算语义学方法）和心理学。

资金支持

本研究得到了马克斯·普朗克学会研究资助*CALC3*（JML，https:// digling.org/calc/）以及 ERC Consolidator Grant *ProduSemy*（JML，资助编号 101044282，见 https: // cordis.europa.eu/project/id/101044282）的支持。

致谢

我感谢Robert Forkel就部分代码库提供的宝贵意见，以及创建《洲际词典系列》的分段版本。

补充材料

本研究所用的数据和代码托管在 GitHub 上，可通过 https: // github.com/lingpy/pacs/releases/tag/v0.1（版本 0.1）访问。论文发表后，数据和代码也将存档于 Zenodo。

数据可用性声明

本研究的相关数据和代码均免费公开。本研究所用的数据和代码托管在 GitHub 上，可通过 https://github.com/lingpy/pacs/releases/ tag/v0.1（版本 0.1）访问。论文发表后，数据和代码也将存档于 Zenodo。

参考文献

Allen, B. (2007). 白语方言调查 （达拉斯：国际语言学暑期学院）

Anderson, C., Tresoldi, T., Chacon, T. C., Fehn, A.-M., Walworth, M., Forkel, R., 等. (2018). 语音转写系统跨语言数据库。波兹南语言学会议年鉴 ´4, 21–53

Bao, H., Hauer, B., 和 Kondrak, G. (2021). 关于普遍共词化。载于第十一届全球词网会议论文集 （南非大学（南非大学）：全球词网协会），1–7

Bastian、M.、Heymann、S. 和 Jacomy、M.（2009年）。Gephi：用于探索和操作网络的开源软件。载于第三届国际*AAAI*网络日志与社交媒体会议论文集 （人工智能促进协会），1–2

Cs´ardi、G. 和 Nepusz、T.（2006）。用于复杂网络研究的igraph软件包。国际期刊。《复杂系统》

Dehmer、M.、Emmert-Streib、F.、Graber、A. 和 Salvador、A.（2011）。网络生物学应用统计学。系统生物学方法。引言 （魏恩海姆：Wiley-Blackwell）

Dellert, J., Daneyko, T., M¨unch, A., Ladygina, A., Buch, A., Clarius, N., 等（2020）。Northeuralex：北欧亚广覆盖词汇数据库。《语言资源与评估》54，273–301。doi:10.1007/s10579-019-09480-6

Di Natale, A., Pellert, M., 和 Garcia, D.（2021）。共词化网络编码情感意义。《情感科学》 doi:10.1007/s42761-021-00033-1 Durbin, R., Eddy, S. R., Krogh, A., 和 Mitchinson, G.（2002[1998]）。生物序列分析：蛋白质和核酸的概率模型（剑桥：剑桥大学出版社），第7版。Ferraz Gerardi, F., Reichert, S., Aragon, C., List, J.-M., Forkel, R., 和 Wientzek, T.（2021）。*TuLeD*：*Tupʼıan* 词汇数据库 *[*数据集，版本 *0*。*11]*（莱比锡：马克斯·普朗克进化人类学研究所）。 doi:10.5281/zenodo.4629306 Forkel, R. 和 List, J.-M.（2020）。CLDFBench：让你的跨语言数据更上一层楼。载于第十二届语言资源与评测国际会议论文集（卢森堡：欧洲语言资源协会（ELRA）），6997–7004 Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., 等（2018）。跨语言数据格式：推动比较语言学中的数据共享与再利用。《科学数据》5，1–10 Fran˛cois, A.（2008）。语义地图与共词化类型学：交织跨语言的多义网络。载于从多义到语义变化，M. Vanhove 编（阿姆斯特丹：Benjamins）。163–215 Geisler, H., Forkel, R., 和 List, J.-M.（2021）。《瑞士罗曼语方言语音表》（tppsr）的数字化回溯标准化版本。载于方言变异的新视角，M. Avanzi、N. LoVecchio、A. Millour 和 A. Thibault 编（斯特拉斯堡：ʹ语言学与语文学出版社）。13–36 Gusfield, D.（1997）。字符串、树和序列算法（剑桥：剑桥大学出版社）[数据集] Hagberg, A.（2009）。NetworkX：复杂网络高效软件。由作者发布。Harvill, J., Girju, R., 和 Hasegawa-Johnson, M.（2022）。Syn2Vec：用于词汇语义相似度的同义词集共词化图。载于*2022*年北美计算语言学协会分会年会：人类语言技术会议论文集（美国西雅图：计算语言学协会），5259–5270。doi:10.18653/v1/2022.naacl-main.386 Haspelmath, M. 和 Tadmor, U.（2009）。世界借词数据库（莱比锡：马克斯·普朗克进化人类学研究所） Hoehndorf, R., Kelso, J., 和 Herre, H.（2009）。生物序列本体。《*BMC*生物信息学》10。 doi:10.1186/1471-2105-10-377 Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Mucha, P. J., Forkel, R., 等（2019年）。情绪语义既体现文化变异，也展现普遍结构。《科学》366，1517–1522

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., 和 Lindquist, K.（2022）。从文本到思维：分析语言如何推动心理科学发展。《心理科学展望》17，805–826。Key, M. R. 和 Comrie, B.（2016）。《洲际词典系列》（莱比锡：马克斯·普朗克进化人类学研究所）Koch, P.（2001）。从认知和语言学角度看词汇类型学。载于《语言类型学与语言共性》（柏林和纽约：de Gruyter），《语言学与传播科学手册》第20.2卷。1142–1178。Kruskal, J. B. 和 Liberman, M.（1999[1983]）。对称时间扭曲问题：从连续到离散。载于《时间扭曲、字符串编辑与大分子：序列比较的理论与实践》，D. Sankoff 和 J. B. Kruskal 编（斯坦福：CSLI 出版社）。重印版，第125–161页。List, J.-M.（2014）。《历史语言学中的序列比较》（杜塞尔多夫：杜塞尔多夫大学出版社）List, J.-M.（2019a）。计算多样性语言学中的未决问题：结论与展望。《系统发育网络的谱系世界》6。List, J.-M.（2019b）。语义多义性类型学（计算多样性语言学中的未决问题10）。《系统发育网络的谱系世界》6。List, J.-M.（2022）。如何使用 CL Toolkit 计算共词化（语言学中的如何做X 10）。《实践中的计算机辅助语言比较》5。List, J.-M., Anderson, C., Tresoldi, T., 和 Forkel, R.（2023）。跨语言转写系统*[*数据集，版本 *2.3*。*0]*（耶拿：马克斯·普朗克进化人类学研究所）。doi:10.5281/zenodo.3515744。List, J.-M. 和 Forkel, R.（2021）。*CL Toolkit*：用于处理跨语言数据的 *Python* 库 *[*软件库，版本 *0.1*。*1]*（日内瓦：Zenodo）。doi:10.5281/zenodo.5156028。List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., 和 Gray, R. D.（2022）。Lexibank：包含经过计算的语音和词汇特征的标准化词表公共存储库。《科学数据》9，1–31。List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., 和 Forkel, R.（2018）。CLICS2：一个改进版的跨语言共词化数据库，借助跨语言数据格式汇集词汇数据。《语言类型学》22，277–306。List, J.-M., Terhalle, A., 和 Urban, M.（2013）。运用网络方法加强对跨语言多义现象的分析。载于第十届国际计算语义学会议论文集——短文（斯特劳兹堡：计算语言学协会），347–353

Li ́u Lı̀ ̌i, W ́angH ́ongzh ̄ong, and B ̌ai Y ́ıng (2007). ***Xi ̀and ̀ai H ̀any ̌u f ̄angy ́an h ́ex ̄ınc ́ı, t ̀ezh ̄eng c ́ij ́ı*** /《现代汉语方言基本词汇词语和特色方言词语集》 (/ (N ́anj ̄ıng:F ̀enghu ́ang)

Mayer, T., List, J.-M., Terhalle, A., and Urban, M. (2014). 跨语言共词化模式的交互式可视化。载于可视化作为语言资源开发、使用和评估中的附加价值。作为国际语言资源与评估会议的一部分组织的研讨会 (计算语言学协会)。1–8

Newman, M. E. J. (2010). 网络：入门（牛津：牛津大学出版社）

Rosvall, M. and Bergstrom, C. T. (2008). 复杂网络中的随机游走图揭示了社群结构。美国国家科学院院刊105, 1118–1123

Rzymski, C., Tresoldi, T., Greenhill, S., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., et al. (2020). 跨语言共词化数据库：对跨语言多义性的可复现分析。《科学数据》7,1–12

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D. 等（2003）。Cytoscape：用于集成生物分子相互作用网络模型的软件环境。《基因组研究》13, 2498–2504

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., 和 Ideker, T.（2011）。Cytoscape 2.8：数据集成和网络可视化的新功能。生物信息学 27, 431–432. doi:10.1093/bioinformatics/ btq675

Spearman, C.（1904）。两个事物之间关联的证明与测量。《美国心理学杂志》15, 72–101

Urban, M.（2011）。显性标记与语义变化方向性的非对称性。《历史语言学杂志》1,3–47

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., 等（2020年）。SciPy 1.0：Python 科学计算的基础算法。《自然·方法》17，261–272。doi:10.1038/s41592-019-0686-2

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., 等（2016）。论人类词汇语义学的普遍结构。美国国家科学院院刊doi:10.1073/pnas.1520752113